Paruošė:
Agnė Griniūtė

Vilniaus Universitetas
Matematikos ir informatikos fakultetas
Informacinių sistemų inžinerija
2 k., 1 g.
2020m.

Statistiniai duomenų analizės metodai

Savarankiškas darbas

Logistinė regresija

Turinys

1. Duomenų rinkinio aprašymas

Duomenų rinkinys išsaugotas pavadinimu diabetes2.csv. Duomenų analizei atlikti pasirinktas duomenų rinkinys (šaltinis *https://www.kaggle.com/kandij/diabetes-dataset*) sudarytas iš 9 rodmenų ir 724 stebėjimų. Duomenys surinkti "National Institute of Diabetes and Digestive and Kidney Diseases" organizacijos. Duomenų rinkinyje pateikta informacija apie moterų nuo 21 metų fiziologinius duomenis ir jų sergamumą diabetu. Sergamumas šioje logistinėje regresijoje bus priklausomas kintamasis, kuris gali įgyti tik reikšmes 0 arba 1.

Informacija apie duomenų rinkinio "*diabetes2*" rodiklius:

- Nėštumų skaičius (angl. Pregnancies);
- Gliukozės kiekis (angl. Glucose);
- Kraujo spaudimas (angl. Blood pressure);
- Odos storumas (angl. Skin thickness) milimetrais;
- Insulino kiekis (angl. Insulin);
- Kūno masės indeksas (angl. BMI);
- Diabeto paveldimumo rodiklis (angl. Diabetes pedigree function);
- Amžius (angl. Age);
- Rezultatas (angl. Outcome), parodantis ar moteris serga diabetu. 0 - jei neserga, 1 - jei serga.

Dirbdama RStudio aplinkoje visų pirma paruošiu duomenų rinkinį analizei:

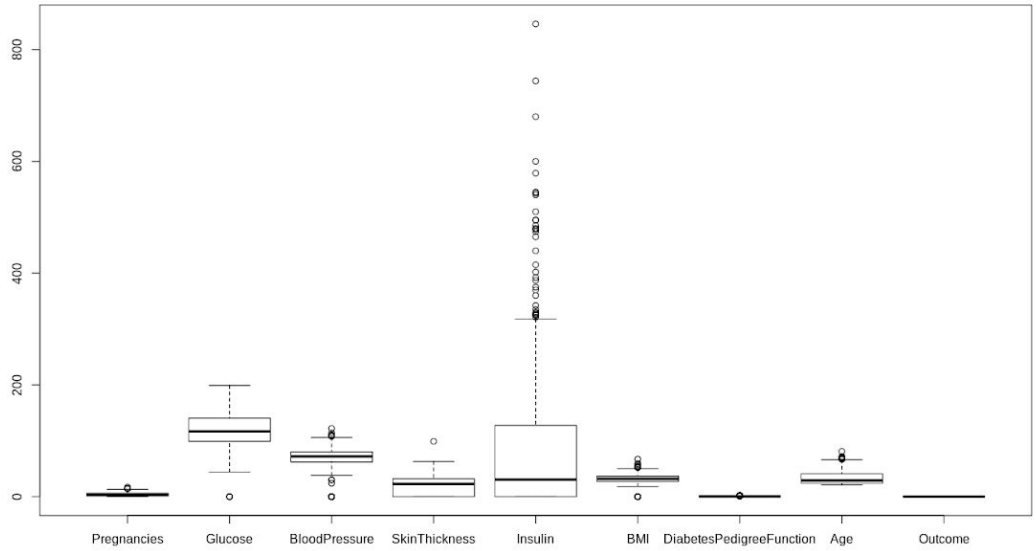
- Patikrinu ar nėra praleistų reikšmių. Praleistų reikšmių nėra, todėl duomenų rinkinys tinkamas.

```
> col1 <- mapply(anyNA, diabetes2)
> print(col1)
            Pregnancies                 Glucose           BloodPressure           SkinThickness
                  FALSE                   FALSE                   FALSE                   FALSE
                Insulin                     BMI DiabetesPedigreeFunction                     Age
                  FALSE                   FALSE                   FALSE                   FALSE
                Outcome
                  FALSE
```

1 pav. Tikrinama ar nėra praleistų reikšmių.

- Patikrinu ar duomenų rinkinys turi išskirčių. Kaip matyti iš 2 paveikslėlio, duomenų rinkinyje išskirčių gana daug, tačiau neskubu jų trinti. Visų pirma išsiaiškinu kokios tai išskirtys. Daugelyje rodiklių randu išskirtis, kurių reikšmė lygi 0. Šias išskirtis ištrinu,

nes tokie rodikliai kaip kraujo spaudimas, gliukozės kiekis ar kūno masės indeksas negali būti lygūs nuliui. Panaikinus šias išskirtis, vis dar išlieka kitos išskirtys, bet jų nenaikinu, nes pavyzdžiui insulino ar diabeto paveldimumo rodiklis gali turėti didelę reikšmė sudaromam modeliui, todėl būtų netikslinga ištrinti šias reikšmes ir taip suklaidinti kuriamą modelį.



2 pav. Ieškoma išskirčių.

2. Naudojamų rodiklių aprašomoji statistika

Atlieku rodiklių aprašomąją statistiką bei randu kiekvieno rodiklio standartinį nuokrypį. Gauti rezultatai pavaizduoti 3 ir 4 paveikslėliuose:

```
> summary(diabetes2)
  Pregnancies        Glucose       BloodPressure   SkinThickness      Insulin            BMI
 Min.   : 0.000   Min.   : 44.00   Min.   : 24.0   Min.   : 0.00   Min.   :  0.00   Min.   :18.20
 1st Qu.: 1.000   1st Qu.: 99.75   1st Qu.: 64.0   1st Qu.: 0.00   1st Qu.:  0.00   1st Qu.:27.50
 Median : 3.000   Median :117.00   Median : 72.0   Median :24.00   Median : 48.00   Median :32.40
 Mean   : 3.866   Mean   :121.88   Mean   : 72.4   Mean   :21.44   Mean   : 84.49   Mean   :32.47
 3rd Qu.: 6.000   3rd Qu.:142.00   3rd Qu.: 80.0   3rd Qu.:33.00   3rd Qu.:130.50   3rd Qu.:36.60
 Max.   :17.000   Max.   :199.00   Max.   :122.0   Max.   :99.00   Max.   :846.00   Max.   :67.10
 DiabetesPedigreeFunction      Age           Outcome
 Min.   :0.0780           Min.   :21.00   Min.   :0.0000
 1st Qu.:0.2450           1st Qu.:24.00   1st Qu.:0.0000
 Median :0.3790           Median :29.00   Median :0.0000
 Mean   :0.4748           Mean   :33.35   Mean   :0.3439
 3rd Qu.:0.6275           3rd Qu.:41.00   3rd Qu.:1.0000
 Max.   :2.4200           Max.   :81.00   Max.   :1.0000
```

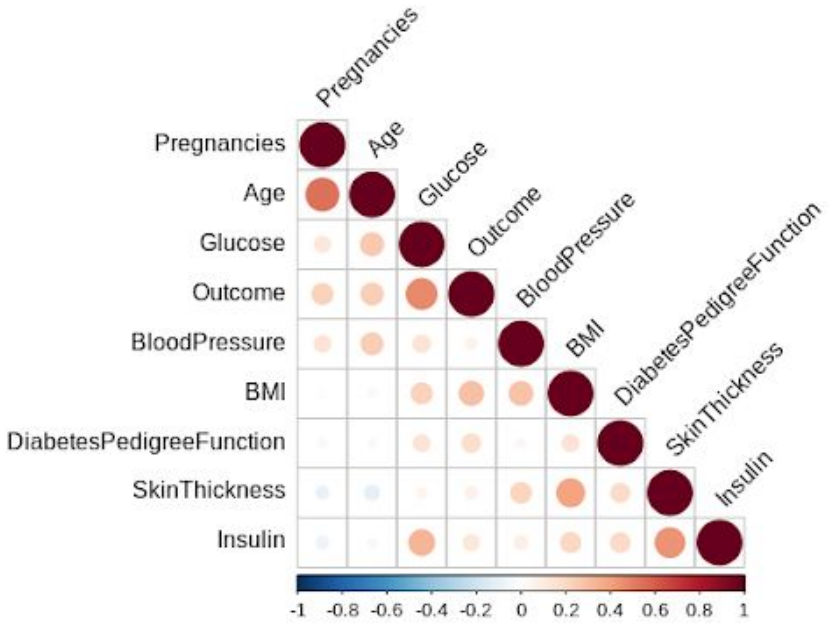
3 pav. Aprašomoji statistika

```
> sd(diabetes2$Pregnancies)
[1] 3.362803
> sd(diabetes2$Glucose)
[1] 30.75003
> sd(diabetes2$BloodPressure)
[1] 12.37987
> sd(diabetes2$SkinThickness)
[1] 15.73276
> sd(diabetes2$Insulin)
[1] 117.0165
> sd(diabetes2$BMI)
[1] 6.888941
> sd(diabetes2$DiabetesPedigreeFunction)
[1] 0.332315
> sd(diabetes2$Age)
[1] 11.76539
> sd(diabetes2$Outcome)
[1] 0.475344
```

4 pav. Standartinis nuokrypis.

3. Koreliacija tarp rodiklių

Kaip matyti iš 5 paveikslėlio, nepriklausomi rodikliai tarpusavyje stipriai nekoreliuoja, vadinasi turimi rodikliai tinkami regresijos modeliui sudaryti, multikolinearumo nėra. Stipriausiai koreliuoja amžius ir nėštumų skaičius, todėl reikėtų vengti abu rodiklius įtraukti į modelį.



5 pav. Koreliacija tarp rodiklių.

4. Sudaromas modelis

Prieš sudarant modelį, duomenų rinkinys “diabetes2” buvo atsitiktinai padalytas į dvi dalis: apmokymų duomenų rinkinį (“trainData”), kuris sudaro 80%, ir testavimo duomenų rinkinį (“testData”), kuris sudaro 20% pirminio duomenų rinkinio. Šis veiksmas suteiks galimybę objektyviai įvertinti modelio tikslumą su duomenimis, kurių modelis dar nematė mokymosi metu.

Visų pirma sudarau logistinės regresijos modelį, į jį įtraukdama visus nepriklausomus rodiklius, bet iš 6 paveikslėlyje pateiktos informacijos matyti, kad tokie rodikliai kaip amžius, kraujo spaudimas, odos storumas ir insulino kiekis nėra statistiškai reikšmingi, nes p > 0.05. Tikrinama hipotezė

$$\begin{cases} H_0 : b = 0, \\ H_1 : b \neq 0. \end{cases}$$

Šiuo atveju nulinė hipotezė neatmetama, tai reiškia, kad Y nepriklauso nuo X ir modelį reikia taisyti.

```
Call:
glm(formula = Outcome ~ Pregnancies + Age + Glucose + BloodPressure +
    BMI + DiabetesPedigreeFunction + SkinThickness + Insulin,
    family = binomial, data = diabetes2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6374  -0.7152  -0.4079   0.7214   2.4149

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -9.0572456  0.8276228 -10.944  < 2e-16 ***
Pregnancies              0.1161430  0.0335394   3.463 0.000534 ***
Age                      0.0159706  0.0098765   1.617 0.105871
Glucose                  0.0369312  0.0039120   9.441  < 2e-16 ***
BloodPressure           -0.0105289  0.0087177  -1.208 0.227137
BMI                      0.0943048  0.0169173   5.574 2.48e-08 ***
DiabetesPedigreeFunction 1.0076042  0.3109647   3.240 0.001194 **
SkinThickness            0.0001526  0.0071103   0.021 0.982879
Insulin                 -0.0011241  0.0009266  -1.213 0.225066
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 931.94  on 723  degrees of freedom
Residual deviance: 671.13  on 715  degrees of freedom
AIC: 689.13

Number of Fisher Scoring iterations: 5
```

6 pav. Modelio aprašymas.

Sudarau naują logistinės regresijos modelį (žr. 7 pav.), šiuo atveju įtraukdama tik statistiškai reikšmingus regresorius - nėštumų skaičių, gliukozės kiekį, kūno masės indeksą ir diabeto paveldimumo rodiklį.

Parametrų įverčiai:

α = -8.97;

β_1 = 0.14;

β_2 = 0.04;

β_3 = 0.09;

β_4 = 0.75.

Gaunama lygtis:

z(x) = -8.97 + 0.14 Pregnancies + 0.04 Glucose + 0.09 BMI + 0.75 DiabetesPedigreeFunction;

```
> LogM_2 = glm(Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction, data = trainData, family = binomial)
> summary(LogM_2)

Call:
glm(formula = Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction,
    family = binomial, data = trainData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6531  -0.7234  -0.3939   0.7179   2.2038

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.969230   0.794679 -11.287  < 2e-16 ***
Pregnancies              0.144441   0.031843   4.536 5.73e-06 ***
Glucose                  0.035230   0.003864   9.118  < 2e-16 ***
BMI                      0.086827   0.017041   5.095 3.48e-07 ***
DiabetesPedigreeFunction 0.753146   0.330514   2.279   0.0227 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 745.12  on 578  degrees of freedom
Residual deviance: 540.81  on 574  degrees of freedom
AIC: 550.81

Number of Fisher Scoring iterations: 5
```

7 pav. Naujo modelio aprašymas.

Modelio regresorių galimybių santykiai ir pasikliautinieji intervalai

Regresorių galimybių santykiai šiuo atveju yra gana maži, labiausiai išsiskiria diabeto paveldimumo rodiklis.

```
> exp(LogM_2$coefficients)
            (Intercept)             Pregnancies                 Glucose                     BMI
           0.0001272662            1.1553936932            1.0358580196            1.0907076207
DiabetesPedigreeFunction
           2.1236712808
```

8 pav. Regresorių galimybių santykiai

Pasikliautinieji intervalai šiame modelyje nėra dideli.

```
> exp(confint.default(LogM_2))
                                2.5 %       97.5 %
(Intercept)                2.680902e-05 0.0006041505
Pregnancies                1.085488e+00 1.2298015275
Glucose                    1.028043e+00 1.0437320072
BMI                        1.054881e+00 1.1277512966
DiabetesPedigreeFunction   1.111100e+00 4.0590217418
```

9 pav. Pasikliautinieji intervalai

5. Modelio tinkamumas

1. Modelio suderinamumo kriterijus

```
> LogM.tuscias <- glm(Outcome ~ 1, data = trainData, family = binomial())
> anova(LogM.tuscias, LogM_2, test = "Chisq")
Analysis of Deviance Table

Model 1: Outcome ~ 1
Model 2: Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       578     745.12
2       574     540.81  4   204.31 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

10 pav. Modelio suderinamumo kriterijus.

Atmetama nulinė hipotezė, kuri teigia, jog nė vienas regresorius modelyje nėra reikšmingas.

2. Akaikės informacinis kriterijus (AIC)

Šio modelio AIC yra lygus 550.81 ir jis yra mažesnis už tuščio modelio AIC (< 745.12), kas rodo, jog modelis yra tinkamas.

3. Determinacijos koeficientas

```
> rkv <- 1 - LogM_2$deviance / LogM_2$null.deviance
> rkv
[1] 0.2742025
```

11 pav. Determinacijos koeficientas

Determinacijos koeficientas nėra didelis, tačiau didesnis nei 0.2, o tai leidžia daryti išvadas, kad modelis yra tinkamas.

6. Prognozavimas

Visų pirma prognozavimas buvo atliekamas pasitelkiant apmokymo duomenis. Gauti rezultatai atvaizduoti lentelėje. Kaip matoma, modelis teisingai prognozavo 334 sveikus pacientus (iš 380), taigi sveikų pacientų teisingas klasifikavimas yra 0.88. Sergančių pacientų teisingas klasifikavimas lygus 0.58. Bendras teisingai klasifikuotų pacientų procentas: 0.77.

```
> progTrain = predict(LogM_2, type = "response")
> table(trainData$Outcome, progTrain > 0.5)

   FALSE TRUE
  0   334   46
  1    85  114
> #sensitivity
> 114/199
[1] 0.5728643
> #specifity
> 334/380
[1] 0.8789474
> #accuracy
> 448/579
[1] 0.7737478
```

12 pav. Duomenų rinkinio "trainData" prognozavimo tikslumas.

Siekiant įsitikinti modelio tikslumu, prognozavimas buvo atliktas duomenų rinkiniui "trainData". Šiuo atveju sveikų pacientų teisingas klasifikavimas lygus 0.92. Sergančių: 0.42. Bendras teisingai klasifikuotų pacientų procentas: 0.75.

```
> progTest = predict(LogM_2, type = "response", newdata = testData)
> table(testData$Outcome, progTest >= 0.5)

    FALSE TRUE
  0    88    7
  1    29   21
> #sensitivity
> 21/50
[1] 0.42
> #specify
> 88/95
[1] 0.9263158
> #accuracy
> 109/145
[1] 0.7517241
```

13 pav. Duomenų rinkinio "testData" prognozavimo tikslumas.

7. Modelio aprašymas

Iš logistinės regresijos modelio pirmiausia buvo pašalinti nereikšmingi ir modeliui įtakos nedarantys regresoriai: amžius, kraujo spaudimas, odos storumas ir insulino kiekis. Gautas galutinis modelis yra statistiškai reikšmingas.

- Modelio suderinamumo kriterijus: 204 ($p < 0.05$);
- Determinacijos koeficientas: 0.27;
- AIC: 550.81;
- Regresoriaus *(Pregnancies)* galimybių santykis 1.16 (95% pasikliautinis intervalas [1.09; 1.22];
- Regresoriaus *(Glucose)* galimybių santykis 1.04 (95% pasikliautinis intervalas [1.02; 1.04];
- Regresoriaus *(BMI)* galimybių santykis 1.09 (95% pasikliautinis intervalas [1.05; 1.13];
- Regresoriaus *(DiabetesPedigreeFunction)* galimybių santykis 2.12 (95% pasikliautinis intervalas [1.11; 4.06];

Naudojantis sudarytu modeliu teisingai klasifikuoti 92.63% sveiki pacientai ir 42% sergančių pacientų.

8. Išvados

Visų pirma atlikau duomenų rinkinio diabetes2.csv priešanalizę: patikrinau ar nėra praleistų reikšmių, radau išskirtis. Atmetusi statistiškai nereikšmingus regresorius ir atlikus logistinės regresijos tinkamumo analizę, įvertinau modelio suderinamumo kriterijų, determinacijos koeficientą, AIC rodiklį. Analizės rezultatai patikino, jog sudarytas modelis yra tinkamas. Gauti prognozavimo rezultatai taip pat leido padaryti išvadas, jog modelis atlieka gana tikslias prognozes.

9. Priedai

Programinis kodas

```
#patikrinu ar duomenu rinkinyje nevyrauja viena is reiksmiu
table(diabetes2$Outcome)

# tikrinu ar nera praleistu reiksmiu
col1 <- mapply(anyNA, diabetes2)
print(col1)

#tikrinu ar nera isskirciu
boxplot(diabetes2)
outlier_values1 <- boxplot.stats(diabetes2$Pregnancies)$out
print(outlier_values1)
outlier_values2 <- boxplot.stats(diabetes2$Glucose)$out
print(outlier_values2)
outlier_values3 <- boxplot.stats(diabetes2$BloodPressure)$out
print(outlier_values3)
outlier_values4 <- boxplot.stats(diabetes2$SkinThickness)$out
print(outlier_values4)
outlier_values5 <- boxplot.stats(diabetes2$Insulin)$out
print(outlier_values5)
outlier_values6 <- boxplot.stats(diabetes2$BMI)$out
print(outlier_values6)
outlier_values7 <- boxplot.stats(diabetes2$DiabetesPedigreeFunction)$out
print(outlier_values7)
outlier_values8 <- boxplot.stats(diabetes2$Age)$out
print(outlier_values8)

#pasalinu išskirtis, kurios lygios 0
diabetes2 <- diabetes2[-which (diabetes2$Glucose == 0), ]
diabetes2 <- diabetes2[-which (diabetes2$BloodPressure == 0), ]
diabetes2 <- diabetes2[-which (diabetes2$BMI == 0), ]
boxplot(diabetes2)

summary(diabetes2)
sd(diabetes2$Pregnancies)
sd(diabetes2$Glucose)
sd(diabetes2$BloodPressure)
sd(diabetes2$SkinThickness)
sd(diabetes2$Insulin)
sd(diabetes2$BMI)
sd(diabetes2$DiabetesPedigreeFunction)
sd(diabetes2$Age)

#tikrinama koreliacija
source("http://www.sthda.com/upload/rquery_cormat.r")
rquery.cormat(diabetes2)

#duomenu rinkiniu padalijimas
library(caTools)
set.seed(99)
split = sample.split(diabetes2$Outcome, SplitRatio = 0.80)
split
trainData = subset(diabetes2, split == TRUE)
testData = subset(diabetes2, split == FALSE)
```

```
trainData
testData

#sudaromas modelis
LogM_1 = glm(Outcome ~ Pregnancies + Age + Glucose + BloodPressure + BMI +
DiabetesPedigreeFunction + SkinThickness + Insulin, data = trainData, family =
binomial)
summary(LogM_1)

#sudaromas naujas modelis
LogM_2 = glm(Outcome ~ Pregnancies + Glucose + BMI +
DiabetesPedigreeFunction, data = trainData, family = binomial)
summary(LogM_2)

#regresoriu galimybiu santykiai ir pasikliautinieji intervalai
exp(LogM_2$coefficients)
exp(confint.default(LogM_2))

#modelio suderinamumo kriterijus
LogM.tuscias <- glm(Outcome ~ 1, data = trainData, family = binomial())
anova(LogM.tuscias, LogM_2, test = "Chisq")

#determinacijos koeficientas
rkv <- 1 - LogM_2$deviance / LogM_2$null.deviance
rkv

#prognozavimas
progTrain = predict(LogM_2, type = "response")
table(trainData$Outcome, progTrain > 0.5)

#sensitivity
114/199
#specifity
334/380
#accuracy
448/579

#modelio prognoziu tikslumas
progTest = predict(LogM_2, type = "response", newdata = testData)
table(testData$Outcome, progTest >= 0.5)

#sensitivity
21/50
#specifity
88/95
#accuracy
109/145
```