SVEUČILIŠTE U ZAGREBU
FAKULTET ORGANIZACIJE I INFORMATIKE
VARAŽDIN

Petar Matišić
Filip Antunović

TEŽINA

SEMINAR

Varaždin, 2023.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Petar Matišić
Filip Antunović

Matični brojevi: 0016145882, 0016143851

Studij: Informacijsko i programsko inženjerstvo

TEŽINA

SEMINAR

Mentorice:

prof. dr. sc. Jasminka Dobša
Maja Buhin Pandur, prof. math.

Varaždin, svibanj 2023.

Petar Matišić
Filip Antunović

Izjava o izvornosti

Izjavljujemo da je ovaj seminar izvorni rezultat našeg rada te da se u izradi istoga nismo koristili drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autori potvrdili prihvaćanjem odredbi u sustavu FOI Radovi

Sažetak

U ovom istraživanju, koristeći skup podataka o tjelesnim mjerenjima, izračunat je indeks tjelesne mase (BMI) koristeći varijable težine i visine. Statistička analiza uključivala je opisivanje varijabli i njihovo grafičko prikazivanje, kao i izračun matrice korelacija. Normalnost distribucije svih varijabli je ispitana. Varijabla BMI je rekodirana u kategorije (pothranjenost, idealna težina, prekomjerna tjelesna masa i pretilost), a zatim je provedena analiza varijance (ANOVA) za ispitivanje razlika u mjerenjima tijela u odnosu na te kategorije. Nadalje, definiran je model linearne regresije s postotkom tjelesne masti kao zavisnom varijablom, a ostale kvantitativne varijable kao nezavisne. Parametri modela, uključujući koeficijent determinacije i p-vrijednosti, su interpretirani. Konačno, prikazani su i analizirani reziduali regresijskog modela kako bi se procijenile pretpostavke regresijske analize.

Ključne riječi: statistička analiza, ANOVA, regresija

Sadržaj

1. Uvod

U ovom seminarskom radu provodi se detaljna analiza i istraživanje skupa podataka na temelju procjene postotka tjelesne masti kod muškaraca. Poznavanje informacije o potkožnom masnom tkivu trebalo bi biti u interesu svakog pojedinca jer igra veliku ulogu u procjeni zdravlja individue. Vjerodostojna mjerenja potkožnog masnog tkiva najčešće su izrazito skupa pa se sve više naglasak stavlja na metode koje se bave procjenom iznosa te vrijednosti. Neke od metoda procjene koje se tada koriste su metoda procjene kaliperom mjerenjem debljine kožnih nabora, izračun mjerenjima opsega dijelova tijela (trbuh i sl.), izračuni mjerenjem preklapanja kožnih nabora. Istraživanje se sastoji od više zadataka, koji uključuju izračunavanje novih varijabli, statističku analizu, vizualizaciju podataka, izračun matrice korelacija, testiranje normalnosti distribucije, analizu varijance i regresijsku analizu.

Skup podataka [1] koji potječe iz izvora dostupnog na `http://lib.stat.cmu.edu/datasets/bodyfat` sadrži procjene postotka tjelesne masti kod 252 muškarca, dobivene podvodnim mjerenjem težine i raznim mjerenjima obujma tijela. Cilj je pružiti metode za procjenu postotka tjelesne masti koje su jednostavnije i jeftinije od direktnih mjerenja. Skup podataka uključuje varijable kao što su gustoća tijela određena podvodnim mjerenjem, postotak tjelesne masti izračunat pomoću Siri-jeve jednadžbe, dob, težina, visina, te opseg vrata, prsa, trbuha, bokova, bedara, koljena, gležnja, bicepsa, podlaktice i zapešća. Ovi podaci mogu se koristiti u edukativne svrhe za ilustraciju tehnika višestruke regresije, te su korisni u istraživanju i razvoju jednostavnih metoda za procjenu tjelesne kompozicije.

U nastavku je prikazana tablica 1. sa podatcima varijabli koje se odnose na skup podataka.

Tablica 1: Popis i opis varijabli

Varijabla (HR)	Varijabla (EN)	Tip varijable	Modaliteti
Gustoća (podvodno mjerenje)	Density (underwater)	Kontinuirana	-
Postotak tjelesne masti (Siri)	Percent body fat (Siri)	Kontinuirana	-
Dob (godine)	Age (years)	Kontinuirana	-
Težina (lbs)	Weight (lbs)	Kontinuirana	-
Visina (inči)	Height (inches)	Kontinuirana	-
Opseg vrata (cm)	Neck circ. (cm)	Kontinuirana	-
Opseg prsa (cm)	Chest circ. (cm)	Kontinuirana	-
Opseg trbuha (cm)	Abdomen 2 circ. (cm)	Kontinuirana	-
Opseg bokova (cm)	Hip circ. (cm)	Kontinuirana	-
Opseg bedra (cm)	Thigh circ. (cm)	Kontinuirana	-
Opseg koljena (cm)	Knee circ. (cm)	Kontinuirana	-
Opseg gležnja (cm)	Ankle circ. (cm)	Kontinuirana	-
Opseg bicepsa (istegnut, cm)	Biceps (extended) circ. (cm)	Kontinuirana	-
Opseg podlaktice (cm)	Forearm circ. (cm)	Kontinuirana	-
Opseg zapešća (cm)	Wrist circ. (cm)	Kontinuirana	-

2. Zadatci

U ovom poglavlju se radi obrada, točnije razne analize i testiranja za određene zadatke. Korištene metode navedene su za pojedini zadatak. Da bi se moglo ispravno raditi s podacima, prvo je potrebno napraviti "čišćenje" podataka kako bi oni bili spremni za obradu. Pritom su korišteni jezici Python (za "čišćenje" podataka) i R (za obradu i analizu). Slijedi ukratko pojašnjenje postupka "čišćenja" podataka.

Isječak kôda 1: Učitavanje i obrada podataka

```
import pandas as pd # Uvoz pandas biblioteke za manipulaciju podacima

# Broj redaka na početku datoteke koji ne sadrže podatke i treba ih preskočiti
skip_rows = 117

# Ukupan broj redaka u datoteci
total_lines = 381

# Izračun broja redaka na kraju datoteke koje treba preskočiti
end_skip = total_lines - 370

# Definiranje naslova stupaca za DataFrame
headers = [
    "Density",
    "Percent body fat",
    "Age",
    "Weight",
    "Height",
    "Neck",
    "Chest",
    "Abdomen 2",
    "Hip",
    "Thigh",
    "Knee",
    "Ankle",
    "Biceps",
    "Forearm",
    "Wrist"
]

# Učitavanje podataka iz .txt datoteke u DataFrame, preskačući nepotrebne redove
df = pd.read_csv('bodyfat.txt', delimiter="\s+", skiprows=skip_rows, skipfooter=
 end_skip, names=headers, engine='python')

# Spremanje DataFrame-a u .csv datoteku koristeći točku-zarez kao separator
df.to_csv('data.csv', sep=";", index=False)

# Ispis prvih 5 redaka DataFrame-a
print(df.head())
```

U ovom kodu, glavni cilj je učitati skup podataka iz tekstualne datoteke, očistiti ga i spremiti u CSV format za daljnju analizu. Na početku, uvozi se pandas biblioteka koja je neophodna za manipulaciju podacima. Zatim se postavljaju varijable koje određuju koliko redaka treba preskočiti na početku i na kraju datoteke, jer ovi redovi ne sadrže relevantne podatke. Nakon toga, definiraju se naslovi stupaca koji će biti korišteni u DataFrame-u. Podaci se učitavaju iz tekstualne datoteke preskačući nepotrebne redove i dodajući definirane naslove stupaca. Konačno, DataFrame se sprema u CSV datoteku koristeći točku-zarez kao separator, a prvih pet redaka se ispisuje za pregled.

Isječak kôda 2: Učitavanje i obrada podataka

```
# Učitavanje skupa podataka
data <- read.csv('data.csv', sep = ";", dec = ".", header = TRUE)

# Provjera strukture podataka
str(data)
head(data)

# Zamjena NA vrijednosti srednjim vrijednostima
data$`Weight`[is.na(data$`Weight`)] <- mean(data$`Weight`, na.rm = TRUE)
data$`Height`[is.na(data$`Height`)] <- mean(data$`Height`, na.rm = TRUE)
```

Ovaj isječak koda, koji je napisan u R programskom jeziku, uključuje učitavanje skupa podataka iz CSV datoteke u radnu memoriju. Učitavanje se vrši pomoću funkcije `read.csv`, pri čemu se specificira separator (točka-zarez) i decimalni znak (točka), a parametar `header` postavljen na `TRUE` ukazuje da prvi redak datoteke sadrži nazive stupaca. Nakon učitavanja, koriste se funkcije `str` i `head` kako bi se ispitala struktura podataka i prikazalo prvih nekoliko redaka. Zatim, kod provjerava postoje li u stupcima `Weight` i `Height` nedostajuće (NA) vrijednosti, i ako je to slučaj, one se zamjenjuju srednjim vrijednostima odgovarajućih stupaca, što je čest pristup u obradi nedostajućih podataka.

2.1. Zadatak A

Prvi zadatak (a) zahtijevao je izračunavanje indeksa tjelesne mase (BMI) na temelju varijabli težine (Weight) i visine (Height) pomoću formule

$$\text{BMI} = \frac{\text{Weight}}{\text{Height}^2} \times 703.$$

BMI je dodan u skup podataka kao nova varijabla.

Isječak kôda 3: Izračun BMI

```
# a) Izračun BMI
data$BMI <- (data$`Weight` / (data$`Height`)^2) * 703
```

Nakon izračuna, možemo primijetiti kako je dodan stupac `BMI`.

```
> head(data)
  Density Percent.body.fat Age Weight Height Neck Chest Abdomen.2   Hip Thigh Knee Ankle Biceps Forearm Wrist      BMI
1  1.0708             12.3  23 154.25  67.75 36.2  93.1      85.2  94.5  59.0 37.3  21.9   32.0    27.4  17.1 23.62446
2  1.0853              6.1  22 173.25  72.25 38.5  93.6      83.0  98.7  58.7 37.3  23.4   30.5    28.9  18.2 23.33205
3  1.0414             25.3  22 154.00  66.25 34.0  95.8      87.9  99.2  59.6 38.9  24.0   28.8    25.2  16.6 24.66632
4  1.0751             10.4  26 184.75  72.25 37.4 101.8      86.4 101.2  60.1 37.3  22.8   32.4    29.4  18.2 24.88078
5  1.0340             28.7  24 184.25  71.25 34.4  97.3     100.0 101.9  63.2 42.2  24.0   32.2    27.7  17.7 25.51485
6  1.0502             20.9  24 210.25  74.75 39.0 104.5      94.4 107.8  66.0 42.0  25.6   35.7    30.6  18.8 26.45263
```

Slika 1: Prikaz podataka i stupca `BMI` (desno)

2.2. Zadatak B

Zadatku (b), provedeno je opisivanje varijabli skupa podataka, uključujući i novu varijablu BMI, te grafički prikaz istih. Korištena je funkcija `summary` za prikaz osnovnih statističkih mjera, dok su za grafički prikaz korištene funkcije `plot`, `hist` i `boxplot`.

Isječak kôda 4: Opis varijabli

```
# b) Opisivanje varijabli
summary(data)

# Odabrati proizvoljnu varijablu za prikaz
plot(density(data$BMI))
hist(data$Percent.body.fat)
boxplot(data$Weight)
plot(data$Weight, data$Height) # Usporedba dvaju varijabli

# b) Grafički prikaz
pairs(data)
```

U nastavku su prikazani rezultati navedenih naredbi. Za primjer prikaza po vrsti grafa uzete su neke od ključnih varijabli kao što su `BMI`, `Percent.body.fat` i sl. Nakon toga prikazane su sve varijable korištenjem funkcije `pairs`.

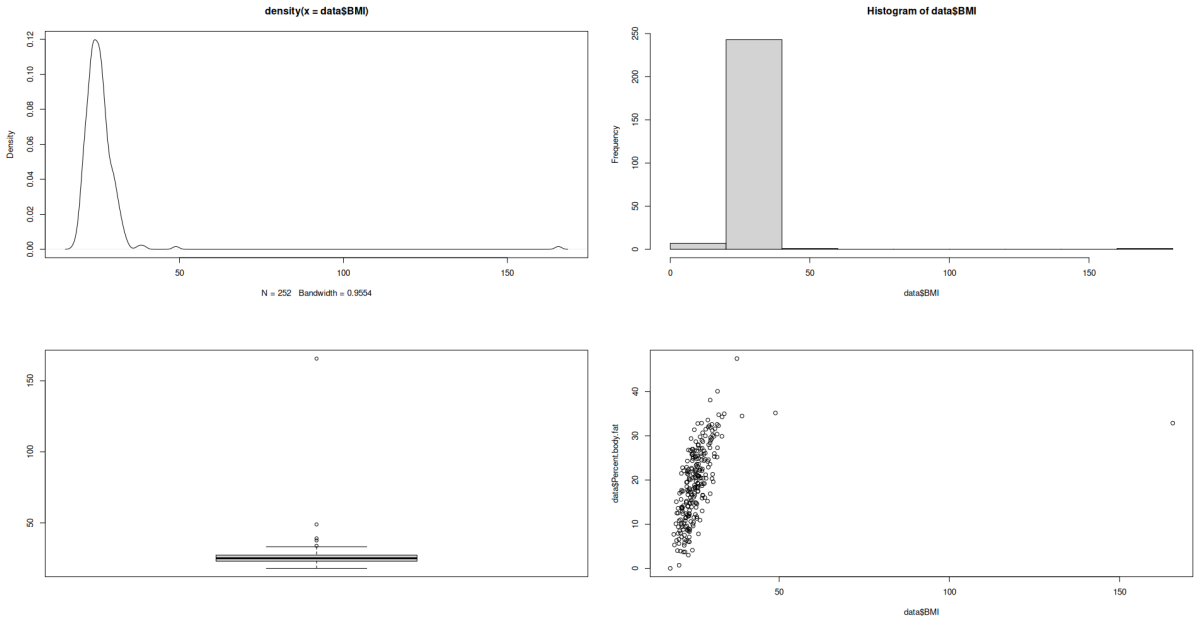
```
> summary(data)
    Density      Percent.body.fat      Age             Weight          Height          Neck            Chest      
 Min.   :0.995   Min.   : 0.00    Min.   :22.00   Min.   :118.5   Min.   :29.50   Min.   :31.10   Min.   : 79.30  
 1st Qu.:1.041   1st Qu.:12.47    1st Qu.:35.75   1st Qu.:159.0   1st Qu.:68.25   1st Qu.:36.40   1st Qu.: 94.35  
 Median :1.055   Median :19.20    Median :43.00   Median :176.5   Median :70.00   Median :38.00   Median : 99.65  
 Mean   :1.056   Mean   :19.15    Mean   :44.88   Mean   :178.9   Mean   :70.15   Mean   :37.99   Mean   :100.82  
 3rd Qu.:1.070   3rd Qu.:25.30    3rd Qu.:54.00   3rd Qu.:197.0   3rd Qu.:72.25   3rd Qu.:39.42   3rd Qu.:105.38  
 Max.   :1.109   Max.   :47.50    Max.   :81.00   Max.   :363.1   Max.   :77.75   Max.   :51.20   Max.   :136.20  
   Abdomen.2          Hip             Thigh            Knee            Ankle           Biceps          Forearm     
 Min.   : 69.40   Min.   : 85.0   Min.   :47.20   Min.   :33.00   Min.   :19.1   Min.   :24.80   Min.   :21.00  
 1st Qu.: 84.58   1st Qu.: 95.5   1st Qu.:56.00   1st Qu.:36.98   1st Qu.:22.0   1st Qu.:30.20   1st Qu.:27.30  
 Median : 90.95   Median : 99.3   Median :59.00   Median :38.50   Median :22.8   Median :32.05   Median :28.70  
 Mean   : 92.56   Mean   : 99.9   Mean   :59.41   Mean   :38.59   Mean   :23.1   Mean   :32.27   Mean   :28.66  
 3rd Qu.: 99.33   3rd Qu.:103.5   3rd Qu.:62.35   3rd Qu.:39.92   3rd Qu.:24.0   3rd Qu.:34.33   3rd Qu.:30.00  
 Max.   :148.10   Max.   :147.7   Max.   :87.30   Max.   :49.10   Max.   :33.9   Max.   :45.00   Max.   :34.90  
     Wrist             BMI        
 Min.   :15.80   Min.   : 18.02  
 1st Qu.:17.60   1st Qu.: 23.03  
 Median :18.30   Median : 25.09  
 Mean   :18.23   Mean   : 25.94  
 3rd Qu.:18.80   3rd Qu.: 27.33  
 Max.   :21.40   Max.   :165.60  
```

Slika 2: Opis varijabli

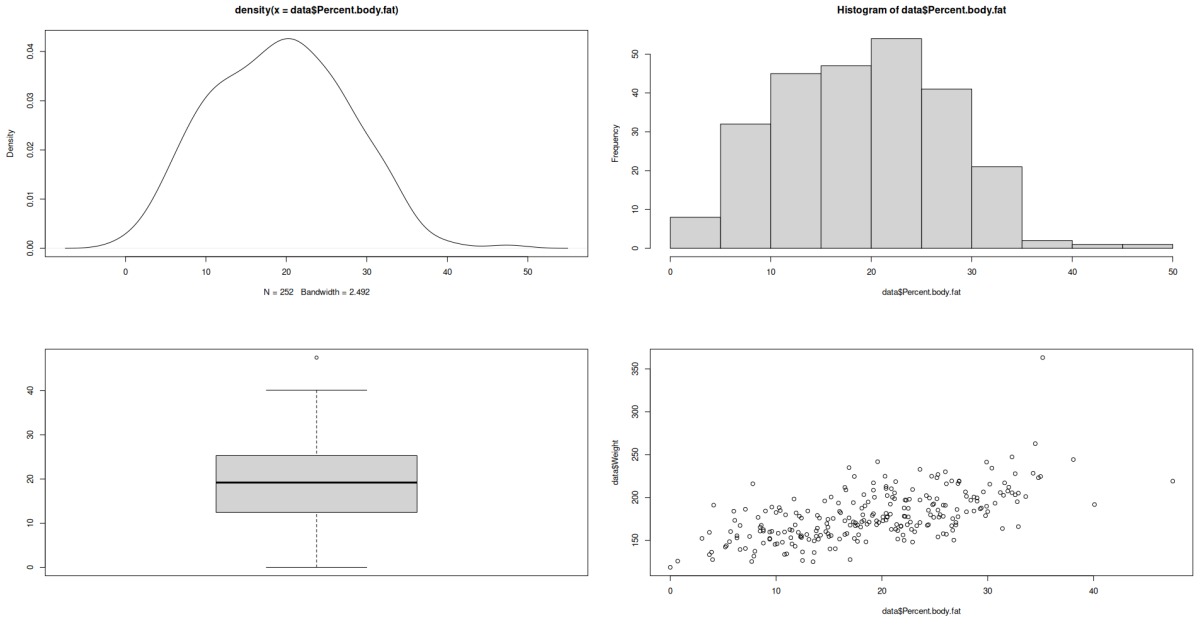
U nastavku su navedene interpretacije pojedine varijable uključujući i BMI.

- **Gustoća (Density)**: Vrijednosti gustoće variraju od 0.995 do 1.109, s prosječnom vrijednošću od 1.056. Većina vrijednosti je koncentrirana oko medijana koji iznosi 1.055.
- **Postotak tjelesne masnoće (Percent body fat)**: Postotak tjelesne masnoće ima širok raspon od 0.00% do 47.50%, s prosjekom od približno 19.15%. Medijan iznosi 19.20%.
- **Dob (Age)**: Dob ispitanika varira od 22 do 81 godine, s prosječnom dobi od približno 44.88 godina. Medijan iznosi 43 godine.
- **Težina (Weight)**: Težina se kreće od 118.5 do 363.1 lbs. Prosjek težine je 178.9 lbs. Medijan težine je 176.5 lbs, što ukazuje na to da su podaci prilično simetrični.
- **Visina (Height)**: Visina ispitanika varira između 29.50 i 77.75 inča, s prosječnom visinom od približno 70.15 inča.
- **Opseg vrata (Neck)**: Opseg vrata kreće se od 31.1 do 51.2 cm, s prosječnom vrijednošću od oko 38.0 cm.
- **Opseg prsa (Chest)**: Opseg prsa varira od 79.3 do 136.2 cm, s prosjekom oko 100.8 cm.
- **Opseg trbuha (Abdomen.2)**: Vrijednosti opsega trbuha se kreću od 69.4 do 148.1 cm, s prosječnom vrijednošću od 92.6 cm. Medijan je približno 91.0 cm.
- **Opseg kukova (Hip)**: Opseg kukova varira od 85.0 do 147.7 cm, s prosječnom vrijednošću od oko 99.9 cm.
- **Opseg bedra (Thigh)**: Vrijednosti opsega bedra kreću se između 47.2 i 87.3 cm, s prosječnom vrijednošću od 59.4 cm.
- **Opseg koljena (Knee)**: Vrijednosti opsega koljena variraju od 33.0 do 49.1 cm, s prosječnom vrijednošću od približno 38.6 cm.
- **Opseg gležnja (Ankle)**: Vrijednosti opsega gležnja kreću se od 19.1 do 33.9 cm, s prosječnom vrijednošću od 23.1 cm.
- **Opseg nadlaktice (Biceps)**: Opseg nadlaktice varira od 24.8 do 45.0 cm, s prosječnom vrijednošću od 32.3 cm.
- **Opseg podlaktice (Forearm)**: Vrijednosti opsega podlaktice kreću se od 21.0 do 34.9 cm, s prosječnom vrijednošću od 28.7 cm.
- **Opseg zapešća (Wrist)**: Vrijednosti opsega zapešća variraju od 15.8 do 21.4 cm, s prosječnom vrijednošću od 18.2 cm.
- **BMI (BMI)**: BMI se kreće od 18.02 do 165.60, što je širok raspon. Prosjek BMI-a je oko 25.94, što se generalno smatra kao normalna vrijednost.

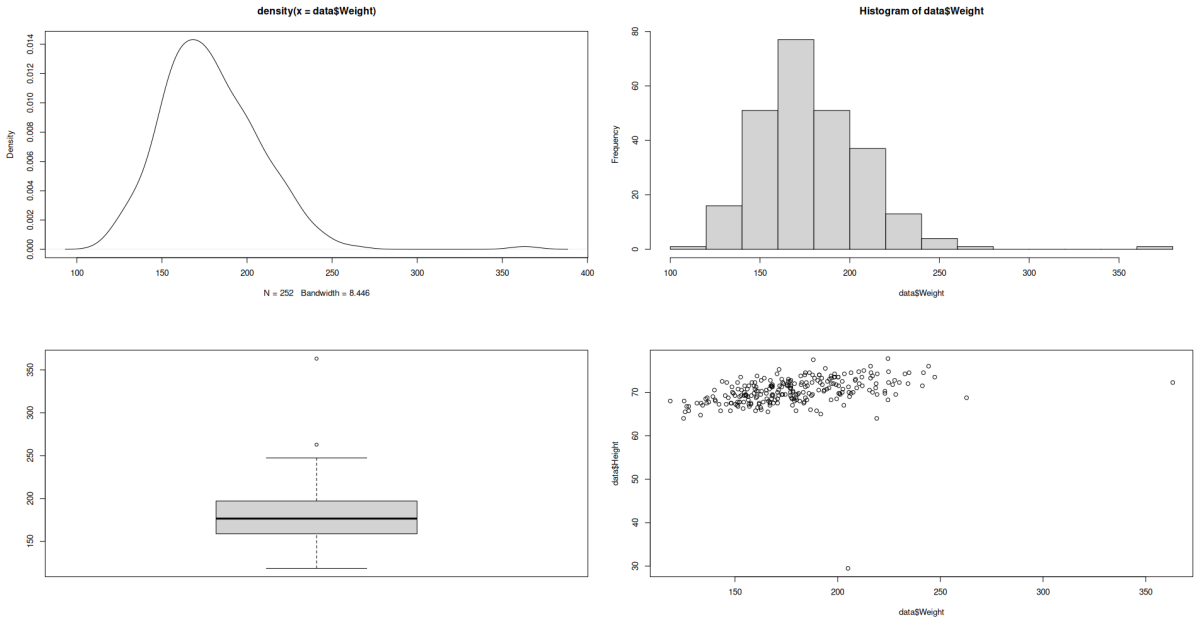
U nastavku su prikazani grafički prikazi navedenih primjera varijabli.



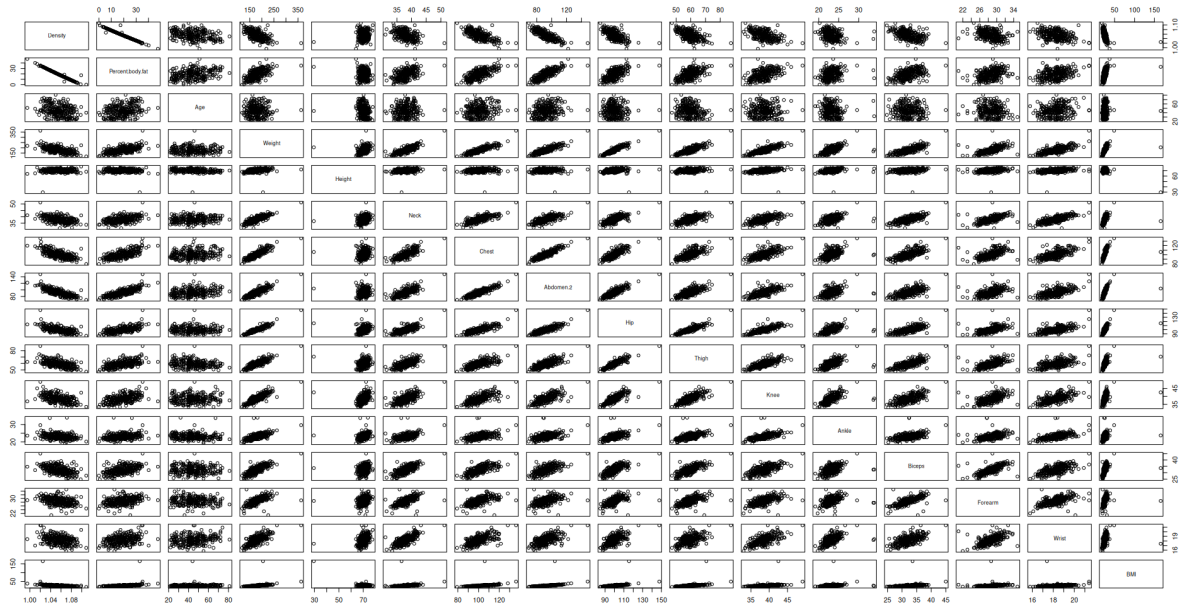
Slika 3: Grafički prikaz varijable `BMI`



Slika 4: Grafički prikaz varijable `Percent.body.fat`



Slika 5: Grafički prikaz varijable `Weight`



Slika 6: Grafički prikaz svih varijabli

2.3. Zadatak C

U zadatku (c), izračunata je matrica korelacija između svih numeričkih varijabli skupa podataka. Za izračun matrice korelacija korištena je funkcija `cor`, a za grafički prikaz funkcija `corrplot`. Rezultati su interpretirani u kontekstu odnosa među varijablama.

Isječak kôda 5: Matrica korelacije

```r
# Izdvajanje samo numeričkih varijabli
numeric_data <- data[sapply(data, is.numeric)]

# c) Izračun matrice korelacija samo za numeričke varijable
cor_matrix <- cor(numeric_data, use = "pairwise.complete.obs")

# c) Prikaz matrice korelacija
print(cor_matrix)

# c) Grafički prikaz matrice korelacija
corrplot(cor_matrix, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```

```
> print(cor_matrix)
                Density Percent.body.fat         Age      Weight      Height       Neck      Chest   Abdomen.2
Density      1.00000000      -0.98778240 -0.27763721 -0.59406188  0.09788114 -0.4729664 -0.6825987 -0.79895463
Percent.body.fat -0.98778240  1.00000000  0.29145844  0.61241400 -0.08949538  0.4905919  0.7026203  0.81343228
Age          -0.27763721       0.29145844  1.00000000 -0.01274609 -0.17164514  0.1135052  0.1764497  0.23040942
Weight       -0.59406188       0.61241400 -0.01274609  1.00000000  0.30827854  0.8307162  0.8941905  0.88799494
Height        0.09788114      -0.08949538 -0.17164514  0.30827854  1.00000000  0.2537099  0.1348918  0.08781291
Neck         -0.47296636       0.49059185  0.11350519  0.83071622  0.25370988  1.0000000  0.7848350  0.75407737
Chest        -0.68259865       0.70262034  0.17644968  0.89419052  0.13489181  0.7848350  1.0000000  0.91582767
Abdomen.2    -0.79895463       0.81343228  0.23040942  0.88799494  0.08781291  0.7540774  0.9158277  1.00000000
Hip          -0.60933143       0.62520092 -0.05033212  0.94088412  0.17039426  0.7349579  0.8294199  0.87406618
Thigh        -0.55309098       0.55960753 -0.20009576  0.86869354  0.14843561  0.6956973  0.7298586  0.76662393
Knee         -0.49504035       0.50866524  0.01751569  0.85316739  0.28605321  0.6724050  0.7194964  0.73717888
Ankle        -0.26489003       0.26596977 -0.10505810  0.61368542  0.26474369  0.4778924  0.4829879  0.45322269
Biceps       -0.48710872       0.49327113 -0.04116212  0.80041593  0.20781557  0.7311459  0.7279075  0.68498272
Forearm      -0.35164842       0.36138690 -0.08505555  0.63030143  0.22864922  0.6236603  0.5801727  0.50331609
Wrist        -0.32571598       0.34657486  0.21353062  0.72977489  0.32206533  0.7448264  0.6601623  0.61983243
BMI          -0.36444320       0.37139475  0.03990184  0.39061954 -0.63798238  0.2662978  0.3833683  0.41494702
                    Hip       Thigh       Knee       Ankle      Biceps     Forearm       Wrist         BMI
Density      -0.60933143 -0.5530910 -0.49504035 -0.2648900 -0.48710872 -0.35164842 -0.3257160 -0.36444320
Percent.body.fat  0.62520092  0.5596075  0.50866524  0.2659698  0.49327113  0.36138690  0.3465749  0.37139475
Age          -0.05033212 -0.2000958  0.01751569 -0.1050581 -0.04116212 -0.08505555  0.2135306  0.03990184
Weight        0.94088412  0.8686935  0.85316739  0.6136854  0.80041593  0.63030143  0.7297749  0.39061954
Height        0.17039426  0.1484356  0.28605321  0.2647437  0.20781557  0.22864922  0.3220653 -0.63798238
Neck          0.73495788  0.6956973  0.67240498  0.4778924  0.73114592  0.62366027  0.7448264  0.26629783
Chest         0.82941992  0.7298586  0.71949640  0.4829879  0.72790748  0.58017273  0.6601623  0.38336835
Abdomen.2     0.87406618  0.7666239  0.73717888  0.4532227  0.68498272  0.50331609  0.6198324  0.41494702
Hip           1.00000000  0.8964098  0.82347262  0.5583868  0.73927252  0.54501412  0.6300895  0.46201247
Thigh         0.89640979  1.0000000  0.79917030  0.5397971  0.76147745  0.56684218  0.5586848  0.43275096
Knee          0.82347262  0.7991703  1.00000000  0.6116082  0.67870883  0.55589819  0.6645073  0.36411126
Ankle         0.55838682  0.5397971  0.61160820  1.0000000  0.48485454  0.41904999  0.5661946  0.21033103
Biceps        0.73927252  0.7614774  0.67870883  0.4848545  1.00000000  0.67825513  0.6321264  0.31090518
Forearm       0.54501412  0.5668422  0.55589819  0.4190500  0.67825513  1.00000000  0.5855883  0.21520047
Wrist         0.63008954  0.5586848  0.66450729  0.5661946  0.63212642  0.58558825  1.0000000  0.19018729
BMI           0.46201247  0.4327510  0.36411126  0.2103310  0.31090518  0.21520047  0.1901873  1.00000000
```

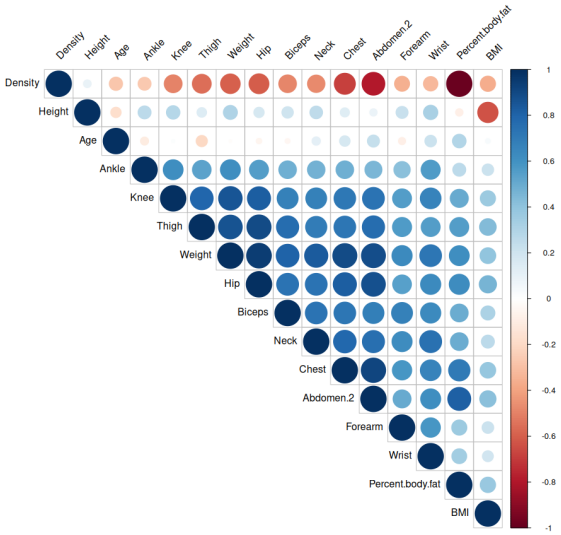
Slika 7: Matrica korelacije

Iz matrice korelacija vidimo nekoliko značajnih odnosa između varijabli:

- Postoji vrlo visoka negativna korelacija (-0.99) između `Density` (gustoće) i `Percent body fat` (postotka tjelesne masnoće), što ukazuje na to da kako gustoća tijela raste, postotak tjelesne masnoće opada, i obrnuto. To je očekivano jer je veća gustoća tijela povezana s manjom količinom masnoće.
- `Weight` (težina) ima visoku pozitivnu korelaciju s mnogim varijablama koje se odnose na tjelesne dimenzije kao što su `Chest` (opseg prsa), `Abdomen.2` (opseg trbuha), `Hip` (opseg kukova), `Thigh` (opseg bedra), itd. Ovo sugerira da kako težina osobe raste, dimenzije tijela također imaju tendenciju povećanja.
- `Height` (visina) ima negativnu korelaciju s `BMI` (-0.64), što ukazuje na to da više osobe imaju tendenciju nižeg BMI-a, a niže osobe imaju tendenciju višeg BMI-a.

- `Age` (dob) ima nisku do umjerenu pozitivnu korelaciju s varijablama kao što su `Percent body fat` i `Neck`. To može ukazivati na to da kako osoba stari, postotak tjelesne masnoće i opseg vrata mogu imati tendenciju povećanja.
- Opseg trbuha (`Abdomen.2`) ima snažnu pozitivnu korelaciju s postotkom tjelesne masnoće (0.81), što znači da osobe s većim opsegom trbuha imaju tendenciju većeg postotka tjelesne masnoće.

U nastavku je prikazan grafički prikaz matrice korelacije.



Slika 8: Grafički prikaz matrice korelacije

Ovi rezultati mogu biti korisni za razumijevanje kako se različite tjelesne mjere odnose jedna na drugu i kako su povezane s tjelesnom masnoćom i gustoćom.

2.4. Zadatak D

Zadatak (d) uključivao je provjeru normalnosti distribucije za sve promatrane varijable pomoću Shapiro-Wilkovog testa.

Isječak kôda 6: Ispitivanje normalnosti

```
# d) Ispitivanje normalnosti samo za numeričke varijable
apply(numeric_data, 2, function(x) shapiro.test(x)$p.value)
```

```
> apply(numeric_data, 2, function(x) shapiro.test(x)$p.value)
       Density Percent.body.fat            Age         Weight         Height 
  6.570746e-01   1.648601e-01   1.043298e-03   1.709433e-08   3.225256e-21 
          Neck          Chest      Abdomen.2            Hip          Thigh 
  4.914854e-05   1.175082e-04   9.831292e-06   3.018896e-10   1.071571e-05 
          Knee          Ankle         Biceps       Forearm          Wrist 
  3.303666e-03   8.782113e-15   4.635486e-02   4.821288e-02   6.377029e-02 
           BMI 
  3.899302e-30 
```

Slika 9: Ispitivanje normalnosti pomoću Shapiro-Wilkovog testa

P-vrijednosti dobivene iz Shapiro-Wilk testa normalnosti sugeriraju sljedeće:

- Varijabla `Density` ima p-vrijednost od 0.66, što je veće od 0.05, ukazujući na to da podaci mogu biti normalno distribuirani.
- Varijabla `Percent body fat` ima p-vrijednost od 0.16, što je također veće od 0.05, sugerirajući normalnu distribuciju.
- Varijabla `Age` ima p-vrijednost od 0.001, što je manje od 0.05, ukazujući na to da podaci vjerojatno nisu normalno distribuirani.
- Većina ostalih varijabli, uključujući `Weight`, `Height`, `Neck`, `Chest`, i druge imaju p-vrijednosti manje od 0.05, što također ukazuje na to da njihovi podaci vjerojatno nisu normalno distribuirani.

Ukratko, većina varijabli u datasetu vjerojatno nije normalno distribuirana, osim `Density` i `Percent body fat`.

2.5. Zadatak E

U zadatku (e), formirana je nova varijabla BMI_faktor, rekodiranjem varijable BMI u kategorije: Pothranjenost, Idealna težina, Prekomjerna tjelesna masa i Pretilost.

Isječak kôda 7: Formiranje nove varijable BMI_faktor

```
# e) Formiranje nove varijable BMI_faktor
data$BMI_faktor <- cut(data$BMI,
                       breaks = c(-Inf, 20, 25, 30, Inf),
                       labels = c("Pothranjenost", "Idealna tezina", "
                        Prekomjerna tjelesna masa", "Pretilost"))
```

U nastavku u prikazu podataka je moguće primijetiti varijablu `BMI_faktor`.

```
> head(data)
  Density Percent.body.fat Age Weight Height Neck Chest Abdomen.2   Hip Thigh Knee Ankle Biceps Forearm Wrist
1  1.0708             12.3  23 154.25  67.75 36.2  93.1      85.2  94.5  59.0 37.3  21.9   32.0    27.4  17.1
2  1.0853              6.1  22 173.25  72.25 38.5  93.6      83.0  98.7  58.7 37.3  23.4   30.5    28.9  18.2
3  1.0414             25.3  22 154.00  66.25 34.0  95.8      87.9  99.2  59.6 38.9  24.0   28.8    25.2  16.6
4  1.0751             10.4  26 184.75  72.25 37.4 101.8      86.4 101.2  60.1 37.3  22.8   32.4    29.4  18.2
5  1.0340             28.7  24 184.25  71.25 34.4  97.3     100.0 101.9  63.2 42.2  24.0   32.2    27.7  17.7
6  1.0502             20.9  24 210.25  74.75 39.0 104.5      94.4 107.8  66.0 42.0  25.6   35.7    30.6  18.8
       BMI               BMI_faktor
1 23.62446           Idealna tezina
2 23.33205           Idealna tezina
3 24.66632           Idealna tezina
4 24.88078           Idealna tezina
5 25.51485 Prekomjerna tjelesna masa
6 26.45263 Prekomjerna tjelesna masa
```

Slika 10: Prikaz podataka i stupca `BMI_faktor` (dolje desno)

2.6. Zadatak F

Zadatak (f) uključivao je primjenu jednofaktorske analize varijance (ANOVA) kako bi se istražilo postoje li razlike u različitim varijablama u ovisnosti o modalitetima varijable BMI_faktor. Ako su ispunjene pretpostavke za provedbu ANOVA-e, korištena je ta metoda, inače je korišten odgovarajući neparametarski test. Kada su pronađene razlike, post hoc test je korišten za ispitivanje između kojih modaliteta postoje razlike.

Isječak kôda 8: Primjena ANOVA-e i neparametarskog testa

```
for (var in variables) {
  # Ispitivanje pretpostavaka za ANOVA
  print(paste("Testing for variable:", var))

  # Ako postoji više od jedne grupe u BMI faktoru za trenutnu varijablu
  if (length(unique(data[!is.na(data[[var]]) & !is.na(data$BMI_faktor),]$BMI_
   faktor)) > 1) {

    # Shapiro-Wilkov test za normalnost
    shapiro_p_value <- shapiro.test(data[[var]])$p.value
    print(paste("Shapiro-Wilk test p-value for", var, ":", shapiro_p_value))

    # Bartlettov test za homogenost varijanci
    bartlett_p_value <- bartlett.test(data[[var]] ~ data$BMI_faktor)$p.value
    print(paste("Bartlett's test p-value for", var, ":", bartlett_p_value))

    # Ako su pretpostavke zadovoljene, provodi se ANOVA
    if (shapiro_p_value > 0.05 && bartlett_p_value > 0.05) {
      # Izvođenje ANOVA
      anova_result <- aov(data[[var]] ~ data$BMI_faktor, data = data)
      print(summary(anova_result))

      # Post-hoc test ako je ANOVA značajna
      if (summary(anova_result)[[1]][["Pr(>F)"]][1] < 0.05) {
        print(paste("ANOVA is significant for", var, ". Performing Tukey HSD
         test."))
        print(TukeyHSD(anova_result))
      }

      # Ako pretpostavke nisu zadovoljene, provodi se Kruskal-Wallisov test
    } else {
      print(paste("Assumptions not met for", var, ". Performing Kruskal-Wallis
       test."))
      print(kruskal.test(data[[var]], data$BMI_faktor))
    }
  }
}
```

Varijabla	Shapiro-Wilk p-vrij.	Bartlett p-vrij.	Test	Rezultat
Percent.body.fat	0.165	0.585	ANOVA	$F(3, 248) = 71.31$, $p < 2 \times 10^{-16}$***
			Tukey HSD	Vidjeti dolje
Neck	4.91×10^{-5}	0.002	Kruskal-Wallis	$\chi^2(3) = 120.71$, $p < 2.2 \times 10^{-16}$
Chest	0.000118	0.012	Kruskal-Wallis	$\chi^2(3) = 177.27$, $p < 2.2 \times 10^{-16}$
Abdomen.2	9.83×10^{-6}	0.000171	Kruskal-Wallis	$\chi^2(3) = 177.02$, $p < 2.2 \times 10^{-16}$
Hip	3.02×10^{-10}	4.09×10^{-10}	Kruskal-Wallis	$\chi^2(3) = 145.92$, $p < 2.2 \times 10^{-16}$
Thigh	1.07×10^{-5}	0.000110	Kruskal-Wallis	$\chi^2(3) = 122.89$, $p < 2.2 \times 10^{-16}$
Knee	0.0033	0.0087	Kruskal-Wallis	$\chi^2(3) = 100.06$, $p < 2.2 \times 10^{-16}$
Ankle	8.78×10^{-15}	0.739	Kruskal-Wallis	$\chi^2(3) = 57.061$, $p = 2.494 \times 10^{-12}$
Biceps	0.046	0.0029	Kruskal-Wallis	$\chi^2(3) = 116.15$, $p < 2.2 \times 10^{-16}$
Forearm	0.048	0.0047	Kruskal-Wallis	$\chi^2(3) = 98.518$, $p < 2.2 \times 10^{-16}$
Wrist	0.064	0.026	Kruskal-Wallis	$\chi^2(3) = 79.59$, $p < 2.2 \times 10^{-16}$

Tablica 2: Rezultati analize podataka s obzirom na kategorije BMI faktora

Tukey HSD rezultati za `Percent.body.fat`:

- Idealna tezina - Pothranjenost: $diff = 6.22$, $p = 0.049$
- Prekomjerna tjelesna masa - Pothranjenost: $diff = 14.45$, $p < 0.001$
- Pretilost - Pothranjenost: $diff = 22.63$, $p < 0.001$
- Prekomjerna tjelesna masa - Idealna tezina: $diff = 8.23$, $p < 0.001$
- Pretilost - Idealna tezina: $diff = 16.41$, $p < 0.001$
- Pretilost - Prekomjerna tjelesna masa: $diff = 8.17$, $p < 0.001$

Rezultati iz tablice 2. prikazuju ispitivanje varijabli u odnosu na kategorije BMI faktora (Pothranjenost, Idealna težina, Prekomjerna tjelesna masa, Pretilost) koristeći ANOVA analizu.

- Za varijablu `Percent.body.fat`, oba testa (Shapiro-Wilk i Bartlett) imaju p-vrijednosti veće od 0.05, što ukazuje na to da su pretpostavke zadovoljene. ANOVA je provedena i pokazala je da postoji značajna razlika u postotku tjelesne masti među različitim kategorijama BMI faktora (p-vrijednost $< 2 \times 10^{-16}$). Tukey HSD post-hoc test je dodatno proveden kako bi se utvrdilo između kojih kategorija BMI faktora postoje značajne razlike.
- Za sve ostale varijable (`Neck`, `Chest`, `Abdomen.2`, `Hip`, `Thigh`, `Knee`, `Ankle`, `Biceps`, `Forearm`, `Wrist`), p-vrijednosti Shapiro-Wilk testa ili Bartlettovog testa su manje od 0.05, što ukazuje na to da pretpostavke nisu zadovoljene. Zbog toga je za te varijable korišten Kruskal-Wallisov test. Sve varijable su pokazale značajne razlike među kategorijama BMI faktora (sve p-vrijednosti su bile iznimno niske).

Ukratko, rezultati ukazuju na to da postoji značajna razlika u mjerenjima među različitim kategorijama BMI faktora za sve ispitivane varijable.

2.7. Zadatak G

Zadatak (g) se odnosio na definiranje modela regresije gdje je zavisna varijabla postotak tjelesne masti, a nezavisne varijable su varijable za koje se smatra da su najviše povezane s postotkom tjelesne masti, na temelju prethodnih analiza. Za procjenu parametara modela korištena je metoda najmanjih kvadrata. Provjerena je adekvatnost modela, procjena reziduala i provjera pretpostavki regresijskog modela. Na temelju rezultata predviđanja, izračunata je greška predviđanja, a rezultati su vizualno prikazani.

Isječak kôda 9: Definiranje modela regresije i provođenje analiza

```
# g) Linearna regresija
data$BMI_faktor <- as.factor(data$BMI_faktor)

# Definiranje modela regresije
model <- lm(`Percent.body.fat` ~ . - BMI, data = data)

# Prikaz rezultata modela regresije
summary(model)

# Izbor modela na temelju AIC
step_model <- step(model, direction = "both")
summary(step_model)

# Prikaz koeficijenata modela
coef(step_model)

# Dijagnostički grafovi
par(mfrow = c(2, 2)) # za prikaz više grafova na jednom prozoru
plot(step_model, which = 1:4) # Prikaz svih dijagnostičkih grafova

# Test normalnosti reziduala
shapiro.test(resid(step_model))

# Test homoskedastičnosti (pogledajte p-vrijednost)
bptest(step_model)

# Test autocorrelation (Durbin-Watson test)
dwtest(step_model)

# Multicollinearity test
vif(step_model) # Variance Inflation Factors
```

Rezultati linearne regresije u tablici 3. ukazuju na to kako varijable utječu na postotak tjelesne masti. Model ima visoku prilagodbu podacima s višestrukim R^2 od 0.9784, što ukazuje na to da objašnjava približno 97.84% varijacije u postotku tjelesne masti. F-statistika je 623.9 s p-vrijednosti manje od 2.2e-16, što ukazuje na statističku značajnost modela.

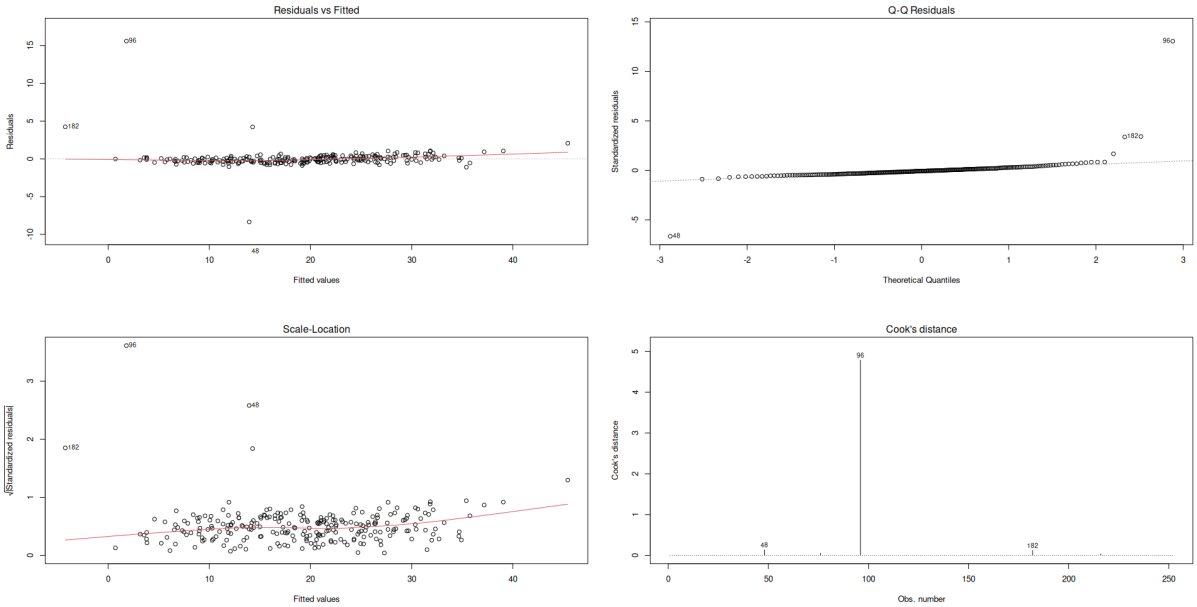
No, kad promotrimo koeficijente, uočavamo da neke varijable nemaju statistički značajan utjecaj na postotak tjelesne masti. Na primjer, varijable kao što su dob, visina, i težina nemaju značajan utjecaj (p-vrijednosti veće od 0.05). Najznačajniji prediktor postotka tjelesne masti je

gustoća (koeficijent = -411.5) koja ima visoku negativnu korelaciju.

Varijabla	**Koeficijent**	**Standardna greška**	**t-vrijednost**	**p-vrijednost**
(Intercept)	450.4	10.73	41.964	$< 2e-16$
Gustoća	-411.5	8.313	-49.500	$< 2e-16$
Dob	0.0144	0.009688	1.486	0.139
Težina	0.00935	0.01604	0.583	0.561
Visina	-0.0117	0.03062	-0.383	0.702
⋮	⋮	⋮	⋮	⋮

Tablica 3: Rezultati linearne regresije

Kroz korak-po-korak odabir modela temeljen na AIC-u (Akaikeovom informacijskom kriteriju), model se pojednostavljuje uklanjajući varijable koje najmanje pridonose. Ovo pomaže u stvaranju manje kompleksnog modela koji bolje generalizira na nove podatke.



Slika 11: Prikaz svih dijagnostičkih grafova

3. Zaključak

U ovom seminarskom radu, fokusirali smo se na detaljnu analizu i istraživanje skupa podataka koji obuhvaća procjene postotka tjelesne masti kod 252 muškaraca [1]. Kroz razne metodologije, uključujući statističku analizu, vizualizaciju podataka, matricu korelacija, ANOVA testove, te regresijske modele, cilj je bio razumjeti kako različite karakteristike tijela utječu na postotak tjelesne masti.

Kao prvi korak, izračunali smo indeks tjelesne mase (BMI) koristeći varijable težine i visine, što je rezultiralo novom varijablom koja se dodaje u skup podataka. Zatim smo opisali varijable skupa podataka, uključujući BMI, i vizualizirali ih putem raznih grafova.

Matrica korelacija, izračunata korištenjem funkcije `cor`, pokazala je kako su različite numeričke varijable povezane. To nam je omogućilo uvid u odnose između varijabli, kao što su visina, težina, dob, i postotak tjelesne masti.

Nakon toga, proveli smo Shapiro-Wilkov test kako bismo provjerili normalnost distribucije promatranih varijabli. Ovo je bitno jer neke statističke tehnike zahtijevaju normalnu distribuciju podataka.

Također, implementirali smo analizu varijance (ANOVA) kako bismo istražili mogu li se razlike u postotku tjelesne masti pripisati različitim kategorijama BMI-a. U slučajevima gdje pretpostavke za ANOVA nisu ispunjene, korišteni su odgovarajući neparametarski testovi. Post hoc analiza provedena je za uočavanje modaliteta gdje postoje značajne razlike.

Nadalje, konstruirali smo model regresije gdje je postotak tjelesne masti zavisna varijabla. Korištenjem metode najmanjih kvadrata, ocijenili smo parametre modela. Utvrdili smo adekvatnost modela, procijenili reziduale, i proveli provjeru pretpostavki regresijskog modela. Ovaj model regresije pokazao se učinkovitim u predviđanju postotka tjelesne masti na temelju mjerenja tijela.

Kroz ovu sveobuhvatnu analizu, dobili smo dubinski uvid u kako karakteristike tijela, poput težine, visine, i opsega, utječu na postotak tjelesne masti. Rezultati su posebno relevantni za stručnjake u području tjelesnog zdravlja, trenere, i pojedince zainteresirane za razumijevanje tjelesne kompozicije.

Kao preporuku za buduća istraživanja, bilo bi korisno proširiti skup podataka dodatnim varijablama kao što su prehrana, tjelesna aktivnost, i genetski faktori. Osim toga, istraživanje bi trebalo uključivati i žensku populaciju, s obzirom na razlike u distribuciji tjelesne masti između spolova. Kroz analizu u ovom radu, smatramo da su postavljeni lijepi temelji za daljnje istraživanje, nudeći metodologije i uvide koji mogu pridonijeti primjerice boljem razumijevanju i upravljanju tjelesnom kompozicijom i zdravljem.

Popis literature

[1] R. W. Johnson. „Lists estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men," Department of Mathematics & Computer Science, South Dakota School of Mines & Technology. (2023.), adresa: `http://lib.stat.cmu.edu/datasets/bodyfat` (pogledano 22. 5. 2023.).

Popis slika

Popis tablica

Popis isječaka koda

Prilozi

1. obrada.py

Kod napravljen u programskom jeziku Python za svrhu "čišćenja" podataka, kako bi se dobio ispravan skup podataka za analizu. Poveznica na repozitorij sustava za verzioniranje na kojem se nalazi izvorni kod. (Potrebno je kliknuti na ovaj tekst.)

```
import pandas as pd

# Broj redaka koje treba preskočiti
skip_rows = 117

# Ukupan broj redaka u datoteci
total_lines = 381

# Izračunaj broj redaka koje treba preskočiti na kraju
end_skip = total_lines - 370

# Definiraj naslove stupaca
headers = [
    "Density",
    "Percent body fat",
    "Age",
    "Weight",
    "Height",
    "Neck",
    "Chest",
    "Abdomen 2",
    "Hip",
    "Thigh",
    "Knee",
    "Ankle",
    "Biceps",
    "Forearm",
    "Wrist"
]

# Učitaj .txt datoteku preskačući zadane retke i dodaj naslove stupaca
df = pd.read_csv('bodyfat.txt', delimiter="\s+", skiprows=skip_rows, skipfooter=
 end_skip, names=headers, engine='python')

# Pretvori u .csv datoteku koristeći ";" kao separator
df.to_csv('data.csv', sep=";", index=False)

# Prikazuje prvih 5 redaka
print(df.head())
```

2. analiza.R

Kod napravljen u programskom jeziku R za svrhu analize podataka. Poveznica na repozitorij sustava za verzioniranje na kojem se nalazi izvorni kod. (Potrebno je kliknuti na ovaj tekst.)

```
1  if (!require("tidyverse")) install.packages("tidyverse")
2  if (!require("corrplot")) install.packages("corrplot")
3  if (!require("lmtest")) install.packages("lmtest")
4  if (!require("car")) install.packages("car")
5  library(tidyverse)
6  library(corrplot)
7  library(lmtest)
8  library(car)
9
10 # Učitavanje skupa podataka
11 data <- read.csv('data.csv', sep = ";", dec = ".", header = TRUE)
12
13 # Provjera strukture podataka
14 str(data)
15 head(data)
16
17 # Zamjena NA vrijednosti srednjim vrijednostima
18 data$`Weight`[is.na(data$`Weight`)] <- mean(data$`Weight`, na.rm = TRUE)
19 data$`Height`[is.na(data$`Height`)] <- mean(data$`Height`, na.rm = TRUE)
20
21 # a) Izračun BMI
22 data$BMI <- (data$`Weight` / (data$`Height`)^2) * 703
23
24 # b) Opisivanje varijabli
25 summary(data)
26
27 # Odabrati proizvoljnu varijablu za prikaz
28 plot(density(data$Weight))
29 hist(data$Weight)
30 boxplot(data$Weight)
31 plot(data$Weight, data$Height) # Usporedba dvaju varijabli
32
33 # b) Grafički prikaz
34 pairs(data)
35
36 # Izdvajanje samo numeričkih varijabli
37 numeric_data <- data[sapply(data, is.numeric)]
38
39 # c) Izračun matrice korelacija samo za numeričke varijable
40 cor_matrix <- cor(numeric_data, use = "pairwise.complete.obs")
41
42 # c) Prikaz matrice korelacija
43 print(cor_matrix)
44
45 # c) Grafički prikaz matrice korelacija
46 corrplot(cor_matrix, type = "upper", order = "hclust",
47          tl.col = "black", tl.srt = 45)
```

```r
# d) Ispitivanje normalnosti samo za numeričke varijable
apply(numeric_data, 2, function(x) shapiro.test(x)$p.value)

# e) Formiranje nove varijable BMI_faktor
data$BMI_faktor <- cut(data$BMI,
                       breaks = c(-Inf, 20, 25, 30, Inf),
                       labels = c("Pothranjenost", "Idealna tezina", "
                        Prekomjerna tjelesna masa", "Pretilost"))

# f) Lista varijabli za provođenje ANOVA
variables <- c("Percent.body.fat", "Neck", "Chest", "Abdomen.2", "Hip", "Thigh",
  "Knee", "Ankle", "Biceps", "Forearm", "Wrist")

# f) Provjera distribucije BMI faktora
table(data$BMI_faktor)

for (var in variables) {
  # Ispitivanje pretpostavaka za ANOVA
  print(paste("Testing for variable:", var))

  # Ako postoji više od jedne grupe u BMI faktoru za trenutnu varijablu
  if (length(unique(data[!is.na(data[[var]]) & !is.na(data$BMI_faktor),]$BMI_
   faktor)) > 1) {

    # Shapiro-Wilkov test za normalnost
    shapiro_p_value <- shapiro.test(data[[var]])$p.value
    print(paste("Shapiro-Wilk test p-value for", var, ":", shapiro_p_value))

    # Bartlettov test za homogenost varijanci
    bartlett_p_value <- bartlett.test(data[[var]] ~ data$BMI_faktor)$p.value
    print(paste("Bartlett's test p-value for", var, ":", bartlett_p_value))

    # Ako su pretpostavke zadovoljene, provodi se ANOVA
    if (shapiro_p_value > 0.05 && bartlett_p_value > 0.05) {
      # Izvođenje ANOVA
      anova_result <- aov(data[[var]] ~ data$BMI_faktor, data = data)
      print(summary(anova_result))

      # Post-hoc test ako je ANOVA značajna
      if (summary(anova_result)[[1]][["Pr(>F)"]][1] < 0.05) {
        print(paste("ANOVA is significant for", var, ". Performing Tukey HSD
         test."))
        print(TukeyHSD(anova_result))
      }

      # Ako pretpostavke nisu zadovoljene, provodi se Kruskal-Wallisov test
    } else {
      print(paste("Assumptions not met for", var, ". Performing Kruskal-Wallis
       test."))
      print(kruskal.test(data[[var]], data$BMI_faktor))
    }
  }
}
```

```r
# g) Linearna regresija
data$BMI_faktor <- as.factor(data$BMI_faktor)

# Definiranje modela regresije
model <- lm(`Percent.body.fat` ~ . - BMI, data = data)

# Prikaz rezultata modela regresije
summary(model)

# Izbor modela na temelju AIC
step_model <- step(model, direction = "both")
summary(step_model)

# Prikaz koeficijenata modela
coef(step_model)

# Dijagnostički grafovi
par(mfrow = c(2, 2)) # za prikaz više grafova na jednom prozoru
plot(step_model, which = 1:4) # Prikaz svih dijagnostičkih grafova

# Test normalnosti reziduala
shapiro.test(resid(step_model))

# Test homoskedastičnosti (pogledajte p-vrijednost)
bptest(step_model)

# Test autocorrelation (Durbin-Watson test)
dwtest(step_model)

# Multicollinearity test
vif(step_model) # Variance Inflation Factors
```