Navodila za uporabo programskega okolja R

Bivariatna analiza

KAZALO VSEBINE

1 Bivariatna analiza 3
2 Proučevanje povezanosti 3
2.1 Povezanost za nominalni tip para spremenljivk 3
2.2 Povezanost za ordinalni tip para spremenljivk 4
2.3 Povezanost za intervalni / razmernostni tip para spremenljivk 5
3 Proučevanje povezanosti v R 6
3.1 Kontingenčna tabela 6
3.2 Hi-kvadrat test 10
3.3 Kontingenčni koeficienti 11
3.4 Spearmanov korelacijski koeficient rangov 11
3.5 Pearsonov korelacijski koeficient 12
4 Proučevanje odvisnosti 12
5 Proučevanje odvisnosti v R 14

Kazalo tabel

Tabela 1: Primer kontingenčne tabele (Območje bivanja glede na spol anketirancev) 6

Kazalo izpisov

Izpis 1: Prikaz kontingenčne tabele (Spol - Območje bivanja) 6
Izpis 2: Robne frekvence kontingenčne tabele s pomočjo funkcije `margin.table()` 7
Izpis 3: Kontingenčna tabela, dopolnjena z marginalnimi frekvencami 7
Izpis 4: Izračun deležev v kontingenčni tabeli s pomočjo funkcije `prop.table()` 8
Izpis 5: Kontingenčne tabele z deleži, dopolnjene z marginalnimi deleži 8
Izpis 6: Kontingenčna tabela s pomočjo funkcije `CrossTable()` (paket `gmodels`) 9
Izpis 7: Hi-kvadrat test povezanosti med spremenljivkama Spol in Območje bivanja s pomočjo funkcije `chisq.test()` 10
Izpis 9: Spearmanov korelacijski koeficient rangov za spremenljivki Izobrazba in Število let šolanja 11

Izpis 11: Pearsonov korelacijski koeficient za spremenljivki Starost in Število let šolanja.....12
Izpis 10: Rezultat bivariatne linearne regresije (Zadovoljstvo~Starost).....16
Izpis 11: Podroben rezultat izvedene bivariatne regresijske analize (Zadovoljstvo~Starost).....16

1 BIVARIATNA ANALIZA

Kadar proučujemo primere, ko na vsaki enoti gledamo po dve spremenljivki hkrati, pravimo, da izvajamo bivariatno analizo. Pri tem se osredotočamo na proučevanje relacij med spremenljivkami.

Poznamo dve pomembni relaciji:

- **povezanost**: $X \longleftrightarrow Y$
    - razumemo kot relacijo, ko se vrednosti obeh spremenljivk spreminjajo hkrati;
    - obe spremenljivki sta v tem primeru enakovredni;
    - namen študija povezanosti je izračunati ustrezno mero, ki vrednoti jakost povezanosti dveh spremenljivk.

- **odvisnost:** $X \longrightarrow Y$
    - razumemo kot relacijo, kjer vrednosti ene spremenljivke vplivajo na vrednosti druge spremenljivko, vpliva v drugo smer pa ni;
    - ena spremenljivka je torej odvisna od druge;
    - namen študija odvisnosti je pridobiti nova spoznanja o odvisnosti ter napovedovanje vrednosti odvisne spremenljivke pri izbrani vrednosti neodvisne spremenljivke.

2 PROUČEVANJE POVEZANOSTI

Mere povezanosti ločimo glede na tip spremenljivk:

- NOMINALNI tip para spremenljivk:
    - ena od spremenljivk je nominalna: χ^2 test
- ORDINALNI tip para spremenljivk:
    - ena spremenljivka je ordinalna, druga pa ordinalna ali boljša**:** Spearmanov koeficient korelacije rangov
- RAZMERNOSTNI/INTERVALNI tip para spremenljivk:
    - obe spremenljivki sta številski: Pearsonov koeficient korelacije

2.1 Povezanost za nominalni tip para spremenljivk

V primerih, kadar proučujemo povezanost med dvema spremenljivkama, pri čemer je vsaj ena od spremenljivk nominalna (opisna), kot mero povezanosti računamo **χ^2 test**.

Povezanost preverimo po naslednjem postopku:

1. Podatke uredimo v kontingenčno tabelo.
2. Izračunamo tabelo pričakovanih frekvenc.

3. Izračunamo hi-kvadrat test: normirano razliko med ugotovljenimi (dejanskimi) frekvencami f_{ij} in pričakovanimi frekvencami f_{ij}'.

χ^2-test sloni na primerjavi empiričnih (dejanskih) frekvenc s pričakovanimi frekvencami. V tem primeru so pričakovane frekvence tiste frekvence, ki bi bile v kontingenčni tabeli, če spremenljivki ne bi bili medsebojno povezani. To pa pomeni, da bi bili porazdelitvi ene spremenljivke glede na drugo *enaki*. Če spremenljivki nista povezani, so verjetnosti hkratne zgoditve posameznih vrednosti prve in druge spremenljivke enake produktu verjetnosti posameznih vrednosti. Vrednost χ^2-testa je vedno **pozitivna**.

2.2 Povezanost za ordinalni tip para spremenljivk

V primerih, kadar proučujemo povezanost med dvema spremenljivkama, pri čemer je vsaj ena od spremenljivk ordinalna, kot mero povezanosti računamo **Spearmanov korelacijski koeficient rangov**.

Povezanost preverimo po naslednjem postopku:

1. Podatke uredimo v tabelo in jih rangiramo glede na na prvo in nato še glede na drugo spremenljivko. V tabeli imamo torej:
   - v prvi stolpec vpišemo enote,
   - v drugi stolpec zaporedno mesto enote v ranžirni vrsti glede na prvo obravnavano lastnost (spremenljivko),
   - v tretji stolpec pa zaporedno mesto enote v ranžirni vrsti glede na drugo obravnavano lastnost (spremenljivko).
2. Izračunamo Spearmanov korelacijski koeficient rangov: meri povezanost med rangi obravnavanih spremenljivk.

Spearmanov koeficient korelacije rangov lahko zavzame vrednosti v intervalu [-1,1].

Pomen predznaka vrednosti:

- Pozitivna linearna povezanost: $\rho_s > 0$

  Z večanjem rangov po prvi spremenljivki se večajo tudi rangi po drugi spremenljivki.

- Negativna linearna povezanost: $\rho_s < 0$

  Z večanjem rangov po prvi spremenljivki se manjšajo rangi po drugi spremenljivki.

- Ni povezanosti med spremenljivkama: $\rho_s = 0$

Same vrednosti pa interpretiramo na naslednji način:

Absolutna vrednost $\|\rho_S\|$ ali $\|r_s\|$	**Pomen**
od 0.5 do 1	močna povezanost
od 0.3 do 0.5	srednje močna povezanost
od 0.1 do 0.3	šibka povezanost
od 0 do 0.1	ni povezanosti

2.3 Povezanost za intervalni / razmernostni tip para spremenljivk

V primerih, kadar proučujemo povezanost med dvema spremenljivkama, pri čemer sta obe spremenljivki intervalni ali razmernostni, povezanost ugotavljamo s pomočjo kovariance med proučevanima spremenljivkama. Kovarianco nato standardiziramo, s čimer torej kot mero povezanosti računamo **Pearsonov korelacijski koeficient**.

Pearsonov korelacijski koeficient lahko zavzame vrednosti v intervalu [-1,1].

Pomen predznaka vrednosti:

- Pozitivna linearna povezanost: $\rho > 0$

  Z večanjem vrednosti prve spremenljivke se večajo vrednosti tudi druge spremenljivke.

- Negativna linearna povezanost: $\rho < 0$

  Z večanjem vrednosti prve spremenljivke se vrednosti druge spremenljivke manjšajo.

- Ni linearne povezanosti: $\rho = 0$

Same vrednosti pa interpretiramo na naslednji način:

Absolutna vrednost $\|\rho\|$ ali $\|r\|$	**Pomen**
od 0.5 do 1	močna povezanost
od 0.3 do 0.5	srednje močna povezanost
od 0.1 do 0.3	šibka povezanost
od 0 do 0.1	ni povezanosti

3 PROUČEVANJE POVEZANOSTI V R

3.1 Kontingenčna tabela

Za proučevanje povezanosti za nominalni tip para spremenljivk izhajamo iz kontingenčne tabele. Kontingenčna tabela je pravzaprav dvorazsežna frekvenčna tabela, ki prikazuje seštevek enot glede na vrednosti opazovanih spremenljivk. Primer kontingenčne tabele za naše podatke prikazuje Tabela 1.

Tabela 1: Primer kontingenčne tabele (Območje bivanja glede na spol anketirancev)

	Veliko mesto	Primestje	Manjše mesto	Vas	Kmetija
Moški	43	128	127	249	58
Ženski	70	114	146	382	56

Kontingenčno tabelo v R pripravimo s pomočjo funkcije `table()`, ki ima naslednjo obliko:

```
table(A,B)
```

Pri tem je: `A` spremenljivka, katere vrednosti bodo prikazane v vrsticah kontingenčne tabele

`B` spremenljivka, katere vrednosti bodo prikazane v stolpcih kontingenčne tabele

Na naših podatkih torej kontingenčno tabelo pripravimo kot:

```
table(podatki2$Spol,podatki2$Obmocje)
```

Izpis 1: Prikaz kontingenčne tabele (Spol - Območje bivanja)

```
        veliko mesto primestje majno mesto vas kmetija
 moski            43       128         127 279      58
 zenski           70       114         146 382      56
```

Če tabelo shranimo v nek objekt (denimo `tabela`), lahko izpišemo tudi marginalne frekvence s pomočjo funkcije `margin.table()` ter deleže s pomočjo funkcije `prop.table()`:

```
tabela<-table(podatki2$Spol,podatki2$Obmocje)
```

Skupno število enot, prikazanih v kontingenčni tabeli, dobimo s pomočjo ukaza:

```
margin.table(tabela)
```

Vsoto po vrsticah kontingenčne tabele dobimo s pomočjo ukaza:

```
margin.table(tabela,1)
```

Vsoto po stolpcih kontingenčne tabele dobimo s pomočjo ukaza:

```
margin.table(tabela,2)
```

Izpis 2: Robne frekvence kontingenčne tabele s pomočjo funkcije `margin.table()`

```
> margin.table(tabela)
[1] 1403

> margin.table(tabela,1)

 moski zenski
   635    768

> margin.table(tabela,2)

veliko mesto    primestje  majno mesto          vas      kmetija
         113          242          273          661          114
```

Marginalne frekvence lahko dodamo kontingenčni tabeli s pomočjo funkcije `addmargins()`:

```
addmargins(tabela)
```

Izpis 3: Kontingenčna tabela, dopolnjena z marginalnimi frekvencami

```
         veliko mesto primestje majno mesto  vas kmetija  Sum
  moski            43       128         127  279      58  635
  zenski           70       114         146  382      56  768
  Sum             113       242         273  661     114 1403
```

V kontingenčni tabeli frekvence prikažemo kot deleže s pomočjo ukaza:

```
prop.table(tabela)
```

Deleže po vrsticah kontingenčne tabele izračunamo s pomočjo ukaza:

```
prop.table(tabela,1)
```

Deleže po stolpcih kontingenčne tabele izračunamo s pomočjo ukaza:

```
prop.table(tabela,2)
```

Izpis 4: Izračun deležev v kontingenčni tabeli s pomočjo funkcije `prop.table()`

```
> prop.table(tabela)

         veliko mesto  primestje majno mesto        vas    kmetija
  moski    0.03064861 0.09123307  0.09052031 0.19885959 0.04133999
  zenski   0.04989309 0.08125445  0.10406272 0.27227370 0.03991447
> prop.table(tabela,1)

         veliko mesto  primestje majno mesto        vas    kmetija
  moski    0.06771654 0.20157480  0.20000000 0.43937008 0.09133858
  zenski   0.09114583 0.14843750  0.19010417 0.49739583 0.07291667
> prop.table(tabela,2)

         veliko mesto primestje majno mesto       vas   kmetija
  moski     0.3805310 0.5289256   0.4652015 0.4220877 0.5087719
  zenski    0.6194690 0.4710744   0.5347985 0.5779123 0.4912281
```

Tudi v kontingenčno tabelo s prikazanimi deleži lahko dodamo marginalne deleže (vsote deležev) s pomočjo funkcije `addmargins()`. Ker pa je v funkciji privzeto, da izračuna vsote po stolpcih in vrsticah tabele za vse spremenljivke, moramo v primeru, da želimo izračunati deleže le glede na vsoto ene spremenljivke, podati tudi dodatne indekse:

Skupen delež enot, prikazanih v kontingenčni tabeli, dobimo s pomočjo ukaza:

```
addmargins(prop.table(tabela))
```

Vsoto deležev glede na vrstično spremenljivko kontingenčni tabeli dobimo s pomočjo ukaza:

```
addmargins(prop.table(tabela,1),2)
```

Vsoto deležev glede na stolpčno spremenljivko v kontingenčni tabeli dobimo s pomočjo ukaza:

```
addmargins(prop.table(tabela,2),1)
```

Izpis 5: Kontingenčne tabele z deleži, dopolnjene z marginalnimi deleži

```
> addmargins(prop.table(tabela))

         veliko mesto  primestje majno mesto        vas    kmetija        Sum
  moski    0.03064861 0.09123307  0.09052031 0.19885959 0.04133999 0.45260157
  zenski   0.04989309 0.08125445  0.10406272 0.27227370 0.03991447 0.54739843
  Sum      0.08054170 0.17248753  0.19458304 0.47113329 0.08125445 1.00000000
> addmargins(prop.table(tabela,1),2)

         veliko mesto  primestje majno mesto        vas    kmetija        Sum
  moski    0.06771654 0.20157480  0.20000000 0.43937008 0.09133858 1.00000000
  zenski   0.09114583 0.14843750  0.19010417 0.49739583 0.07291667 1.00000000
> addmargins(prop.table(tabela,2),1)
```

```
         veliko mesto primestje majno mesto       vas   kmetija
  moski     0.3805310 0.5289256   0.4652015 0.4220877 0.5087719
  zenski    0.6194690 0.4710744   0.5347985 0.5779123 0.4912281
  Sum       1.0000000 1.0000000   1.0000000 1.0000000 1.0000000
```

Kontingenčne tabele pa lahko pripravimo tudi s pomočjo funkcije `CrossTable()`, ki jo najdemo v paketu `gmodels` in ima naslednjo obliko:

```
CrossTable(A,B)
```

Pri tem je: `A` spremenljivka, katere vrednosti bodo prikazane v vrsticah kontingenčne tabele

`B` spremenljivka, katere vrednosti bodo prikazane v stolpcih kontingenčne tabele

Najprej prenesemo knjižnico v R:

```
install.packages('gmodels')
```

```
load(gmodels)
```

In uporabimo funkcijo:

```
CrossTable(podatki2$Spol,podatki2$Obmocje)
```

Izpis 6: Kontingenčna tabela s pomočjo funkcije `CrossTable()` *(paket* `gmodels`*)*

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  1403


             | podatki2$Obmocje
podatki2$Spol | veliko mesto |    primestje |  majno mesto |          vas |      kmetija |    Row Total |
--------------|--------------|--------------|--------------|--------------|--------------|--------------|
        moski |           43 |          128 |          127 |          279 |           58 |          635 |
              |        1.297 |        3.115 |        0.096 |        1.360 |        0.795 |              |
              |        0.068 |        0.202 |        0.200 |        0.439 |        0.091 |        0.453 |
              |        0.381 |        0.529 |        0.465 |        0.422 |        0.509 |              |
              |        0.031 |        0.091 |        0.091 |        0.199 |        0.041 |              |
--------------|--------------|--------------|--------------|--------------|--------------|--------------|
       zenski |           70 |          114 |          146 |          382 |           56 |          768 |
              |        1.072 |        2.575 |        0.079 |        1.124 |        0.657 |              |
              |        0.091 |        0.148 |        0.190 |        0.497 |        0.073 |        0.547 |
              |        0.619 |        0.471 |        0.535 |        0.578 |        0.491 |              |
              |        0.050 |        0.081 |        0.104 |        0.272 |        0.040 |              |
--------------|--------------|--------------|--------------|--------------|--------------|--------------|
 Column Total |          113 |          242 |          273 |          661 |          114 |         1403 |
              |        0.081 |        0.172 |        0.195 |        0.471 |        0.081 |              |
--------------|--------------|--------------|--------------|--------------|--------------|--------------|
```

Funkcija `CrossTable()` ima tudi druge opcije: izpiše lahko odstotke (po vrsticah, stolpcih, skupno), določimo lahko tudi število decimalnih mest, izračunamo hi-kvadrat, Fisherjev in McNemarjev test povezanosti, prikažemo pričakovane frekvence. To bomo spoznali kasneje.

3.2 Hi-kvadrat test

Za nomialni tip para spremenljivk povezanost proučujemo s pomočjo hi-kvadrat testa. V R hi-kvadrat test izvedemo s pomočjo funkcije `chisq.test()`, ki ima naslednjo splošno obliko:

```
chisq.test(table)
```

Pri tem je: `table` dvorazsežna kontingenčna tabela na spremenljivkah, med katerima proučujemo povezanost.

Za naše podatke izvedemo hi-kvadrat test:

```
chisq.test(tabela)
```

Izpis 7: Hi-kvadrat test povezanosti med spremenljivkama Spol in Območje bivanja s pomočjo funkcije `chisq.test()`

```
        Pearson's Chi-squared test

data:  tabela
X-squared = 12.17, df = 4, p-value = 0.01613
```

Rezultat hi-kvadrat testa lahko tudi shranimo v *objekt*:

```
hikvadrat<-chisq.test(tabela)
```

S pomočjo objekta `hikvadrat` lahko sedaj dostopamo do:

- tabele opazovanih frekvenc:

  ```
  hikvadrat$observed
  ```

- tabele teoretičnih frekvenc:

  ```
  hikvadrat$expected
  ```

3.3 Kontingenčni koeficienti

Ker nam sama vrednost hi-kvadrat testa ne poda informacije o intenzivnosti povezave med opazovanima spremenljivkama, za te namene računamo koeficiente kontingence.

V R za izračun kontingenčnih koeficientov uporabimo paket `vcd`. Paket najprej prenesemo na računalnik:

```
install.packages('vcd')
```

in naložimo v delovno okolje:

```
library(vcd)
```

Za izračun kontingenčnih koeficientov uporabimo funkcijo `assocstats()`, ki izpiše vrednost hi-kvadrat testa, ter poda vrednosti koeficientov Phi, Cramerjev V in C.

```
assocstats(tabela)
```

3.4 Spearmanov korelacijski koeficient rangov

Povezanost za ordinalni tip para spremenljivk proučujemo s pomočjo Spearmanovega korelacijskega koeficienta rangov. V R ga lahko izačunamo s pomočjo funkcije `cor()`, ki ima naslednjo obliko:

```
cor(x, use= , method= )
```

Pri tem je: `x` podatkovni okvir

`use` opredeljuje, na kakšen način naj bodo obravnavani manjkajoči podatki (uporabljali bomo pairwise complete observation)

`method` pa določa tip korelacije, ki jo želimo izračunati (Spearmanov, Pearsonov ali Kendallov korelacijski koeficient)

V naših podatkih bomo izračunali Spearmanov korelacijski koeficient rangov med spremenljivkama Izobrazba (ordinalna spremenljivka) ter Število let šolanja.

```
cor(podatki2$Izob,podatki2$Sol,use="pairwise.complete.obs",method="spearman")
```

Izpis 8: Spearmanov korelacijski koeficient rangov za spremenljivki Izobrazba in Število let šolanja

```
[1] 0.855085
```

3.5 Pearsonov korelacijski koeficient

Za proučevanje povezanosti za intervalni/razmernostni tip para spremenljivk pa izračunamo Pearsonov korelacijski koeficient. Postopek v R je enak kot v primeru izračuna Spearmanovega korelacijskega koeficienta rangov, le da v funkcijah opredelimo, da računamo Pearsonov korelacijski koeficient.

Na naših podatkih preverimo, ali obstaja povezanost med Starostjo in Številom let šolanja anketirancev:

```
cor(podatki2$Starost,podatki2$Sol,use="pairwise.complete.obs",method
="pearson")
```

Izpis 9: Pearsonov korelacijski koeficient za spremenljivki Starost in Število let šolanja

```
[1] -0.2260437
```

4 PROUČEVANJE ODVISNOSTI

Odvisnost med parom intervalnimi/razmernostnimi spremenljivkami proučujemo s pomočjo bivariatne linearne regresije.

V primeru bivariatne linearne regresije proučujemo odvisnost med eno odvisno in eno neodvisno spremenljivko:

$$X \longrightarrow Y$$

Odnos med proučevanima spremenljivkama najlažje predstavimo v razsevnem grafikonu. S pomočjo bivariatne regresijske analize pa želimo opisati odvisnost med spremenljivkama s pomočjo linearne premice. Problem, ki ga rešujemo s pomočjo bivariatne linearne regresije je tore, kako med točke v razsevnem grafikonu vrisati *linearno premico*, ki bo najbolje ponazarjala odvisnost med spremenljivkama?

Regresijsko premico lahko določimo na dva načina:

- **grafično:** regresijska premica je med točke v razsevnem grafikonu postavljena tako, da so odkloni točk od premice čim manjši. Na ta način najbolje ponazarja odvisnost med spremenljivkama;
- **matematično:** regresijska premica je določena z metodo najmanjših kvadratov, pri kateri iščemo takšne vrednosti napovedane odvisne spremenljivke, da bodo kvadrati odklonov pravih vrednosti od teh napovedanih vrednosti čim manjši.

Odvisnost spremenljivke Y od spremenljivke X s pomočjo linearne funkcije regresijske premice zapišemo kot:

$$Y' = \alpha + \beta X$$

Odnos med *X* in *Y* je torej odvisen od dveh parametrov:

- parametra $\boldsymbol{\alpha}$, ki določa, kje regresijska premica seka ordinato, $a = Y'(0)$, (torej Y' ima vrednost a, ko ima *X* vrednost 0);
- parametra $\boldsymbol{\beta}$, ki določa naklon premice (pozitivna ali negativna povezanost in moč povezanosti).

Parameter $\boldsymbol{\beta}$ imenujemo tudi regresijski koeficient. Pove, za koliko se spremeni vrednost *Y*, če se *X* spremeni za eno enoto. V primeru, da je $\beta = 0$, potem *Y* ni odvisna od *X*.

Pogosto nas zanima, kako dober je izračunani regresijski model (regresijska funkcija). Kvaliteto regresijskega modela lahko ocenimo z 2 kazalcema:

- **determinacijskim koeficientom** (delež pojasnjene variance), ki je kazalec kvalitete opisa odvisnosti med spremenljivkama z regresijsko premico in
- **standardno napako ocene**, ki je kazalec kvalitete napovedovanja vrednosti odvisne spremenljivke s pomočjo regresijske premice.

Determinacijski koeficient predstavlja delež pojasnjene variance spremenljivke *Y* s spremenljivko *X* v celotni varianci. Definiran je na intervalu [0,1] in ga izračunamo kot:

$$R = \frac{\sigma_{y'}^2}{\sigma_y^2} = \frac{\sum_{i=1}^{N}(y_i{}' - \mu_Y)^2}{\sum_{i=1}^{N}(y_i - \mu_Y)^2}$$

Pove nam, koliko (kakšen delež) variabilnosti v vrednostih odvisne spremenljivke *Y* lahko pripišemo vrednostim neodvisne spremenljivke *X* in ne kakšnim drugim vplivom.

V primeru bivariatne linearne regresijske odvisnosti je determinacijski koeficient enak kvadratu Pearsonovega koeficienta korelacije:

$$R = \rho^2$$

Standardna napaka ocene predstavlja kvadratni koren iz nepojasnjene variance σ_e^2. Meri razpršenost točk okoli regresijske krivulje. V primeru linearne regresijske odvisnosti je standardna napaka:

$$\sigma_e = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2} = \sigma_Y\sqrt{1-\rho^2}$$

Pove nam, kakšen je standardni odklon dejanskih vrednosti y_i od napovedanih vrednosti y_i'. Pove torej, kako dobro ocenjena regresijska premica napoveduje vrednosti odvisne spremenljivke *Y* s pomočjo vrednosti neodvisne spremenljivke *X*.

5 PROUČEVANJE ODVISNOSTI V R

Preverimo, ali in kako ena neodvisna spremenljivka vpliva na odvisno.

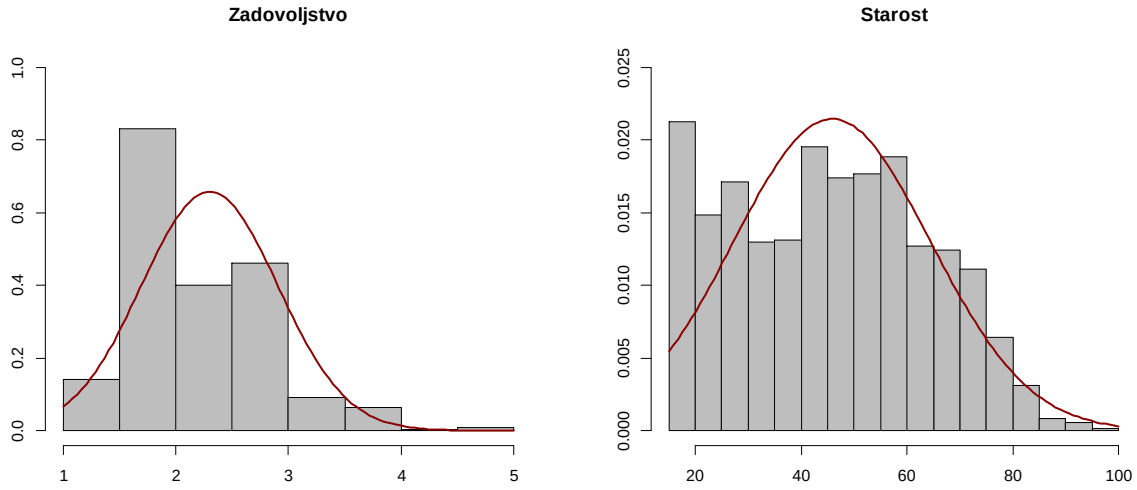
Kot **odvisno spremenljivko** vzemimo latentno spremenljivko, ki smo jo izračunali na osnovi rezultatov izvedene faktorske analize.

Kot **neodvisno spremenljivko** pa vzemimo eno izmed izbranih neodvisnih spremenljivk intervalnega ali razmernostnega tipa.

Kot primer v tem dokumentu bomo proučevali odvisnost med:

- odvisno spremenljivko *Zadovoljstvo* in
- neodvisno spremenljivko *Starost*.

Slika 1: Porazdelitev odvisne (Zadovoljstvo) in neodvisne (Starost) spremenljivke



Torej, zanima nas, ali starost vpliva na zadovoljstvo z življenjem. Če vpliv obstaja, nas zanima tudi, kakšen je ta vpliv.

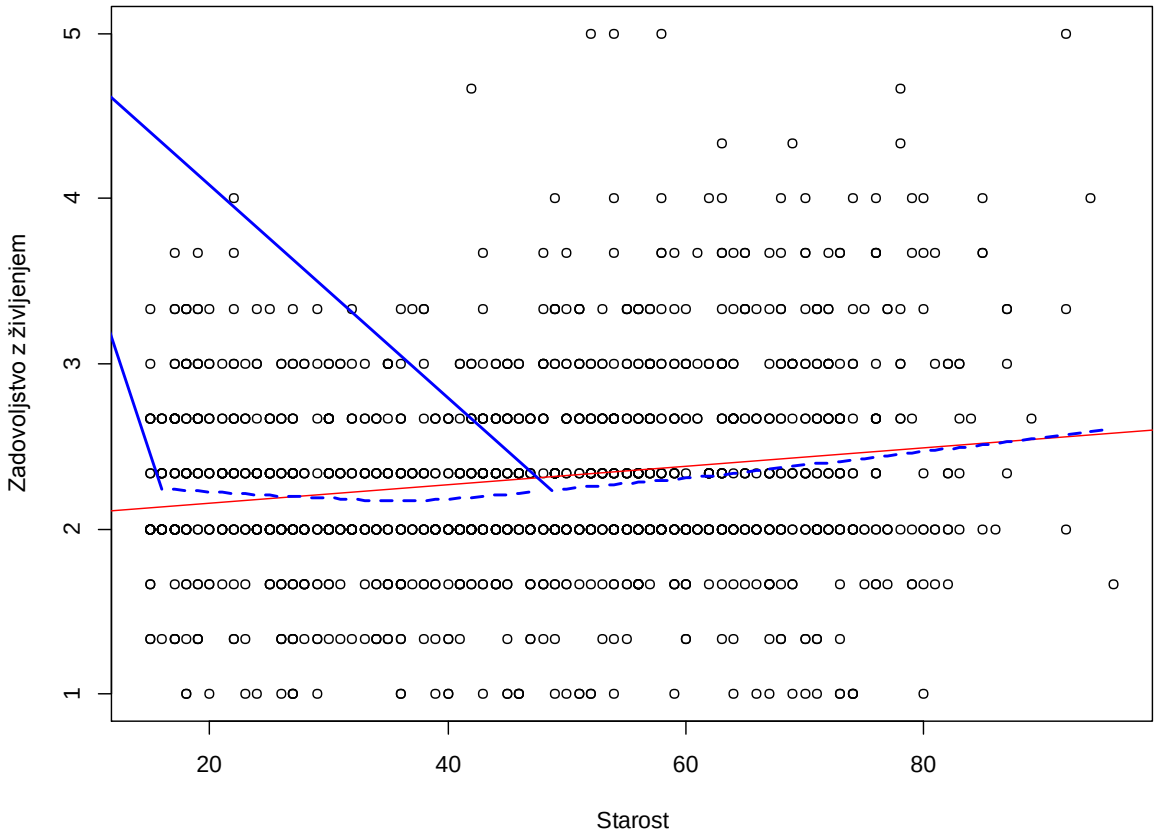
Najprej narišimo razsevni diagram za par izbranih spremenljivk:

```
plot(podatki2$Starost,podatki2$Zadovoljstvo,ylab="Zadovoljstvo z
življenjem",xlab="Starost")
```

Grafu dodamo še ravno regresijsko premico

```
abline(lm(podatki2$Zadovoljstvo~podatki2$Starost), col="red")
```

Slika 2: Odnos med zadovoljstvom z življenjem in starostjo



Za ocenjevanje linearne regresije uporabimo funkcijo `lm()`, ki ima naslednjo osnovno obliko:

```
lm(formula,data)
```

Pri tem: `formula` opisuje predpostavljeni regresijski model

`data` predstavlja podatkovni okvir, ki vključuje spremenljivke v modelu.

Rezultat funkcije je seznam, ki vključuje obsežne informacije o predpostavljenem regresijskem modelu.

Formulo zapišemo v obliki: `Y~X` (znak ~ ločuje odvisno spremenljivko na levi strani od neodvisne spremenljivke na desni strani).

Za naš primer dobimo naslednji rezultat:

```
fit1=lm(Zadovoljstvo~Starost,data=podatki2)
```

Izpis 10: Rezultat bivariatne linearne regresije (Zadovoljstvo~Starost)

```
Call:
lm(formula = Zadovoljstvo ~ Starost, data = podatki2)

Coefficients:
(Intercept)      Starost
    2.04124      0.00562
```

Torej, lahko zapišemo naslednjo regresijsko premico:

$$Zadovoljstvo = 2.041 + 0.0056 \cdot Starost$$

Kar pomeni:

- *b*=0.0056 pozitivna odvisnost: starejši so nekoliko bolj zadovoljni z življenjem kot mlajši. Vsako leto so anketiranci »zadovoljnejši« za 0.0056.
- *a*=2.041 Kolikšno je zadovoljstvo z življenjem ljudi, starih 0 let (v našem primeru to nima smisla). Parameter *a* lahko smatramo tudi kot **prilagoditveno konstanto** (angl. *adjustment constant*).

Preverimo še kakovost izračunanega regresijskega modela. To informacijo pridobimo s pomočjo funkcije `summary()`, ki ima naslednjo osnovno obliko:

```
summary(lm(formula,data))
```

in izpiše področne rezultate za izračunani regresijski model.

Izpis 11: Podroben rezultat izvedene bivariatne regresijske analize (Zadovoljstvo~Starost)

```
Call:
lm(formula = Zadovoljstvo ~ Starost, data = podatki2)

Residuals:
Love learning new things
    Min      1Q  Median      3Q     Max
-1.4909 -0.3560 -0.1013  0.3388  2.6665

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.0412406  0.0425075  48.021  < 2e-16 ***
Starost     0.0056201  0.0008593   6.541 8.57e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5983 on 1401 degrees of freedom

Multiple R-squared:  0.02963,   Adjusted R-squared:  0.02894

F-statistic: 42.78 on 1 and 1401 DF,  p-value: 8.569e-11
```

Ugotovimo lahko, da z variabilnostjo v starosti lahko pojasnimo le 3% variabilnosti zadovoljstva (`Multiple R-squared`).