LINEARNI MODELI 3

STATISTIČKI PRAKTIKUM 2

5

Do sada su nam varijable poticaja bile neprekidne (numeričke). Međutim, neke varijable poticaja mogu biti kvalitativne po prirodi (boja očiju, spol, ...). - zovemo ih *kategorijske varijable* ili *faktori*.

Kako se ovakve varijable poticaja mogu ugraditi u naše modele?

Analiza kovarijance se bavi problemima gdje se pojavljuju kombinacije kvantitativnih i kvalitativnih varijabli poticaja.

Pri ugradnji kvalitativnih varijable poticaja u model

$$Y = Xb + \varepsilon,$$

moramo ih kodirati.

Primjer

- $Y =$ promjena u razini kolesterola
- $x =$ broj godina (kvantitativna varijabla)
- $d = \begin{cases} 0, & \text{ne uzima lijek} \\ 1, & \text{uzima lijek} \end{cases}$ (kvalitativna varijabla)

Zanima nas, primjerice, utječe li uzimanje lijeka značajno na razinu kolesterola i je li taj utjecaj pozitivan ili negativan.

Ovisno o odnosu varijable poticaja x i kvalitativne varijable d biramo jedan od sljedećih linearnih modela:

1. Poseban model za pojedinu grupu (razdvojimo Y i x na grupu $d = 0$ i na grupu $d = 1$)

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

`> model1=lm(y~x)`

Teško testirati utjecaj lijeka.

2. Dva regresijska pravca s istim koeficijentom smjera

$$Y = \beta_0 + \beta_1 x + \beta_2 d + \varepsilon$$

`> model2=lm(y~x+d)`

U modelu se javlja fiksan utjecaj lijeka s obzirom na dob.

3. Različiti pravci za svaku grupu

$$Y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 x \cdot d$$

`> model3=lm(y~x*d)` ili `> model3=lm(y~x+d+x:d)`

Utjecaj lijeka nije fiksan, već ovisi i o godinama - *interakcija*.

Interakcija vs. korelacija

- Korelacija - statistička ovisnost jedne varijable o drugoj
- Interakcija - utjecaj jedne varijable poticaja na zavisnu varijablu ovisi o drugoj varijabli poticaja

Između dvije varijable poticaja može postojati interakcija bez obzira na to postoji li između njih korelacija ili ne.

Testiranje značajnost interakcije možemo napraviti usporedbom odgovarajućih linearnih modela.

Kolinearnost

Kolinearnost između prediktora - jedna varijabla može biti linearno predviđena pomoću ostalih (s određenom "točnošću").

Kolinearnost ne utječe na pouzdanost modela u cjelini (barem ne za dani skup podataka), nego na pojedine prediktore (procijenjeni parametri i p-vrijednosti mogu se značajno promijeniti ako napravimo male promjene u modelu ili podacima).

Posljedica je da ne možemo procijeniti utjecaj prediktora na zavisnu varijablu.

Matricu korelacija dobivamo na sljedeći način:

```
>summary(model, corr=T)
```

Primjer: *Dummy* varijabla s dva stupnja

Podaci za ovaj primjer se sastoje od visina `x`, dužina `y` i stila gradnje `style` srednjovjekovnih katedrala. Neke su romaničkog (r), a druge su gotičkog (g) stila. Podaci su upisani u `cathedral.txt`.

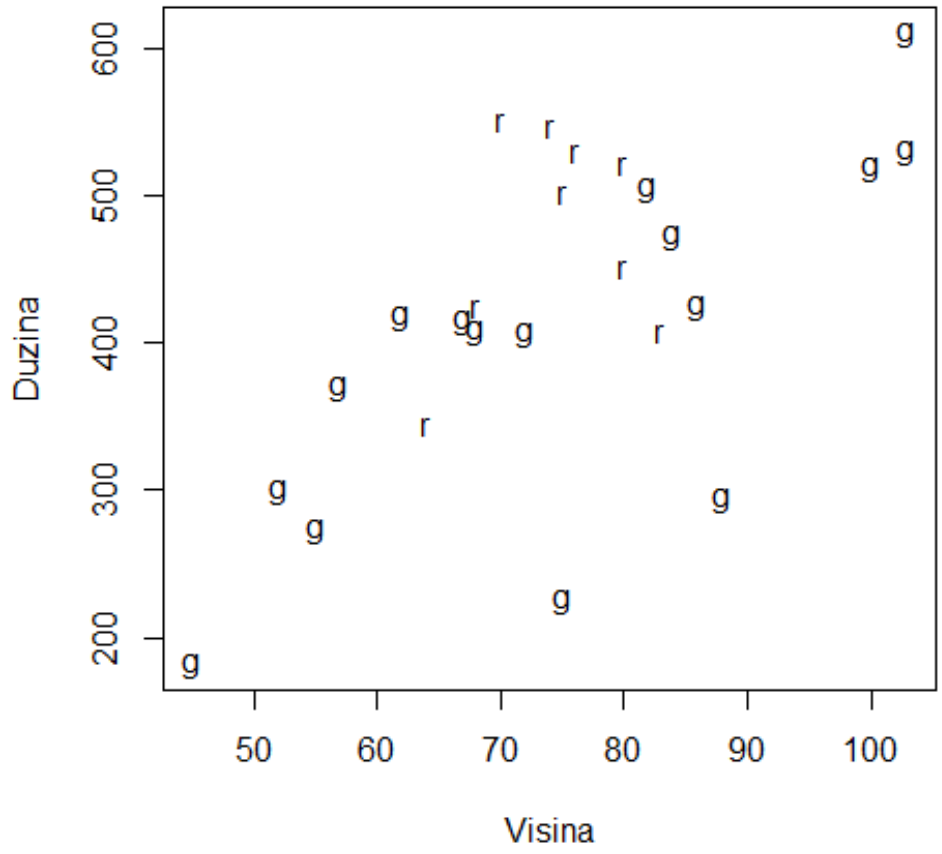
Učitajmo podatke:

```
> k=read.table("cathedral.txt")
> k
              style   x   y
Durham            r  75 502
Gloucester        r  68 425
---
WinchesterG       g 103 530
Salisbury         g  84 473
```

Zanima nas može li se dužina (`y`) opisati pomoću širine (`x`) i stila gradnje.

Grafička analiza:

```
> plot(k$x,k$y,type="n",xlab="Visina",ylab="Duzina")
> text(k$x,k$y,as.character(k$s))
```



Deskriptivna analiza:

```
>  lapply(split(k,k$style),summary)
$g
 style         x                y
 g:16   Min.   : 45.00   Min.   :182.0
 r: 0   1st Qu.: 60.75   1st Qu.:298.8
        Median : 73.50   Median :412.0
        Mean   : 74.94   Mean   :397.4
        3rd Qu.: 86.50   3rd Qu.:481.2
        Max.   :103.00   Max.   :611.0


$r
 style         x                y
 g:0    Min.   :64.00   Min.   :344.0
 r:9    1st Qu.:70.00   1st Qu.:425.0
        Median :75.00   Median :502.0
        Mean   :74.44   Mean   :475.4
        3rd Qu.:80.00   3rd Qu.:530.0
        Max.   :83.00   Max.   :551.0
```

Model:

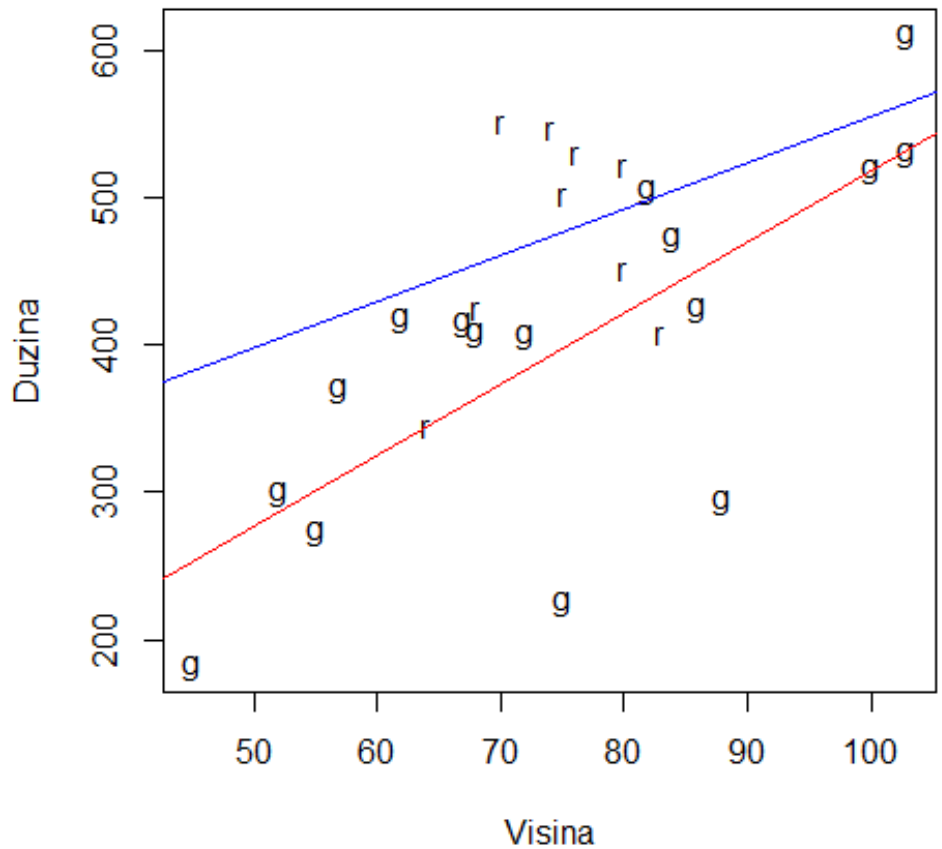
```
> model = lm(y ~ x * style,k)
> summary(model)

Call:
lm(formula = y ~ x * style, data = k)

Residuals:
    Min      1Q  Median      3Q     Max
-172.68  -30.22   23.75   55.78   89.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   37.111     85.675   0.433 0.669317
x              4.808      1.112   4.322 0.000301 ***
styler       204.722    347.207   0.590 0.561733
x:styler      -1.669      4.641  -0.360 0.722657
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.11 on 21 degrees of freedom
Multiple R-squared: 0.5412, Adjusted R-squared: 0.4757
F-statistic: 8.257 on 3 and 21 DF,  p-value: 0.0008072
```

Kako je kodirana varijabla `style` možemo vidjeti iz matrice modela $\mathbf{X}$:

```
> model.matrix(model)
             (Intercept)   x styler x:styler
Durham                 1  75      1        75
Canterbury             1  80      1        80
---
Old.St.Paul            1 103      0         0
Salisbury              1  84      0         0
```

Nacrtajmo pravce koji pripadaju modelu:

```
> abline(model$coef[-c(3,4)],col="red")
> abline(model$coef[1]+model$coef[3],model$coef[2]+model$coef[4],col="blue")
```



Kako je koeficijent uz `x:styler` malen i nije značajan, model se može pojednostavniti.

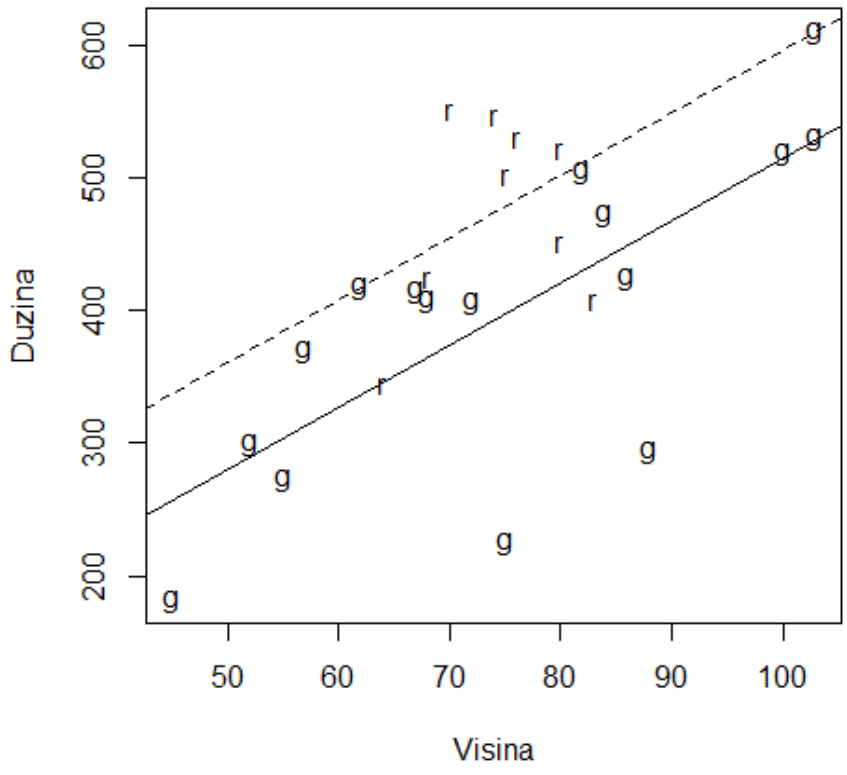
```
> model1=lm(y~x+style,k)
> summary(model1)

Call:
lm(formula = y ~ x + style, data = k)

Residuals:
    Min      1Q  Median      3Q     Max
-172.67  -30.44   20.38   55.02   96.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.298     81.648   0.543   0.5929
x              4.712      1.058   4.452   0.0002 ***
styler        80.393     32.306   2.488   0.0209 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77.53 on 22 degrees of freedom
Multiple R-squared: 0.5384, Adjusted R-squared: 0.4964
F-statistic: 12.83 on 2 and 22 DF,  p-value: 0.0002028
```

Usporedba dvaju modela pokazuje opravdanost naše pretpostavke.

```
> anova(model1,model)
Analysis of Variance Table

Model 1: y ~ x + style
Model 2: y ~ x + style + x:style
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     22 132223
2     21 131413  1       810 0.1294 0.7227
```

Nacrtajmo sada dva pravca koji pripadaju ovim podacima.

```
> abline(model1$coef[-3])
> abline(model1$coef[1]+model1$coef[3],model1$coef[2],lty=2)
```



Zaključak: za istu visinu, Romaničke su katedrale duže 80.39 feet-a i za svako povećanje za 1 foot, oba tipa katedrale će biti oko 4.7 feet-a duže.

Gotičke katedrale su uzete za referentne jer slovo 'g' se nalazi ispred 'r' u abecedi. Slovo 'r' možemo napraviti referentnim.

```
> k$style = relevel(k$sty,ref="r")
> model1=lm(y~x+style,k)
> summary(model1)

 Call:
lm(formula = y ~ x + style, data = k)

Residuals:
    Min      1Q  Median      3Q     Max
-172.67  -30.44   20.38   55.02   96.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  124.690     82.922   1.504   0.1469
x              4.712      1.058   4.452   0.0002 ***
styleg       -80.393     32.306  -2.488   0.0209 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77.53 on 22 degrees of freedom
Multiple R-squared: 0.5384, Adjusted R-squared: 0.4964
F-statistic: 12.83 on 2 and 22 DF,  p-value: 0.0002028
```

Kodiranje kvalitativnih varijabli s više od 2 kategorije

Kodiranje dvostupanjskih faktora nije jedinstveno, a još više je načina za kodiranje višestupanjskih faktora.

Za faktor koji ima k razina, potrebna nam je $k-1$ *umjetna* varijabla za reprezentaciju - jedan parametar se koristi da bi se ocijenio srednji efekt ili možda efekt nekog referentnog nivoa i $k-1$ varijabla nam je potrebna kako bi pokrili preostale slučajeve.

Postoje razne metode kodiranja, a mi ćemo se pozabaviti *tretirajućim kodiranjem* (za svaku kategoriju procjenjujemo njen efekt na Y neovisno od drugih kategorija).

Za testiranje značajnosti kategorijske varijable u modelu, uspoređujemo model bez ijedne umjetne varijable u odnosu na onaj sa svima.

Tretirajuće kodiranje

Faktor koji ima 4 razine bit će kodiran sa 3 umjetne varijable

		Umjetne varijable		
		d_1	d_2	d_3
	1	0	0	0
nivoi	2	1	0	0
	3	0	1	0
	4	0	0	1

Ovakav način prvi nivo tretira kao standardni/referentni (on ima sve umjetne varijable 0 pa je njegov utjecaj vidljiv u slobodnom članu), a ostale uspoređuje u odnosu na njega.

Ovo je standardni način kodiranja umjetnih varijabli u R-u.

Zadatak

U datoteci `twins.txt` nalaze se podaci o rezultatima IQ testiranja za jednojajčane blizance. Jednog blizanca su odgajali stvarni roditelji, a drugog usvojitelji. Dostupni su i podaci o socijalnoj skupini kojoj pripadaju stvarni roditelji.

(i) Grafički usporedite IQ posvojenog i IQ blizanca koji odrastao s biološkim roditeljima, pri tome naznačite socijalnu skupinu roditelja.

(ii) Analizirajte ovisnost IQ posvojenog blizanca o IQ-u blizanca koji odrastao s biološkim roditeljima i socijalnom statusu roditelja. Izaberite najbolji model.

(iii) Testirajte razlikuje li se IQ blizanaca?