Uvod u linearne modele

Testiranje linearne hipoteze

maj 2012. godine

Sadržaj

Opšta formulacija linearnih modela

- Da li se sećate opšte formule (matrične) linearnih modela kojima smo se ovde bavili?

$$y = \mathbf{X}\beta + \epsilon$$

- Šta su osnovni elementi?
  - y je $n \times 1$ vektor zavisne promenljive (odgovora)
  - $\mathbf{X}$ je $n \times p$ matrica modela
  - β je $p \times 1$ vektor parametara
  - ϵ je $n \times 1$ vektor greške
- Kako smo ocenili parametre metodom najmanjih kvadrata?

$$\beta = (X'X)^{-1}X'y$$

Testiranje linearne hipoteze

- Centralno pitanje je šta mi, zapravo, testiramo?!
- Opšti oblik hipoteze linearnih modela jeste:

$$H_0 : \mathbf{L}\beta = c$$

  - L je $q \times p$ matrica hipoteza
  - c je $q \times 1$ vektor za vrednost koju testiramo (što je obično 0)
- Prema takvoj null-hipotezi, možemo odrediti F-vrednost:

$$F_0 = \frac{(L\beta - c)'[L(X'X)^{-1}L']^{-1}(L\beta - c)}{qs^2}$$

TESTIRANJE LINEARNE HIPOTEZE

```
> require(car)
> data(Duncan)
> attach(Duncan)
> duncan.lm <- lm(prestige ~ income + education)

> summary(duncan.lm)

Call:
lm(formula = prestige ~ income + education)
Residuals:
    Min      1Q  Median      3Q     Max
-29.538  -6.417   0.655   6.605  34.641

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.06466    4.27194  -1.420    0.163
income       0.59873    0.11967   5.003 1.05e-05
education    0.54583    0.09825   5.555 1.73e-06

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-squared: 0.8282, Adjusted R-squared:  0.82
F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```


TESTIRANJE LINEARNE HIPOTEZE

```
> linearHypothesis(duncan.lm, hypothesis.matrix=c(0,1,-1))

Linear hypothesis test
Hypothesis:
income - education = 0

Model 1: restricted model
Model 2: prestige ~ income + education

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     43 7518.9
2     42 7506.7  1    12.195 0.0682 0.7952
```

VEŽBE

- oslobodite tabelu “Duncan” i aktivirajte tableu “Davis” u kojoj su podaci i procene visine i težine
- izvedite dve jednostavne regresije *weight repwt*
  - sa svim podacima
  - bez odstupajućeg podatka ispitanika 12
- rekli smo da, idealno, odsečak i nagim treba da budu 0 i 1
- testirajmo ovu pretpostavku
- mala pomoć

```
> diag(2)

     [,1] [,2]
[1,]    1    0
[2,]    0    1
```


REZULTATI VEŽBE

```
> linearHypothesis(davis.lm, diag(2), c(0,1))

Linear hypothesis test
Hypothesis:
(Intercept) = 0
repwt = 1

Model 1: restricted model
Model 2: weight ~ repwt

  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    183 13074
2    181 12828  2    245.97 1.7353 0.1793

> linearHypothesis(davis.lmB, diag(2), c(0,1))

Linear hypothesis test
Hypothesis:
(Intercept) = 0
repwt = 1

Model 1: restricted model
Model 2: weight ~ repwt

  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    182 974.00
2    180 914.31  2    59.691 5.8757 0.003373
```

Vežbe

Pokušajte sada, potpuno samostalno, da sprovedete jednu analizu linearnih modela, postavite hipotezu i testirajte je...

Sadržaj

formula

- `formula` ima levu i desnu stranu razdvojenu znakom '$\sim$'
- Ako je formula $a \sim b$, onda kažemo da a je modelovano kao b ili a je regresirano sa b
- Leva strana mora biti validan izraz koji će postati numerički vektor odgovarajuće dužine
- Na primer, podaci o konformizmu mogu se iskazati kao procenti ili kao log-šanse, umesto kao učestalosti

```
> lm(100*conformity/40 ~ fcategory, data=Moore)
> lm(logit(conformity/40) ~ fcategory, data=Moore)
```


formula

- Desna strana formule može sadržavati faktore i numeričke izraze koji će postati numerički vektori i/ili matrice
- Osnovni operatori imaju drugačije značenje, tj. nisu "zaštićeni":
  - `A + B` nije suma
  - ali `log(A + B)` jeste log sume
  - takođe, funkcija identiteta može "zaštititi" aritmetički izraz `I(A + B)`

```
> lm(prestige ~ I(income + education), data=Duncan)
```

formula

- Još neke mogućnosti koje nam daje R:
  - `a + b`
  - `a - b`, izuzeti b iz a
  - `a : b`
  - `a * b`, jeste `a + b + a:b`
  - `a * b * c - a:b:c`?
  - `b %in% a`, b je ugnežđeno u a
  - ...

subset

- `subset` smo već koristili
- On dopušta još nekoliko varijacija:
  - `subset = sex = 'F'`
  - `subset = 1:100`
  - `subset = -c(3,17)`

contrasts

- `contrasts` dopušta da se specifikuju kontrasti za faktore u linearnom modelu:
  - `contrasts = list(a=contr.sum, b=contr.poly)`

```
> moore.lm <- lm(conformity ~ partner.status + fcategory, data=Moore,
+   contrasts=list(partner.status=contr.sum, fcategory=contr.poly))
> summary(moore.lm)

Call:
lm(formula = conformity ~ partner.status + fcategory, data = Moore,
    contrasts = list(partner.status = contr.sum, fcategory = contr.poly))
Residuals:
    Min      1Q  Median      3Q     Max
-7.7236 -3.1978 -0.1978  2.8831 13.8831

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      12.0821     0.7339  16.462   <2e-16
partner.status1   2.3033     0.7782   2.960   0.0051
fcategory.L      -0.7744     1.3043  -0.594   0.5560
fcategory.Q      -0.5131     1.3153  -0.390   0.6985

Residual standard error: 4.922 on 41 degrees of freedom
Multiple R-squared: 0.1786, Adjusted R-squared: 0.1185
F-statistic: 2.971 on 3 and 41 DF,  p-value: 0.04279
```


Sadržaj

- funkcije za demonstraciju:

```
> source('data/demoUvodLM.R')
```

  - `regConst()`
    - min → minimum za varijablu X
    - max → maksimum za varijablu X
    - a → odsečak na Y-osi
    - b → nagib regresione funkcije
    - it → broj pokušaja (iteracija)
    - m → aritmetička sredina za grešku (epsilon)
    - s → standardna devijacija za grešku (epsilon)

- funkcije za demonstraciju:

```
> source('data/demoUvodLM.R')
```

  - `regVar()`
    - min → minimum za varijablu X
    - max → maksimum za varijablu X
    - a → odsečak na Y-osi
    - b → nagib regresione funkcije
    - it → broj pokušaja (iteracija)
    - m → aritmetička sredina za grešku (epsilon)
    - s → standardna devijacija za grešku (epsilon)
    - g → koeficijent priraštaja greške (epsilon) u funkciji X

- funkcije za demonstraciju:

```
> source('data/demoUvodLM.R')
```

  - `regCorr()`
    - n $\rightarrow$ veličina uzorka
    - m $\rightarrow$ aritmetičke sredine za dve varijable
    - r $\rightarrow$ koeficijent korelacije
    - it $\rightarrow$ broj pokušaja (iteracija)

K R A J