Uvod u linearne modele

Jednostavna linearna regresija

maj 2012. godine

Sadržaj

Osnove linearne regresije

- Osnovna formula jednostavne linearne regresije jeste:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

gde je $\epsilon_i \sim^{iid} N(0, \sigma^2)$

- Dakle, mi tu imamo posla sa sistemom linearnih jednačina, koje bi se, pojedinačno, mogle zapisati kao:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \epsilon_2 \\ &\vdots \quad \vdots \quad \vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \epsilon_n \end{aligned}$$

Matrični zapis

- Prethodno je daleko elegantnije računati s osloncem na matričnu algebru:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

odnosno:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Elementi jednostavne regresione analize

- U regresionoj analizi postoji nekoliko ključnih elemenata – matrica i vektora – sa posebnim ulogama:
  - matrica nacrta ili dizajna (*design matrix*) – nezavisni podaci

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{bmatrix}$$

Elementi jednostavne regresione analize

- U regresionoj analizi postoji nekoliko ključnih elemenata – matrica i vektora – sa posebnim ulogama:
  - vektor parametara (*vector of parameters*)

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Elementi jednostavne regresione analize

- U regresionoj analizi postoji nekoliko ključnih elemenata – matrica i vektora – sa posebnim ulogama:
  - vektor greške (*vector of error terms*)

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Elementi jednostavne regresione analize

- U regresionoj analizi postoji nekoliko ključnih elemenata – matrica i vektora – sa posebnim ulogama:
  - vektor odgovora (*vector of responses*) – zavisni podaci

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Najmanja suma kvadrata

- U rešavanju linearne regresije mi tražimo minimum sume kvadriranih reziduala
- Praktično, tražimo minimum za proizvod matrica

$$\sum \epsilon_i^2 = \begin{bmatrix} \epsilon_1 & \epsilon_2 & \dots & \epsilon_n \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \epsilon' \times \epsilon$$

Najmanja suma kvadrata

- Zamenom u osnovnom izrazu dobijamo da je:

$$\epsilon = \mathbf{Y} - \mathbf{X}\beta$$

pa je:

$$\epsilon'\epsilon = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$$

Najmanja suma kvadrata

- Uzimamo derivat s obzirom na β, pa sledi:

$$\frac{d}{d\beta}((\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)) = -2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)$$

- Tražimo rešenje u 0, za β:

$$-2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = 0$$

dakle:

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'X\beta$$

Najmanja suma kvadrata

- Rešavanje za β daje nam najmanje kvadratno odstupanje:

$$\mathbf{b} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Univerzitetski centar za primenjenu statistiku

Sadržaj

Podsetnik: transponovanje matrice

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

$$\mathbf{X}' = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

Podsetnik: proizvod dve matrice

$$\begin{aligned}\mathbf{AB} &= \begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix}\begin{bmatrix}3 & 0\\ 1 & 2\\ 0 & 1\end{bmatrix}\\ &= \begin{bmatrix}1*3+2*1+3*0 & 1*0+2*2+3*1\\ 4*3+5*1+6*0 & 4*0+5*2+6*1\end{bmatrix}\\ &= \begin{bmatrix}5 & 7\\ 17 & 16\end{bmatrix}\end{aligned}$$

Podsetnik: inverz matrice

$$\mathbf{A} = \begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}$$

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}}\begin{bmatrix}a_{22} & -a_{12}\\ -a_{21} & a_{11}\end{bmatrix}$$

Papir-olovka

- Uzmimo jedan sasvim jednostavan slučaj:

$$\mathbf{X} = \begin{bmatrix} 1 & 4.5 \\ 1 & 7.1 \\ 1 & 9.6 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} 6 \\ 9 \\ 11 \end{bmatrix}$$

Matrično izračunavanje pomoću R-a

```
> X <- matrix(data=c(1.0, 4.5, 1.0, 7.1, 1, 9.6), nrow=3, ncol=2, byrow=TRUE)
> Y <- matrix(data=c(6, 9, 11), nrow=3, ncol=1, byrow=TRUE)

> X

     [,1] [,2]
[1,]    1  4.5
[2,]    1  7.1
[3,]    1  9.6

> Y

     [,1]
[1,]    6
[2,]    9
[3,]   11

> t(X)

     [,1] [,2] [,3]
[1,]  1.0  1.0  1.0
[2,]  4.5  7.1  9.6
```

Matrično izračunavanje pomoću R-a

```
> t(X) %*% X

     [,1]   [,2]
[1,]  3.0  21.20
[2,] 21.2 162.82

> solve(t(X) %*% X)

           [,1]        [,2]
[1,]  4.1727319 -0.54331112
[2,] -0.5433111  0.07688365

> solve(t(X) %*% X) %*% (t(X) %*% Y)

          [,1]
[1,] 1.7303947
[2,] 0.9815479
```


Izračunavanje pomoću R funkcija

```
> solve(t(X) %*% X) %*% (t(X) %*% Y)
```

- Osnovna funkcija za sprovođenje postupka linearne regresije u R-u jeste `lm()`
- Za naš jednostavni slučaj:

```
> lm(Y ~ X[,2])

Call:
lm(formula = Y ~ X[, 2])
Coefficients:
(Intercept)       X[, 2]
     1.7304       0.9815
```

lm() FUNKCIJA

- Naravno, funkcija `lm()` ima daleko veće mogućnosti

```
> summary(lm(Y ~ X[,2]))

Call:
lm(formula = Y ~ X[, 2])
Residuals:
      1       2       3
-0.1474  0.3006 -0.1533

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7304     0.7521   2.301    0.261
X[, 2]        0.9815     0.1021   9.614    0.066

Residual standard error: 0.3682 on 1 degrees of freedom
Multiple R-squared: 0.9893, Adjusted R-squared: 0.9786
F-statistic: 92.43 on 1 and 1 DF,  p-value: 0.06598
```

- Zapamtite još i ovo:

```
> rezultat <- lm(Y ~ X[,2])
> summary(rezultat)
```


VEŽBA 1

- Otvorite tabelu sa podacima 'hsb.txt' i pokušajte da sprovedete jednostavnu linearnu regresiju na oba načina, sa promenljivama:
  - X $\rightarrow$ *read* (čitanje)
  - Y $\rightarrow$ *write* (pisanje)

```
> hsb = read.table('data/hsb.txt', header=TRUE, sep='\t')
> head(hsb)
> attach(hsb)
> X = cbind(1, read)
> Y = write
> solve(t(X) %*% X) %*% (t(X) %*% Y)
> summary(lm(Y ~ X[,2]))
> summary(lm(write ~ read))
```

Sadržaj

Osnove funkcije `lm()`

- Počećemo sa podacima merenja i izveštavanja o visini i težini 200 muškaraca i žena (Davis, 1990):

```
> require(car)

> head(Davis)
> nrow(Davis)
> names(Davis)

> davis.lm <- lm(weight ~ repwt, data=Davis)
```

- Obratite pažnju: sada smo u samoj funkciji `lm()` naveli tabelu sa podacima (*data frame*)
- Pomoću koje funkcije u R-u smo mogli da ovu tabelu aktiviramo i tako učinimo navođenje parametra `data` nepotrebnim?

Rezultati koje daje funkcija lm()

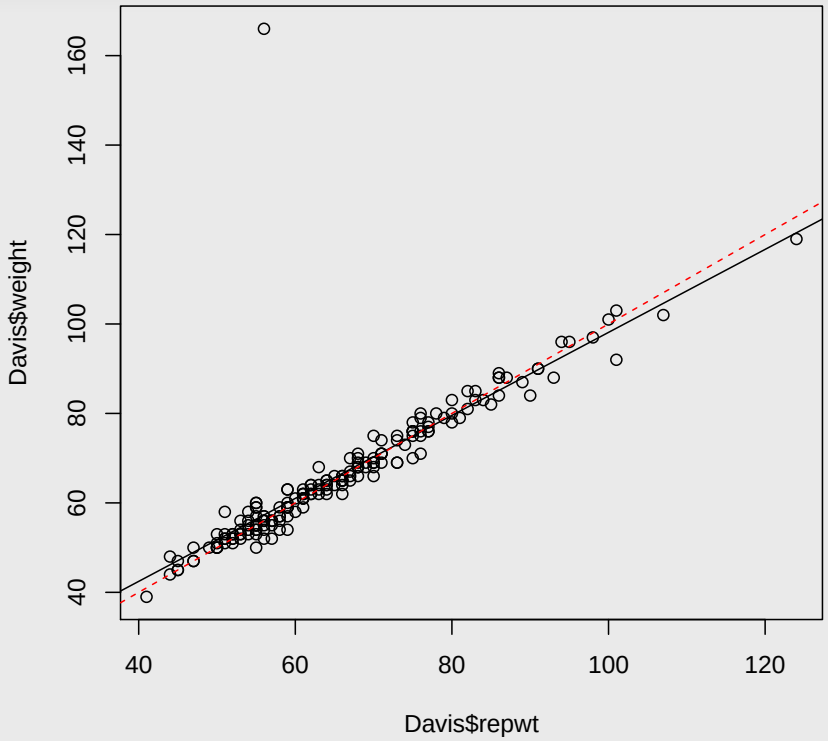
```
> davis.lm

Call:
lm(formula = weight ~ repwt, data = Davis)
Coefficients:
(Intercept)        repwt
     5.3363       0.9278

> summary(davis.lm)

Call:
lm(formula = weight ~ repwt, data = Davis)
Residuals:
    Min      1Q  Median      3Q     Max
 -7.048  -1.868  -0.728   0.601 108.705

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3363     3.0369   1.757   0.0806
repwt         0.9278     0.0453  20.484   <2e-16

Residual standard error: 8.419 on 181 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared: 0.6986, Adjusted R-squared: 0.697
F-statistic: 419.6 on 1 and 181 DF,  p-value: < 2.2e-16

> confint(davis.lm)

                2.5 %    97.5 %
(Intercept) -0.6560394 11.328560
repwt        0.8384665  1.017219
```


Preispitivanja rezultata

- Pošto vizuelno ispitivanje linearnosti odnosa i, uopšte, prirode raspodela nismo obavili prethodno, to moramo učiniti sada:

```
> plot(Davis$repwt, Davis$weight)
> abline(davis.lm)
> abline(0, 1, lty=2, col='red')
```

Dijagram raspršenja za model *weight* ∼ *repwt*



Interaktivni rad sa dijagramom raspršenja

```
> plot(Davis$repwt, Davis$weight)
> identify(Davis$repwt, Davis$weight)
```

- Dakle, utvrdili smo da postoje:
  - jedan odstupajući podatak (*outlier*)
  - jedan istupajući podatak (*extreme value*)

Rezultati nakon kritičke analize modela

- Možemo ukloniti odstupajući podatak i uporediti početni i konačni model:

```
> davis.lm.2 <- lm(weight ~ repwt, data=Davis, subset=-12)
> summary(davis.lm.2)

Call:
lm(formula = weight ~ repwt, data = Davis, subset = -12)
Residuals:
    Min      1Q  Median      3Q     Max
-7.5296 -1.1010 -0.1322  1.1287  6.3891

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.73380    0.81479   3.355 0.000967
repwt        0.95837    0.01214  78.926  < 2e-16

Residual standard error: 2.254 on 180 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared: 0.9719, Adjusted R-squared: 0.9718
F-statistic:  6229 on 1 and 180 DF,  p-value: < 2.2e-16

> cbind(Original=coef(davis.lm), NoCase12=coef(davis.lm.2))

             Original  NoCase12
(Intercept) 5.3362605 2.7338020
repwt       0.9278428 0.9583743
```

- Pokušajte sami sa istupajućim podatkom...

Vežbe

Vežba 2 Sprovedite postupak jednostavne linearne regresije za model *height* $\sim$ *repht*, sa podacima iz tabele 'Davis'

Vežba 3 Sprovedite postupak jednostavne linearne regresije za model *prestige* $\sim$ *income*, sa podacima iz tabele 'Prestige'

Sadržaj

Vrste "uticajnih podataka"

- odstupajuća vrednost (*outlier*)
- ekstremna vrednost (*extreme value*)

- Kakvu vrstu problema mogu proizvesti pomenuti podaci?
- Kako ih možemo dijagnostikovati?
- Kako možemo izolovati/blokirati njihov uticaj?

K R A J