

LINEARNI MODELI

STATISTIČKI PRAKTIKUM 2

2. VJEŽBE

Linearni model

Promatramo jednodimenzionalni linearni model.

$$Y = \beta_0 + \sum_{k=1}^p \beta_k x_k + \varepsilon$$

$x_1, x_2, \dots, x_p$ - varijable poticaja (kontrolirane)

ε - sl. greška

Y - varijabla odaziva

$\beta_0, \beta_1, \dots, \beta_p$ - parametri modela

Više opažanja

U primjeni imamo više opažanja, pa to zapisujemo

$$Y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ki} + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

gdje pretpostavljamo da su greške $\varepsilon_1, \dots, \varepsilon_n$ nezavisne s distribucijom $N(0, \sigma^2)$.

Kraće to zapisujemo u matričnom obliku

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \varepsilon,$$

gdje je $\mathbf{Y} = (Y_1, \dots, Y_n)^\tau$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\tau \sim N(\mathbf{0}, \sigma^2 \mathbb{I})$,
 $\mathbf{b} = (\beta_0, \beta_1, \dots, \beta_p)^\tau$ i

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$$

Metoda najmanjih kvadrata

Ako želimo minimizirati $\|\varepsilon\|_2 = \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|_2$ po $\mathbf{b}$ dobivamo da je najbolja ocjena za $\mathbf{b}$

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y},$$

(uz uvjet da je $\mathbf{X}^T \mathbf{X}$ regularna).

Procijenjene vrijednosti tada su jednake

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}} = \underbrace{\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T}_{\mathbf{H}} \mathbf{Y},$$

a ostaci

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbb{I} - \mathbf{H})\mathbf{Y}.$$

Što sve vrijedi u našem modelu

- ▶ $\hat{\mathbf{b}} \sim N(\mathbf{b}, (\mathbf{X}^\tau \mathbf{X})^{-1} \sigma^2);$
- ▶ $\frac{\hat{b}_i - b_i}{\hat{\sigma} \sqrt{(\mathbf{X}^\tau \mathbf{X})_{ii}^{-1}}} \sim t(n - p - 1);$
- ▶ $\mathbf{e} \sim N(0, (\mathbb{I} - H) \sigma^2);$
- ▶ $\sum_{i=1}^n e_i = 0;$
- ▶ $\hat{\sigma}^2 = \frac{\mathbf{e}^\tau \mathbf{e}}{n - p - 1}$ je nepristrani procjenitelj za σ^2 i vrijedi

$$\frac{(n - p - 1) \hat{\sigma}^2}{\sigma^2} \sim \chi^2(n - p - 1);$$

Goodness of fit

Obično na tri načina ocjenjujemo koliko je model dobar:

- ▶ $R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$, koeficijent determinacije;
- ▶ $\hat{\sigma}^2 = \frac{SSR}{n-p-1}$;
- ▶ $R_a^2 = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2 / (n-p-1)}{\sum_{i=1}^n (Y_i - \bar{Y})^2 / (n-1)}$, prilagođeni R^2 ;
- ▶ $F = \frac{\frac{SSR_0 - SSR}{p - p_0}}{\frac{SSR}{n-p}} \sim F(p - p_0, n - p)$ test značajnosti modela.

Procjena parametara modela

Za procjenu parametara modela koristimo naredbu `lm`, pritom rezultate analize možemo spremiti u neki objekt (specijalnog tipa "`lm`").

$$\text{rez}=\text{lm}(Y\sim x_1+x_2+\dots+x_n)$$

Funkcija `summary` kao argument može primiti objekat tipa `lm` i pritom ispisuje

- ▶ rezultate testova značajnosti parametara
- ▶ R^2 , R_a^2 , $\hat{\sigma}$
- ▶ rezultate testa značajnosti modela

Zadatak

U datoteci `gala.txt` dani su podaci o broju vrsta kornjača na galapagoškom otočju (Johnson, Raven 1973.g). Postoji 30 otoka i 7 varijabli za svaki od njih. Varijable su

- ▶ *Species* - broj vrsta na pojedinom otoku
- ▶ *Endemics* - broj endemskih vrsta
- ▶ *Area* - površina otoka
- ▶ *Elevation* - visina najviše točke na otoku (m)
- ▶ *Nearest* - udaljenost do najbližeg otoka
- ▶ *Scruz* - udaljenost od otoka Santa Cruz
- ▶ *Adjacent* - površina najbližeg (susjednog) otoka

Postoji li linearna ovisnost varijable *Species* o varijablama *Area*, *Elevation*, *Nearest*, *Scruz*, *Adjacent*?

Pojedini podaci

U objektu `rez` pohranjeni su sljedeći podaci o linearnoj regresiji:

```
> names(rez)
[1] "coefficients"      "residuals"          "fitted.values"
[4] "effects"           "R"                   "rank"
[7] "qr"                "family"              "linear.predictors"
[10] "deviance"          "aic"                 "null.deviance"
[13] "iter"              "weights"             "prior.weights"
[16] "df.residual"       "df.null"             "y"
[19] "converged"         "boundary"            "model"
[22] "call"              "formula"             "terms"
[25] "data"              "offset"              "control"
[28] "method"            "contrasts"           "xlevels"
```

Procijenjene ostatci $\mathbf{e}$, vektor koeficijenta $\hat{\mathbf{b}}$ i procijenjene vrijednosti $\hat{\mathbf{y}}$ pohranjeni su redom u objekte

```
> rez$res;
> rez$coef;
> rez$fit.
```

Rezultate poziva funkcije `summary` možemo spremiti u neki objekt

```
> sum=summary(rez)
> names(sum)
[1] "call"          "terms"          "residuals"      "coefficients"
[5] "aliased"        "sigma"          "df"              "r.squared"
[9] "adj.r.squared" "fstatistic"     "cov.unscaled"
```

Procjena varijance grešaka $\hat{\sigma}$, matrica $(\mathbf{X}^T \mathbf{X})^{-1}$, R^2 i R_a^2 pohranjeni su redom u objekte

```
> sum$sig
> sum$cov
> sum$r
> sum$adj.r
```

Zadatak

- (a) Simulirajte podatke za model

$$Y_i = 1 + x_i + 2 \sin(x_i) + \varepsilon_i,$$

gdje je $x_i = i/10$, za $i = 0, 1, \dots, 100$ i $\varepsilon_i \sim N(0, 1)$ nezavisne.

- (b) Simulirane podatke prikažite grafički, zajedno s krivuljom srednje vrijednosti modela

$$y(x) = 1 + x + 2 \sin x.$$

- (c) Na temelju simuliranih podataka procijenite parametre modela

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 \sin(x_i) + \varepsilon_i$$

i

- (d) na prethodni graf dodajte krivulju procijenjenih vrijednosti modela

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 \sin(x).$$

Pouzdani intervali za koeficijente

Odredimo 90% pouzdani interval za svaki od koeficijenata.

```
> confint(sl,level=0.9)
              5 %      95 %
(Intercept) 0.7685881 1.451205
x            0.9296179 1.045557
sin(x)       1.9348879 2.443392
```

(Koristi se statistika $\frac{\hat{b}_i - b_i}{\hat{\sigma} \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}}} \sim t(n - p - 1).$)

Pouzdaní intervali za procjene

Ako želimo procijeniti varijablu odaziva u novoj točki $\mathbf{x}_0$, to je naravno $\hat{y}_0 = \mathbf{x}_0^T \hat{\mathbf{b}}$. No razlikujemo dvije procjene:

1. Procjena srednje vrijednosti $\mathbf{x}_0^T \mathbf{b}$ (model bez šumova, želimo eliminirati greške u mjerenju).
2. Procjena opažanja $\mathbf{x}_0^T \mathbf{b} + \varepsilon$ (model sa šumom, zanimaju nas i greške).

U prvom slučaju $100(1 - \alpha)\%$ pouzdan interval je

$$\hat{y}_0 \pm t_{\alpha/2}(n - p - 1) \hat{\sigma} \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0},$$

a u drugom

$$\hat{y}_0 \pm t_{\alpha/2}(n - p - 1) \hat{\sigma} \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}.$$

U modelu je p varijabli poticaja, ali ukupan broj parametara je $p+1$ (uključujemo slobodni član).

1. Procjena srednje vrijednosti

Procijenimo 80% pouzdani interval za srednje vrijednosti u točkama $(x_i)_{i=0}^{100}$ iz prethodnog zadatka. U varijabli `sl` spremljeni su parametri modela.

```
> pr1=predict(sl,level=0.8,i="c")
> pr1[1:4,]
      fit      lwr      upr
1  1.109897 0.8446996 1.375094
2  1.427205 1.1690792 1.685330
3  1.742329 1.4897851 1.994873
4  2.053108 1.8046079 2.301608
```

Procijenimo srednju vrijednost u točki $xx_0 = 11$ i 60% pouzdani interval.

```
> xx0=data.frame(x=11)
> predict(sl,xx0,level=0.6,i='c')
      fit      lwr      upr
1 9.78424 9.543608 10.02487
> 1+xx0+2*sin(xx0)
      x
1 10.00002
```

2. Procjena pouzdanih intervala za opažanja

Procijenimo 85% pouzdani interval za opažanja u točkama $(x_i)_{i=0}^{100}$.

```
> pr2=predict(s1,level=0.85,i='p')
> pr2[1:4,]
      fit      lwr      upr
1 1.109897 -0.39965520 2.619448
2 1.427205 -0.08079634 2.935206
3 1.742329  0.23552357 3.249135
4 2.053108  0.54715285 3.559063
```

Procijenimo opaženu vrijednost u točki $xx_0 = 11$ i 70% pouzdani interval.

```
> predict(s1,xx0,level=0.7,i='p')
      fit      lwr      upr
1 9.78424 8.680923 10.88756
> 1+xx0+2*sin(xx0)+rnorm(1,0,1)
      x
1 10.27153
```

Zadatak

Prikažite na grafu simulirane podatke i

- (a) 80%-pouzdanu prugu za srednju vrijednost.
- (b) 85%-pouzdanu prugu za opažanja.