LINEARNI MODELI

STATISTIČKI PRAKTIKUM 2

2. VJEŽBE

Linearni model

Promatramo jednodimenzionalni linearni model.

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

$x_1, x_2, \ldots, x_p$ - varijable poticaja (kontrolirane)
ε - sl. greška
Y - varijabla odaziva
$\beta_0, \beta_1, \ldots, \beta_p$ - parametri modela

Više opažanja

U primjeni imamo više opažanja, pa to zapisujemo

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

gdje pretpostavljamo da su greške $\varepsilon_1, \ldots, \varepsilon_n$ nezavisne s distribucijom $N(0, \sigma^2)$.
Kraće to zapisujemo u matričnom obliku

$$\mathsf{Y} = \mathsf{Xb} + \varepsilon,$$

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

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

Metoda najmanjih kvadrata

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

$$\hat{\mathsf{b}} = (\mathsf{X}^\tau \mathsf{X})^{-1} \mathsf{X}^\tau \mathsf{Y},$$

(uz uvjet da je $\mathsf{X}^\tau \mathsf{X}$ regularna).
Procijenjene vrijednosti tada su jednake

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

a ostaci

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

Što sve vrijedi u našem modelu

- $\hat{\mathrm{b}} \sim N(\mathrm{b}, (\mathrm{X}^\tau \mathrm{X})^{-1}\sigma^2)$;
- $\dfrac{\hat{b}_i - b_i}{\hat{\sigma}\sqrt{(\mathrm{X}^\tau \mathrm{X})^{-1}_{ii}}} \sim t(n-p-1)$;
- $\mathrm{e} \sim N(0, (\mathbb{I} - H)\sigma^2)$;
- $\sum_{i=1}^{n} e_i = 0$;
- $\hat{\sigma^2} = \frac{\mathrm{e}^\tau \mathrm{e}}{n-p-1}$ 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 četiri načina ocjenjujemo koliko je model dobar:

- $R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$, koeficijent determinacije;
- $\hat{\sigma}^2 = \frac{\text{SSR}}{n-p-1}$;
- $R_a^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2/(n-p-1)}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2/(n-1)}$, 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`").

```
rez=lm(Y~x_1+x_2+...+x_n)
```

Funkcija `summary` kao argument može primiti objek 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 e, vektor koeficijenta $\hat{b}$ i procijenjene vrijednosti $\hat{y}$ 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 $(X^\tau X)^{-1}$, R^2 i R_a^2 pohranjeni su redom u objekte

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

Pouzdani intervali za koeficijente

Odredimo 90% pouzdani interval za svaki od koeficijenata.

```
> confint(rez,level=0.9)
                    5 %       95 %
(Intercept) -25.70235310 39.83879452
Area         -0.06230034  0.01442366
Elevation     0.22765403  0.41127549
```

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

Pouzdani intervali za procjene

Ako želimo procijeniti varijablu odaziva u novoj točki x_0, to je naravno $\hat{y}_0 = x_0^\tau \hat{b}$. No razlikujemo dvije procjene:

1. Procjena srednje vrijednosti $x_0^\tau b$ (model bez šumova, želimo eliminirati greške u mjerenju).
2. Procjena opažanja $x_0^\tau b + \varepsilon$ (model sa šumom, zanimaju nas i greške).

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

$$\hat{y}_0 \pm t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{x_0^\tau (X^\tau X)^{-1} x_0},$$

a u drugom

$$\hat{y}_0 \pm t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{1 + x_0^\tau (X^\tau X)^{-1} x_0}.$$

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 broja vrsta u svim točkama linearne regresije, tj. za sve otoke. U varijabli `rez` su spremljeni parametri modela.

```
> pr1=predict(rez,level=0.8,i="c")
> pr1[1:2,]
                 fit         lwr       upr
Baltra    116.725946  94.180683 139.27121
Bartolome  -7.273154 -31.560299  17.01399
```

Procijenimo srednju vrijednost i 70% pouzdani interval broja vrsta na otoku A čije vrijednosti parametara iznose redom:

```
Area=10, Elevation=150, Nearest=2, Scruz=40, Adjacent=700.
```

```
> xx0=data.frame(Area=10, Elevation=150, Nearest=2, Scruz=40,
+ Adjacent=700)
> predict(rez,xx0,level=0.7,i='c')
       fit       lwr      upr
1 -7.217512 -26.34593 11.91091
```

2. Procjena pouzdanih intervala za opažanja

Procijenimo 85% pouzdani interval za opažanja u svim točkama regresije, tj. za sve otoke.

```
> pr2=predict(rez,level=0.85,i='p')
> pr2[1:4,]
                 fit         lwr       upr
Baltra    116.725946   22.54609 210.90580
Bartolome  -7.273154 -102.00292  87.45661
Caldwell   29.330659  -64.13032 122.79164
Champion   10.364266  -83.53830 104.26683
```

Procijenimo opaženu vrijednost i 60% pouzdani interval broja vrsta na otoku A.

```
> predict(rez,xx0,level=0.6,i='p')
       fit       lwr      upr
1 -7.217512 -61.70731 47.27229
```

Zadatak

(a) Simulirajte podatke za model

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

gdje je $x_i = i/10$, za $i = 0, 1, \ldots, 100$ 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

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

Zadatak

Prikažite na grafu simulirane podatke i

(a) 80%-*pouzdanu prugu* za srednju vrijednost.

(b) 85%-*pouzdanu prugu* za opažanja.

Zatim procijenite

(c) srednju vrijednost u točki $xx_0 = 11$ i 60% pouzdani interval,

(d) opaženu vrijednost u točki $xx_0 = 11$ i 70% pouzdani interval,

te usporedite sa stvarnim vrijednostima.