Kazalo

1 LINEARNI MODEL V MATRIČNI OBLIKI

V linearnem modelu odzivno spremenljivko y modeliramo na podlagi k regresorjev

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

Krajši in bolj eleganten je zapis v matrični obliki:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \ddots & \vdots & \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

V (1) je $\mathbf{y}$ vektor odzivne spremenljivke, $\mathbf{X}$ je modelska matrika reda $(n \times k+1)$, $\boldsymbol{\beta}$ je vektor parametrov modela velikosti $(k+1)$ in $\boldsymbol{\varepsilon}$ je vektor napak velikosti (n), za katerega velja $E(\boldsymbol{\varepsilon}) = 0$ in $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, $\mathbf{I}$ je enotska diagonalna matrika reda $n \times n$.

1.1 Cenilke parametrov po metodi najmanjših kvadratov

Po metodi najmanjših kvadratov (*Ordinary Least Squares*, OLS) cenilko za $\boldsymbol{\beta}$dobimo z minimiranjem izraza:

$$S(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}))^2. \tag{3}$$

Funkcijo (3) parcialno odvajamo po parametrih β_j, $j = 0, ..., k$, in odvode izenačimo z 0. Dobimo t. i. normalni sistem $k + 1$ linearnih enačb.

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

Rešitev tega sistema so cenilke parametrov, ki jih označimo b_j, $j = 0, ..., k$. Rešitev obstaja, če je $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ nesingularna. To velja,

- če je $n \geq k + 1$; to pomeni, da je število enot vsaj tako veliko kot število ocenjevanih parametrov;
- če nobena spremenljivka ni linearna kombinacija ostalih spremenljivk.

Rešitev je:

$$\boldsymbol{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{4}$$

Izračun cenilk parametrov z matrikami za primer SKT naj služi kot ilustracija in vaja:

```
> tlak<-read.table(file="SKT.txt", header = TRUE)
> model.SKT <- lm(SKT~starost, data=tlak)
> X <- model.matrix(model.SKT)   # modelska matrika
> X[1:5,]   # prvih 5 vrstic

  (Intercept) starost
1           1      41
2           1      60
3           1      41
4           1      47
5           1      66

> t(X) %*% X   # kaj je v tej matriki: n,  vsota x, vsota x*x

            (Intercept) starost
(Intercept)          69    3184
starost            3184  162388

> t(X) %*% tlak$SKT  # kaj je v tej matriki: vsota y, vsota x*y

              [,1]
(Intercept)  10262
starost     488744

> b<-solve(t(X) %*% X) %*% t(X) %*% tlak$SKT; b  # cenilki parametrov

                   [,1]
(Intercept) 103.3490547
starost       0.9833276
```

1.2 Statistične lastnosti cenilk parametrov

Cenilke parametrov dobljene po metodi najmanjših kvadratov so nepristranske:

$$E(\boldsymbol{b}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}E(\mathbf{y}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta}. \tag{5}$$

Edina predpostavka uporabljena pri dokazu je, da je $E(\boldsymbol{\varepsilon}) = 0$. Iz tega sledi, da so cenilke parametrov modela nepristranske tudi, če varianca σ^2 ni konstantna ali če so napake korelirane. Variančno-kovariančna matrika za $\boldsymbol{b}$ je vezana na varianco za $\mathbf{y}$ in na modelsko matriko:

$$Var(\boldsymbol{b}) = ((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})Var(\mathbf{y})((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})^{\mathrm{T}} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}. \tag{6}$$

Porazdelitev parametrov normalnega linearnega modela je **večrazsežna normalna porazdelitev**.

Za ocene parametrov linearnega modela $\mathbf{b}$, dobljene po metodi najmanjših kvadratov, lahko pokažemo, da so **najboljše linearne nepristranske cenilke** za $\boldsymbol{\beta}$ (BLUE, *Best Linear Unbiesed Estimator*), kar pomeni, da imajo med linearnimi cenilkami najmanjšo varianco (Gauss-Markov izrek).

Pokažemo lahko, da je nepristranska cenilka za σ^2

$$\hat{\sigma}^2 = \frac{1}{n-k-1}(\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}). \tag{7}$$

1.3 Matrika H

Poglejmo povezavo med $\hat{\mathbf{y}}$ in $\mathbf{y}$:

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{b} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}. \tag{8}$$

Matrika $\mathbf{H}$, t. i. "hat matrix", je ključna pri izračunu napovedi $\hat{\mathbf{y}}$:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}. \tag{9}$$

Pokažemo lahko, da velja: $\mathbf{H} = \mathbf{H}^{\mathrm{T}} = \mathbf{H}^2$.

Izračun matrike $\mathbf{H}$ za primer SKT:

```
> H<-X %*% solve(t(X) %*% X) %*% t(X)
> dim(H)
```

```
[1] 69 69
```

```
> round(H[1:10,1:10],3)  # izpis prvih 10 stolpcev in 10 vrstic matrike H
```

```
      1     2     3     4     5     6     7     8     9     10
1  0.016 0.010 0.016 0.014 0.008 0.014 0.007 0.016 0.007 0.011
2  0.010 0.027 0.010 0.015 0.032 0.015 0.034 0.012 0.034 0.024
3  0.016 0.010 0.016 0.014 0.008 0.014 0.007 0.016 0.007 0.011
4  0.014 0.015 0.014 0.015 0.016 0.015 0.016 0.014 0.016 0.015
5  0.008 0.032 0.008 0.016 0.040 0.016 0.043 0.010 0.043 0.028
6  0.014 0.015 0.014 0.015 0.016 0.015 0.016 0.014 0.016 0.015
7  0.007 0.034 0.007 0.016 0.043 0.016 0.045 0.010 0.045 0.030
8  0.016 0.012 0.016 0.014 0.010 0.014 0.010 0.015 0.010 0.012
9  0.007 0.034 0.007 0.016 0.043 0.016 0.045 0.010 0.045 0.030
10 0.011 0.024 0.011 0.015 0.028 0.015 0.030 0.012 0.030 0.022
```

Diagonalne člene matrike $\mathbf{H}$ imenujemo vzvodi (*leverages* ali *hatvalues*) in jih dobimo z ukazom `hatvalues`:

```
> round(hatvalues(model.SKT),3)
```

```
    1     2     3     4     5     6     7     8     9    10    11    12    13
0.016 0.027 0.016 0.015 0.040 0.015 0.045 0.015 0.045 0.022 0.037 0.022 0.029
   14    15    16    17    18    19    20    21    22    23    24    25    26
0.021 0.015 0.015 0.015 0.062 0.052 0.055 0.019 0.017 0.016 0.066 0.016 0.020
   27    28    29    30    31    32    33    34    35    36    37    38    39
0.037 0.026 0.049 0.051 0.021 0.031 0.016 0.015 0.025 0.017 0.023 0.026 0.041
   40    41    42    43    44    45    46    47    48    49    50    51    52
0.029 0.018 0.015 0.015 0.037 0.014 0.043 0.016 0.043 0.021 0.035 0.021 0.025
   53    54    55    56    57    58    59    60    61    62    63    64    65
0.024 0.016 0.015 0.015 0.069 0.059 0.062 0.021 0.015 0.018 0.055 0.015 0.018
   66    67    68    69
0.033 0.034 0.043 0.048
```

1.4 Ostanki

V matrični obliki izračunamo ostanke takole:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}. \tag{10}$$

Varianca ostankov $Var(\mathbf{e})$ je ob predpostavki $Var(\epsilon) = \sigma^2\mathbf{I}$:

$$Var(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\ (\sigma^2\mathbf{I})\ (\mathbf{I} - \mathbf{H})^{\mathrm{T}} = \sigma^2(\mathbf{I} - \mathbf{H}). \tag{11}$$

1.5 Inferenca v linearnem modelu

Inferenca v linearnem modelu polnega ranga temelji na tem, da so ocene parametrov **b** porazdeljene po multivariatni normalni porazdelitvi

$$\mathbf{b} \sim N(\boldsymbol{\beta},\quad \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}). \tag{12}$$

Poleg tega velja, da so ocene prametrov **b** neodvisne od ocene variance napak $\hat{\sigma}^2$. Porazdelitev statistike $(n-k-1)\hat{\sigma}^2/\sigma^2$ je χ^2-porazdelitev s stopinjami prostosti $SP = n-k-1$:

$$\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k-1}. \tag{13}$$

1.5.1 Intervalne ocene za parametre modela

Interval zaupanja za posamezen parameter modela β_j, $j = 0, ..., k$, **ob upoštevanju ostalih regresorjev v modelu** imenujemo **parcialni interval zaupanja**. Definiran je na podlagi statistike $\frac{b_j-\beta_j}{\hat{\sigma}\sqrt{a_{jj}}}$, kjer je $\sqrt{a_{jj}}$ diagonalni element matrike $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, to matriko označimo **A**. Velja, da je statistika:

$$\frac{b_j-\beta_j}{\sigma\sqrt{a_{jj}}} \sim N(0,1). \tag{14}$$

Ko izraz (14) delimo s korenom spremenljivke (13) deljene z n-k-1, kjer je (13) porazdeljena po χ^2_{n-k-1} porazdelitvi, dobimo statistiko, ki je porazdeljena po t-porazdelitvi s $SP = n-k-1$:

$$\frac{b_j-\beta_j}{\sigma\sqrt{a_{jj}}} \Big/ \sqrt{\frac{\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2}}{n-k-1}} \sim t(SP = n-k-1), \tag{15}$$

ko zgornji izraz poenostavimo, dobimo

$$\frac{b_j-\beta_j}{\hat{\sigma}\sqrt{a_{jj}}} \sim t(SP = n-k-1). \tag{16}$$

Posledično je $100(1-\alpha)$ % interval zaupanja za β_j ob upoštevanju ostalih napovednih spremenljivk v modelu:

$$(b_j - t_{\alpha/2}(SP = n-k-1)\hat{\sigma}\sqrt{a_{jj}},\ b_j + t_{\alpha/2}(SP = n-k-1)\hat{\sigma}\sqrt{a_{jj}}). \tag{17}$$

Funkcija `confint()` vrne parcialne 95 % intervale zaupanja za vse parametre v modelu. Pri njihovi interpretaciji se je potrebno zavedati, da so to intervali zaupanja za posamezen parameter ob upoštevanju vseh ostalih členov v modelu.

1.5.2 Testiranje domnev o parametrih modela

Za testiranje ničelne domneve $H_0 : \beta_j = \gamma$ ob alternativni domnevi $H_0 : \beta_j \neq \gamma$ in ob upoštevanju vseh ostalih členov v modelu, uporabimo testno statistiko

$$t = \frac{b_j - \gamma}{\hat{\sigma}\sqrt{a_{jj}}}, \tag{18}$$

ki je pod ničelno domnevo porazdeljena po Studentovi porazdelitvi s $SP = n - k - 1$.

Rezultate testiranja posamičnih $k + 1$ ničelnih domnev za parametre modela dobimo v povzetku `lm` modela.

1.5.3 F-test za model

Pri modelu enostavne linearne regresije smo videli, da lahko ničelno domnevo $H_0 : \beta_1 = 0$ preverimo tudi na podlagi F-statistike, ki je pod ničelno domnevo porazdeljena $F(SP_1 = 1, SP_2 = n - k - 1)$. Za linearni model z več napovednimi spremenljivkami na podlagi F-statistike testiramo ničelno domnevo, da so parametri $(\beta_1, \beta_2, ..., \beta_k)$ hkrati enaki nič:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0.$$

Alternativna domneva pravi, da je vsaj en parameter β_j, $j = 1, ..., k$, različen od nič.

Vsoto kvadriranih odklonov za odzivno spremenljivko SS_{yy}, enako kot pri enostavni linearni regresiji, razdelimo na dva dela

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ SS_{yy} &= SS_{model} + SS_{residual} \\ &= (\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - C) + (\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}), \end{aligned} \tag{19}$$

kjer je $C = (\sum_{i=1}^{n} y_i)^2/n$ je t. i. korekcijski člen. F-statistika je definirana z razmerjem

$$F = \frac{SS_{model}/k}{SS_{residual}/(n - k - 1)}. \tag{20}$$

Ob predpostavki $\varepsilon \sim iid\ N(0, \sigma^2)$ je F-statistika porazdeljena po F-porazdelitvi s $SP_1 = SP_{model} = k$ in $SP_2 = SP_{residal} = n - k - 1$.

Tabela 1: Shema tabele ANOVA za spošni linearni regresijski model s k regresorji

Vir variabilnosti	Df	SS	$MS = SS/df$	F
Model	k	SS_{model}	MS_{model}	$MS_{model}/MS_{residual}$
Ostanek (*Residual*)	$n-k-1$	$SS_{residual}$	$MS_{residual}$	
Skupaj	$n-1$	SS_{yy}		

V izpisu povzetka `lm` modela najdemo poleg koeficienta determinacije

$$R^2 = SS_{model}/SS_{yy} = 1 - SS_{residual}/SS_{yy},$$

tudi prilagojeni koeficient determinacije (*Adjusted R-squared*), ki vsebuje tudi informacijo o stopinjah prostosti:

$$R_a^2 = 1 - \frac{\frac{SS_{residual}}{(n-k-1)}}{\frac{SS_{yy}}{(n-1)}} = 1 - \frac{(n-1)\hat{\sigma}^2}{SS_{yy}}. \tag{21}$$

Za koeficient determinacije velja, da se ob vsaki dodani napovedni spremenljivki v modelu njegova vrednost poveča. Za prilagojeni koeficient determinacije to ne velja, ker namesto $SS_{residual}$ v formuli nastopa ocena $\hat{\sigma}^2$. Zato je bolj primeren za primerjavo dveh modelov z različnimi napovednimi spremenljivkami kot navaden R^2. V primerjavi z ostalimi kriteriji za izbiro ustreznega modela, ki jih bomo spoznali v nadaljevanju, je njegova uporaba zastarela.

1.5.4 Območje zaupanja za vse parametre linearnega modela

Ker so ocene parametrov linearnega modela porazdeljene po multivariatni normalni porazdelitvi, lahko pokažemo, da $100(1-\alpha)$ % **območje zaupanja za vse parametre modela hkrati** določimo na podlagi F-statistike:

$$F = \frac{(\mathbf{b}-\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\mathbf{b}-\boldsymbol{\beta})}{(k+1)\hat{\sigma}^2}, \tag{22}$$

ki je porazdeljena po F-porazdelitvi s stopinjami prostosti $SP_1 = k+1$ in $SP_2 = n-k-1$.

$100(1-\alpha)$ % območje zaupanja za $\boldsymbol{\beta}$ predstavlja vse vrednosti za $\boldsymbol{\beta}$, ki ustrezajo pogoju

$$P\left(\frac{(\mathbf{b}-\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\mathbf{b}-\boldsymbol{\beta})}{(k+1)\hat{\sigma}^2} \leq F_\alpha(k+1, n-k-1)\right) = 1-\alpha. \tag{23}$$

1.6 Napovedi linearnega modela

Za vsak y_i, $i = 1, ...n$, imamo vrednosti k napovednih spremenljivk $(x_{i1}, x_{i2}, ..., x_{ik})$. Označimo z $\mathbf{x}_i$ vektor $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ik})^{\mathrm{T}}$ in zapišimo

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i. \tag{24}$$

Zapišimo še napovedano vrednost za odzivno spremenljivko y_* pri vrednostih napovednih spremenljivk $\mathbf{x}_* = (1, x_{*1}, x_{*2}, ..., x_{*k})^{\mathrm{T}}$

$$y_* = \mathbf{x}_*^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_*, \tag{25}$$

napaka ε_* ima pričakovano vrednost 0, varianco σ^2 in je neodvisna od ε_i, $i = 1, 2, , ..., n$. Zanimata nas dva intervala zaupanja, najprej za povprečno napoved $\mathbf{x}_*^{\mathrm{T}}\boldsymbol{\beta}$ in nato še za posamično napoved y_*.

Interval zaupanja za povprečno napoved

Napoved v točki x_* je $x_*^{\mathrm{T}}b$, njena pričakovana vrednost je

$$E(\mathbf{x}_*^{\mathrm{T}}\mathbf{b}) = \mathbf{x}_*^{\mathrm{T}}\boldsymbol{\beta} \tag{26}$$

in njena varianca

$$Var(\mathbf{x}_*^{\mathrm{T}}\mathbf{b}) = \mathbf{x}_*^{\mathrm{T}}Var(\mathbf{b})\mathbf{x}_* = \sigma^2\mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*, \tag{27}$$

$\mathbf{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, kjer je $\mathbf{X}$ je modelska matrika. Ker je napoved $\mathbf{x}_*^{\mathrm{T}}\mathbf{b}$ linearna kombinacija normalno porazdeljenih spremenljivk, tudi zanjo velja, da je porazdeljena normalno

$$\mathbf{x}_*^{\mathrm{T}}\mathbf{b} \sim N(\mathbf{x}_*^{\mathrm{T}}\boldsymbol{\beta}, \quad \sigma^2\mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*). \tag{28}$$

Velja

$$\frac{\mathbf{x}_*^{\mathrm{T}}\mathbf{b} - \mathbf{x}_*^{\mathrm{T}}\boldsymbol{\beta}}{\hat{\sigma}\sqrt{\mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*}} \sim t(SP = n - k - 1), \tag{29}$$

in $(1 - \alpha)100\%$ interval zaupanja za povprečno napoved je

$$\left(\mathbf{x}_*^{\mathrm{T}}\mathbf{b} - t_{1-\frac{\alpha}{2}}(SP = n - k - 1)\hat{\sigma}\sqrt{\mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*}, \quad \mathbf{x}_*^{\mathrm{T}}\mathbf{b} + t_{1-\frac{\alpha}{2}}(SP = n - k - 1)\hat{\sigma}\sqrt{\mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*}\right). \tag{30}$$

Interval zaupanja za posamično napoved

Izrazimo razliko med pravo napovedjo in njeno oceno ter varianco te razlike:

$$y_* - \hat{y}_* = \mathbf{x}_*^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_* - \mathbf{x}_*^{\mathrm{T}}\mathbf{b}. \tag{31}$$

Velja $E(y_* - \hat{y}_*) = 0$ in ε_* in $\mathbf{b}$ sta neodvisna.

$$\begin{aligned} Var(y_* - \hat{y}_*) &= Var(\mathbf{x}_*^{\mathrm{T}}\mathbf{b}) + Var(\varepsilon_*) \\ &= \sigma^2\mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_* + \sigma^2 \\ &= \sigma^2(1 + \mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*). \end{aligned} \tag{32}$$

Tudi tu lahko pokažemo, da je $y_* - \hat{y}_*$ neodvisen od $(n-k-1)\hat{\sigma}^2/\sigma^2$ in velja

$$\frac{y_* - \hat{y}_*}{\hat{\sigma}\sqrt{1 + \mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*}} \sim t(SP = n-k-1), \tag{33}$$

in $(1-\alpha)100\%$ interval zaupanja za posamično napoved je

$$\left(\hat{y}_* - t_{1-\frac{\alpha}{2}}(SP = n-k-1)\hat{\sigma}\sqrt{1 + \mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*},\ \hat{y}_* + t_{1-\frac{\alpha}{2}}(SP = n-k-1)\hat{\sigma}\sqrt{1 + \mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*}\right) \tag{34}$$

1.7 Interpretacija ocen parametrov linearnega modela z več regresorji

Interpretacijo ocen parametrov linearnega modela z več regresorji si poglejmo najprej na primeru dveh številskih regresorjev x_1 in x_2. Zamislimo si pričakovano vrednost tega modela v točki (x_{01}, x_{02}).

$$E(y|x_{01}, x_{02}) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02}, \tag{35}$$

in v točki $(x_{01}, x_{02} + 1)$, kar pomeni, da se pri spremenljivki x_2 premaknemo za eno enoto naprej

$$E(y|x_{01}, x_{02} + 1) = \beta_0 + \beta_1 x_{01} + \beta_2(x_{02} + 1) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \beta_2, \tag{36}$$

iz (35) in (36) sledi

$$E(y|x_{01}, x_{02} + 1) - E(y|x_{01}, x_{02}) = \beta_2. \tag{37}$$

Torej velja, če x_{02} povečamo za eno enoto in ostane izbrana vrednost x_{01} nespremenjena, se pričakovana vrednost y poveča za β_2.

V linearnem modelu z več regresorji ima vsak regresor "pogojni vpliv": če regresor x_j povečamo za eno enoto, se pogojno na konstantne vrednosti vseh ostalih regresorjev v modelu pričakovana vrednost odzivne spremenljivke poveča za β_j enot.

Pogojni vpliv regresorja x_j v modelu z več regresorji je lahko zelo drugačen, kot je njegov "robni" vpliv na odzivno spremenljivko, ko je x_j edini regresor v modelu. Prisotnost ostalih regresorjev lahko povzroči spremembo velikosti, lahko pa tudi spremembo predznaka parametra β_j.

Geometrijska predstavitev enostavne linearne regresije je premica v dvodimenzionalnem prostoru, za model z dvema regresorjema je ravnina v tridimenzionalnem prostoru, za model s k regresorji pa je to hiper ravnina v $k + 1$ dimenzionalnem prostoru.

V regresijski analizi pogosto modeliramo vpliv izbrane spremenljivke na odzivno spremenljivko ob upoštevanju (*controlling for*) določenih t. i. **motečih spremenljivk** (*confounding variables*) v modelu. Zanima nas vpliv te izbrane napovedne spremenljivke, vendar vemo, da je odzivna spremenljivka odvisna tudi od nekaterih drugih spremenljivk, ki pa niso predmet naše raziskave.