Peatükk 8

Mitteparameetriline regressioon

Sageli pakub huvi sõltuva tunnuse Y ja sõltumatu tunnuse X vahelise seose kirjeldamine,

$$Y = g(X) + \varepsilon,\ \mathrm{E}\varepsilon = 0.$$

Parameetrilise regressiooni korral eeldatakse enamasti, et seost iseloomustav funktsioon g sõltub vaid paarist tundmatust parameetrist (näiteks $g(X) = c_0 + c_1 X$, kus c_0 ja c_1 on tundmatud parameetrid) ja seose kirjeldamiseks piisab vaid nende parameetrite hindamisest. Mitteparameetrilise statistika kursuses muidugi sellist eeldust teha ei saa — enamasti ju pole teada, milline on kahe tunnuse vaheline seos tegelikult.

8.1 Nadaraya-Watson'i regressioon (tuumaregressioon)

Üks võimalus kirjeldada seost tunnuste X ja Y vahel on märgata, et $\mathrm{E}(Y|X = x) = g(x)$. Seega otsitakse Y-tunnuse tinglikku keskväärtust tingimusel, et $X = x$ on antud. Tingliku keskväärtuse saab aga avaldada kujul:

$$\begin{aligned} \mathrm{E}(Y|X = x) &= \int_{-\infty}^{\infty} y f_{Y|X=x}(y)dy \\ &= \int_{-\infty}^{\infty} \frac{y f_{Y,X}(y,x)}{f_X(x)}dy, \end{aligned} \tag{8.1}$$

kus $f_{Y|X=x}(y)$ on Y-tunnuse tihedus juhul kui $X = x$ ehk Y-tunnuse tinglik tihedusfunktsioon. Samas oskame juba väga hästi hinnata tihedusfunktsioo-

ne, seega nii $f_{Y,X}$ kui f_X hindamine tuumameetodil peaks olema väga lihtne ülesanne. Tiheduse f_X hinnanguks saame

$$\hat{f}_X(x) = \frac{1}{nh_x} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}\right)$$

ja kasutades kahemõõtmelise tiheduse hindamisel tuumafunktsioonina funktsiooni $K(x, y) = K(x)K(y)$ saame X ja Y-tunnuse ühise tihedusfunktsiooni hinnanguks

$$\hat{f_{X,Y}}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}\right) K\left(\frac{y - y_i}{h_y}\right).$$

Kui pistame saadud tihedusfunktsiooni hinnangud sisse tingliku keskväärtuse leidmiseks mõeldud valemisse 8.1, saame valemi tingliku keskväärtuse hindamiseks:

$$\begin{aligned}
\widehat{\mathrm{E}(Y|X=x)} &= \int_{-\infty}^{\infty} \frac{y\hat{f}_{Y,X}(y,x)}{\hat{f}_X(x)} dy \\
&= \int_{-\infty}^{\infty} \frac{y \frac{1}{nh_x h_y} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right) K\left(\frac{y-y_i}{h_y}\right)}{\frac{1}{nh_x} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right)} dy \\
&= \frac{1}{h_y \sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right)} \int_{-\infty}^{\infty} y \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}\right) K\left(\frac{y - y_i}{h_y}\right) dy \\
&= \frac{1}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right)} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}\right) \int_{-\infty}^{\infty} yK\left(\frac{y - y_i}{h_y}\right) / h_y \, dy \\
&= \frac{1}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right)} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}\right) y_i
\end{aligned}$$

viimane võrdustest kehtib sellepärast, et $K\left(\frac{y-y_i}{h_y}\right)/h_y$ on keskväärtusega y_i juhusliku suuruse tihedusfunktsioon (vaata näiteks tihedusfunktsiooni hindamist käsitlenud peatükis valemile 7.1 eelnenud ja järgnenud arutluskäiku).

Proovime saadud valemit veidi paremini mõista. Vaatame esmalt, milliseks muutub valem siis, kui tuumafunktsiooniks $K(x)$ on ühtlase jaotuse U(-0,5; 0,5) tihedusfunktsioon. Sellisel juhul

$$\cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}\right) = \#\{x_i | \text{kaugus}(x_i, x) < h_x\}$$

ehk tegemist on punktist x kaugusel h_x asuvate vaatluste koguarvuga. Aga

$$\cdot \sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right) y_i = \sum_{\text{kaugus}(x_i,x)<h_x} y_i$$

ehk tegemist on nende vaatluste y-tunnuse väärtuste summaga, mille puhul kaugus x_i ja x vahel on väiksem kui h_x. Seega saame tuumafunktsioonina ühtlast jaotust kasutades tingliku keskväärtuse hinnanguks punktis $X = x$ sellest punktist kuni kaugusel h_x paiknevate vaatluste y-tunnuse väärtuste aritmeetilise keskmise.

Suvalise tuumafunktsiooni korral võime defineerida kaalud w_i:

$$w_i := \frac{K\left(\frac{x-x_i}{h_x}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right)},$$

Kasutades defineeritud kaalusid saab tingliku keskväärtuse hinnangu kirja panna kujul

$$\mathrm{E}(Y|X=x) = \sum w_i y_i,$$

ehk tegemist on kaalutud keskmisega, kusjuures vaatluse (objekti) kaal on seda suurem, mida lähemal on antud objekti X-tunnuse väärtus (x_i) kohale x. Mõistmaks paremini kasutatava kaalu tähendust, võime kaalu kirja panna järgmisel kujul:

$$w_i = \frac{K\left(\frac{x-x_i}{h_x}\right)/h_x}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right)/h_x},$$

Ehk kaalud saame, kui nihutame tuumafunktsioonina kasutatud tiheduse punkti x (ja venitame teda "venitusteguriga" h_x). Selline keskväärtusega x tihedusfunktsioon on (kasutame ka tuumafunktsiooni sümmeetrilisust):

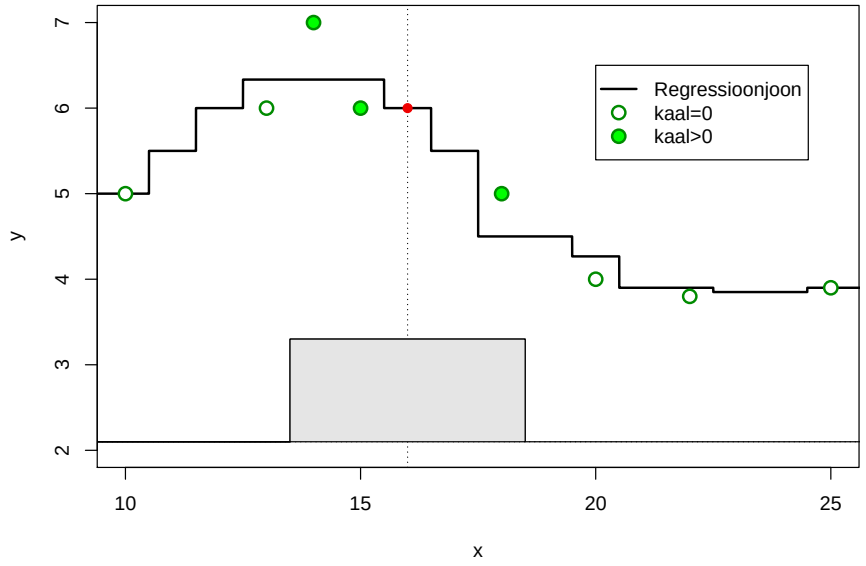
$$K\left(\frac{x_i-x}{h_x}\right)/h_x = K\left(\frac{x-x_i}{h_x}\right)/h_x.$$

Kaalud saame tihedusfunktsiooni väärtuseid normeerides selliselt, et kõigi kaalude summa oleks 1, $\sum w_i = 1$.

Kokkuvõtteks: kui soovime leida regressioonsirge paiknemist punktis x, siis nihutame punkti x kohale mingi sümmeetrilise tihedusfunktsiooni (nn tuumafunktsiooni). Iga vaatlus saab seejärel kaalu, mis on võrdeline sellise punkti x nihutatud tihedusfunktsiooni väärtusega. Tüüpiliselt kasutatakse

tuumafunktsioonina selliseid tihedusfunktsioone, mille puhul tihedusfunktsiooni väärtus on seda väiksem, mida kaugemale oleme jõudnud keskväärtusest (ehk punktist x). Seega omavad punktist x kaugemal paiknevad vaatlused väiksemat mõju regressioonsirge hindamisel antud kohas kui punktile x lähedal paiknevad vaatlused. Vaata ka joonist 8.1.

Joonis 8.1: Nadaraya-Watsoni regressioonsirge ja tema arvutuskäik kohas $x = 16$. Tuumafunktsioonina on kasutatud ühtlast jaotust U(-2.5, 2.5).



8.1.1 Silumisparameetri valik

Saadav regressioonsirge sõltub märkimisväärselt silumisparameetri h_x valikust. Kuidas leida sobivaimat h_x väärtust? Tavaliselt on regressioonjoone leidmise eesmärgiks tulevaste, uute vaatluste korral prognoosida tunnuse x väärtuse abil tunnuse y väärtust. Sellisel juhul võiksime muidugi valida silumiskordaja h_x väärtuse selliselt, et uute vaatluste korral y tunnuse prognoos tuleks võimalikult täpne. Kuidas seda teha? Tavapärane on kasutada ristvalideerimist: valimist eemaldatakse üks vaatlustest ja leitakse regressioonjoone hinnang ilma antud vaatluseta. Seejärel leitakse valimist eemaldatud vaatluse jaoks prognoos kasutades ülejäänud vaatluste põhjal hinnatud regres-

sioonjoont. Samasugust protseduuri korratakse kõigi vaatlustega ja leitakse prognooside ruutvigade keskmine. Valitakse välja selline h_x väärtus, mille puhul prognooside ruutvigade keskmine tuleb kõige väiksem:

$$h_x = argmin \sum_{i=1}^{n} \left(y_i - \hat{g}_{-i}(x_i)\right)^2,$$

kus $\hat{g}_{-i}(x_i)$ on ilma i. vaatlust kasutamata leitud regressioonjoone väärtus kohas x_i.

8.1.2 Servaefekt

Nadaraya-Watsoni regressioon ei ole paraku probleemideta imevahend. Kuna regressioonjoone väärtus mingis punktis on olemasolevate vaatluste (kaalutud) keskmine, siis ei saa regressioonjoon muutuda kunagi väiksemaks kui valimi miinimum või suuremaks kui valimi maksimum. Kui näiteks andmetes esineb selge tõusev trend siis suurimate valimis esinevate x tunnuse väärtuste korral regressioonjoone leidmisel kaasatakse küll ka suurimad y-tunnuse väärtused, aga mingi kaaluga ka veidi väiksematele x-tunnuse väärtustele vastavad veidi väiksemad y-tunnuse väärtused. Seega võib regressioonsirge käitumine olla piirkondades kus vaatlused otsa lõppevad mitterahuldav, vaata ka joonist 8.2.

Joonis 8.2: Servaefekt Nadaraya-Watsoni regressiooni korral.