Statističko učenje

Uvodno predavanje

listopad, 2021

Model

Pretpostavljamo da je $Y = f(X) + \epsilon$ pri čemu je

- $f : \mathbb{R} \to \mathbb{R}$ neka funkcija;
- $X \sim \mathrm{Unif}[1, 4]$;
- $\epsilon \sim \mathrm{N}(0, \sigma^2)$ za neki $\sigma > 0$;
- X i ϵ su nezavisne

Napomena: Uz ove pretpostavke, f je upravo regresijska funkcija, a ireducibilna greška je

$$\mathbb{E}[(Y - f(X))^2 \mid X = x] = \mathbb{E}[\epsilon^2] = \sigma^2 \,, \ \forall x \in \mathbb{R} \,.$$

Simulacije

Za različite kombinacije funkcije f i varijance σ^2, simulirat ćemo skup za učenje $\tau = \{(x^{(i)}, y_i) : i = 1, \ldots, n\}$ uz $n = 100$ te promatrati kako se ponašaju greška na skupu za učenje

$$L_\tau(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x^{(i)}))^2 \,,$$

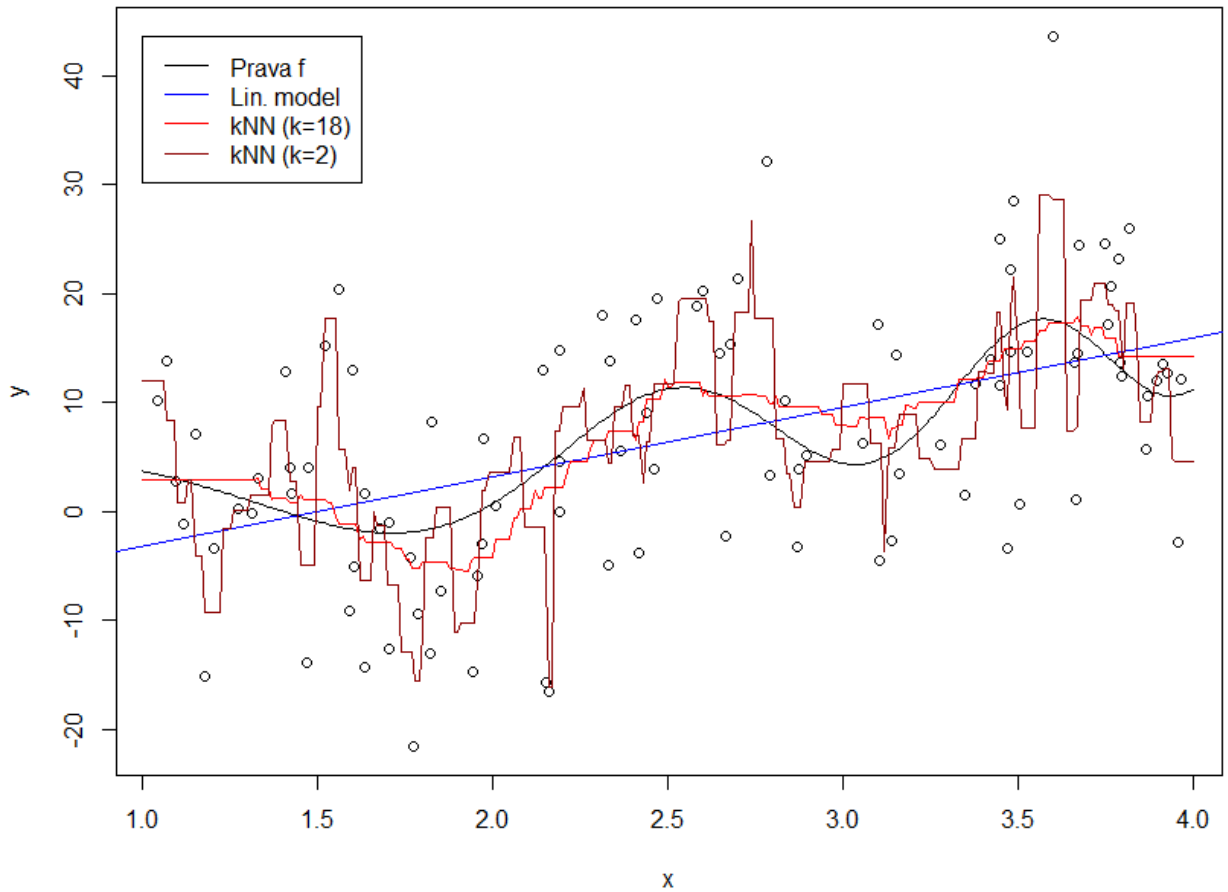
i testna greška

$$L(\hat{f}) = \mathbb{E}[(Y - \hat{f}(X))^2] \,,$$

u (i) linearnoj regresiji, te (ii) kNN metodi za različite k. Testnu grešku $L(\hat{f})$ procijenit ćemo na simuliranom testnom skupu veličine $m = 10000$.

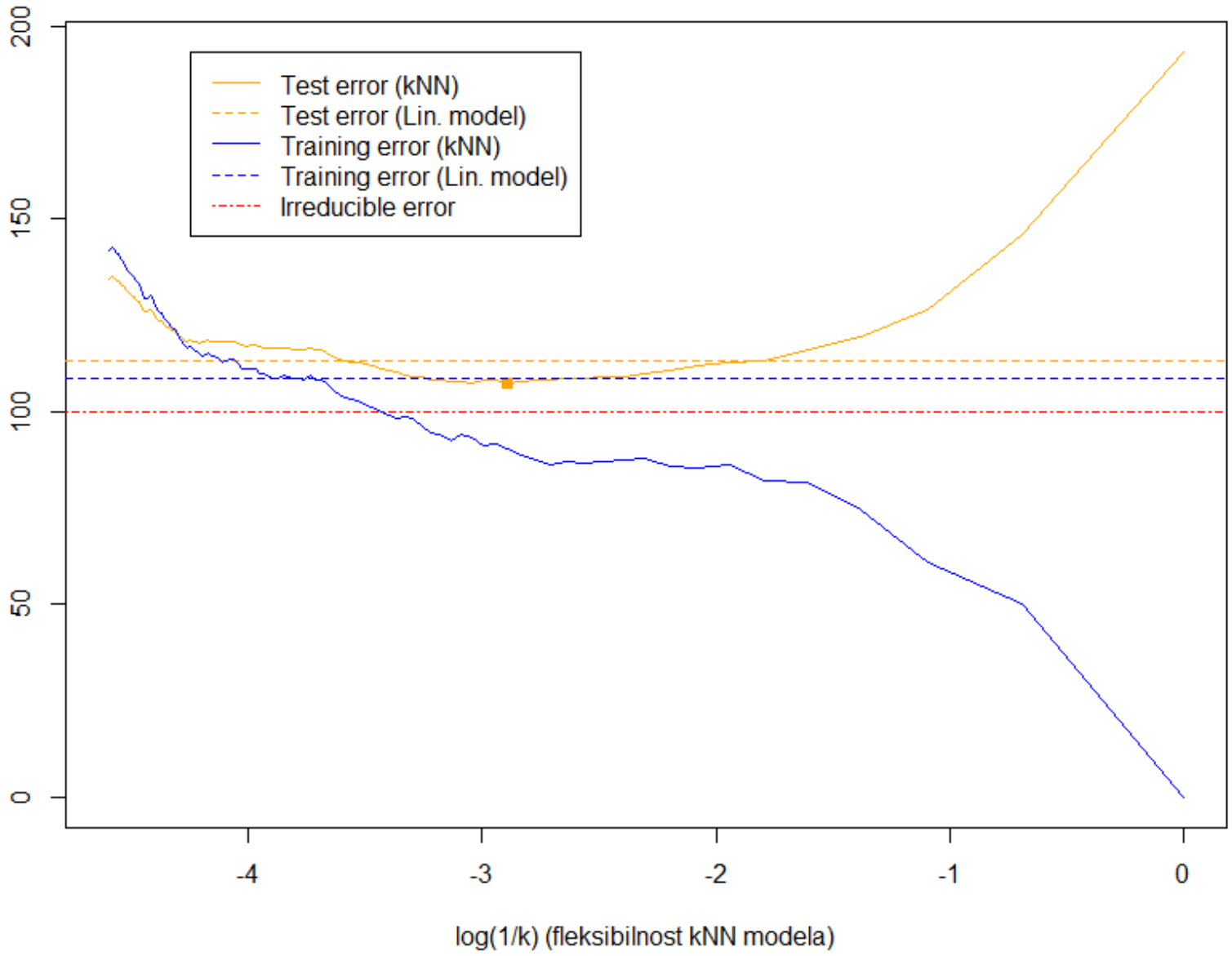
Napomena: U kNN metodi, parametar k kontrolira fleksibilnost modela – povećavanjem k smanjuje se fleksibilnost. Zbog toga ćemo rezultate prikazivati u ovisnosti o $1/k$ i to na log-skali.

Primjer 1 – f nije linearna

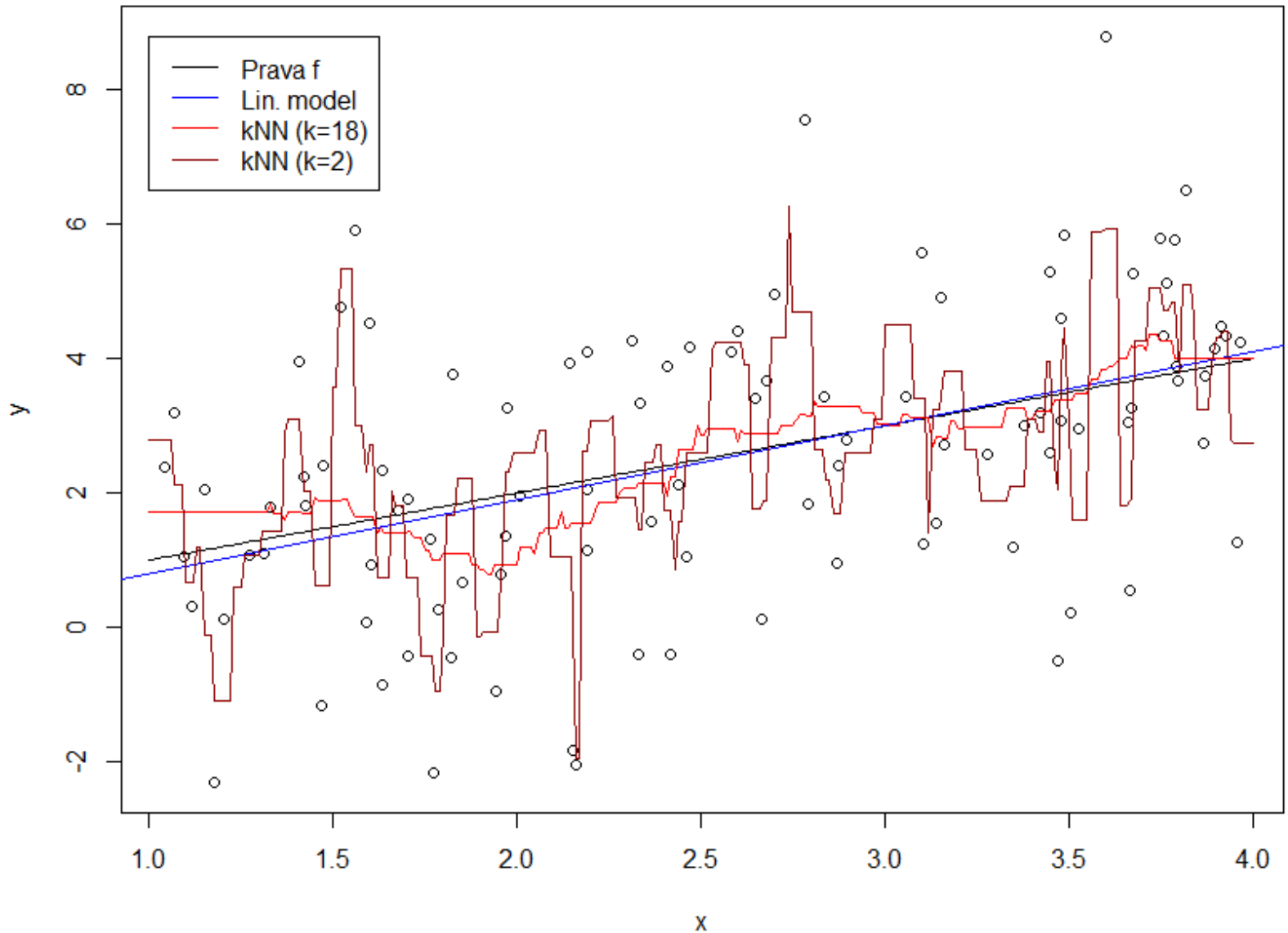


Slika: U slučaju $k = 2$, kNN procjena se "previše" prilagođava podacima – tzv. overfitting. S druge strane, lin. model nije dovoljno fleksibilan da bi dobro procijenio f.

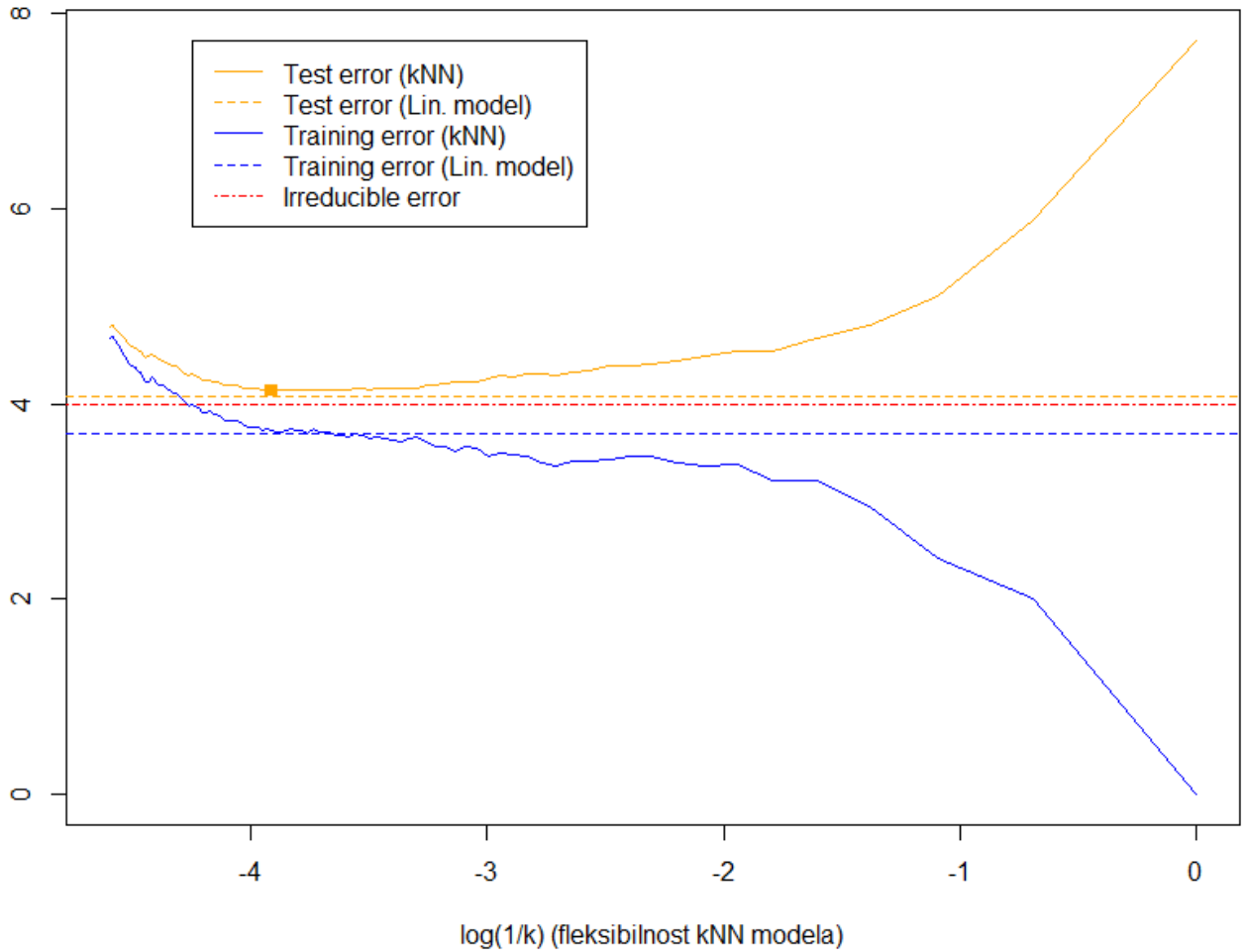
Primjer 1 – "U-oblik" testne greške



Primjer 2 – f je linearna

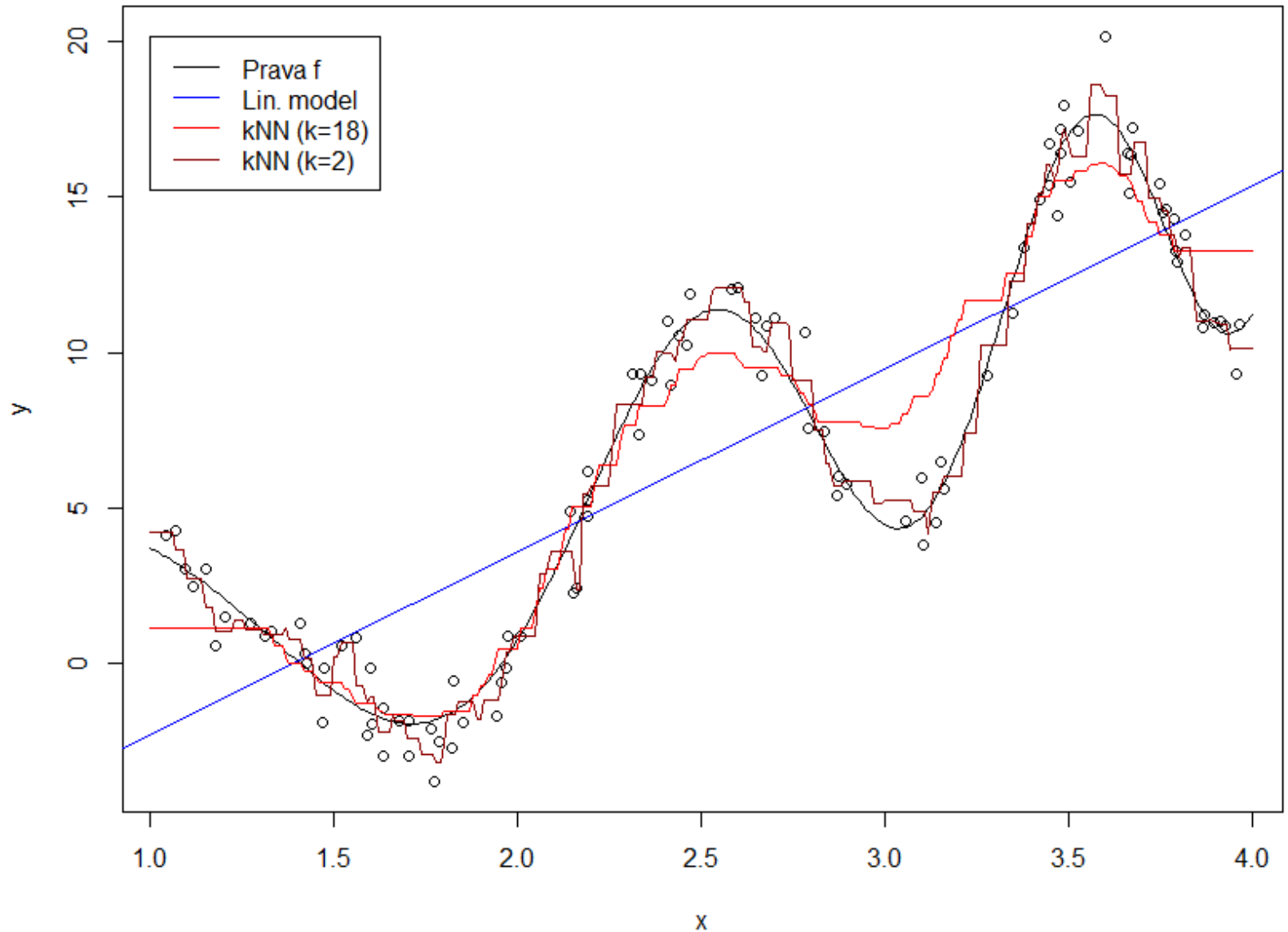


Primjer 2

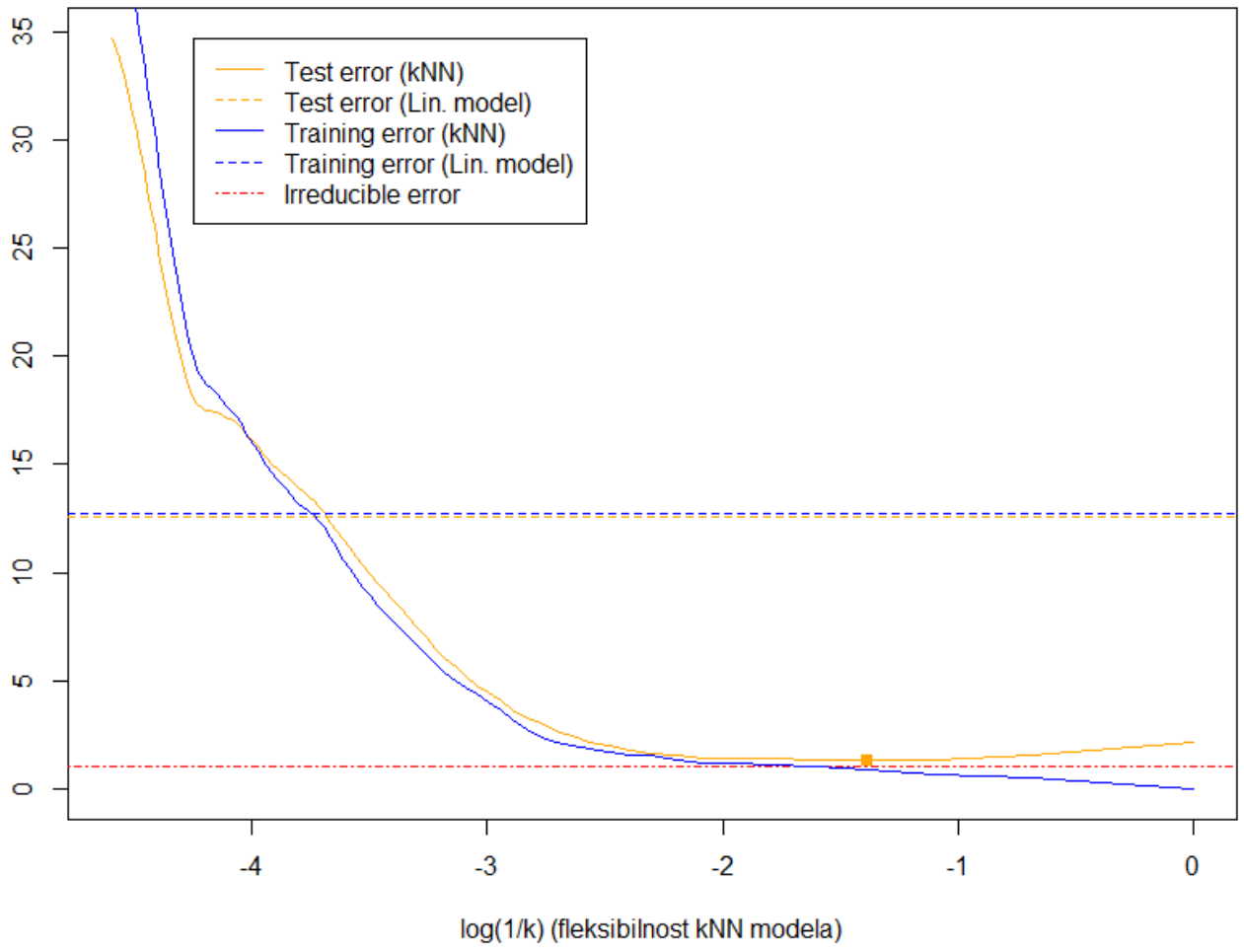


Slika: Budući da je f linearna, linearni model i manje fleksibilan kNN model daju skoro pa optimalan rezultat.

Primjer 3 – f nije linearna i varijanca je mala



Primjer 3

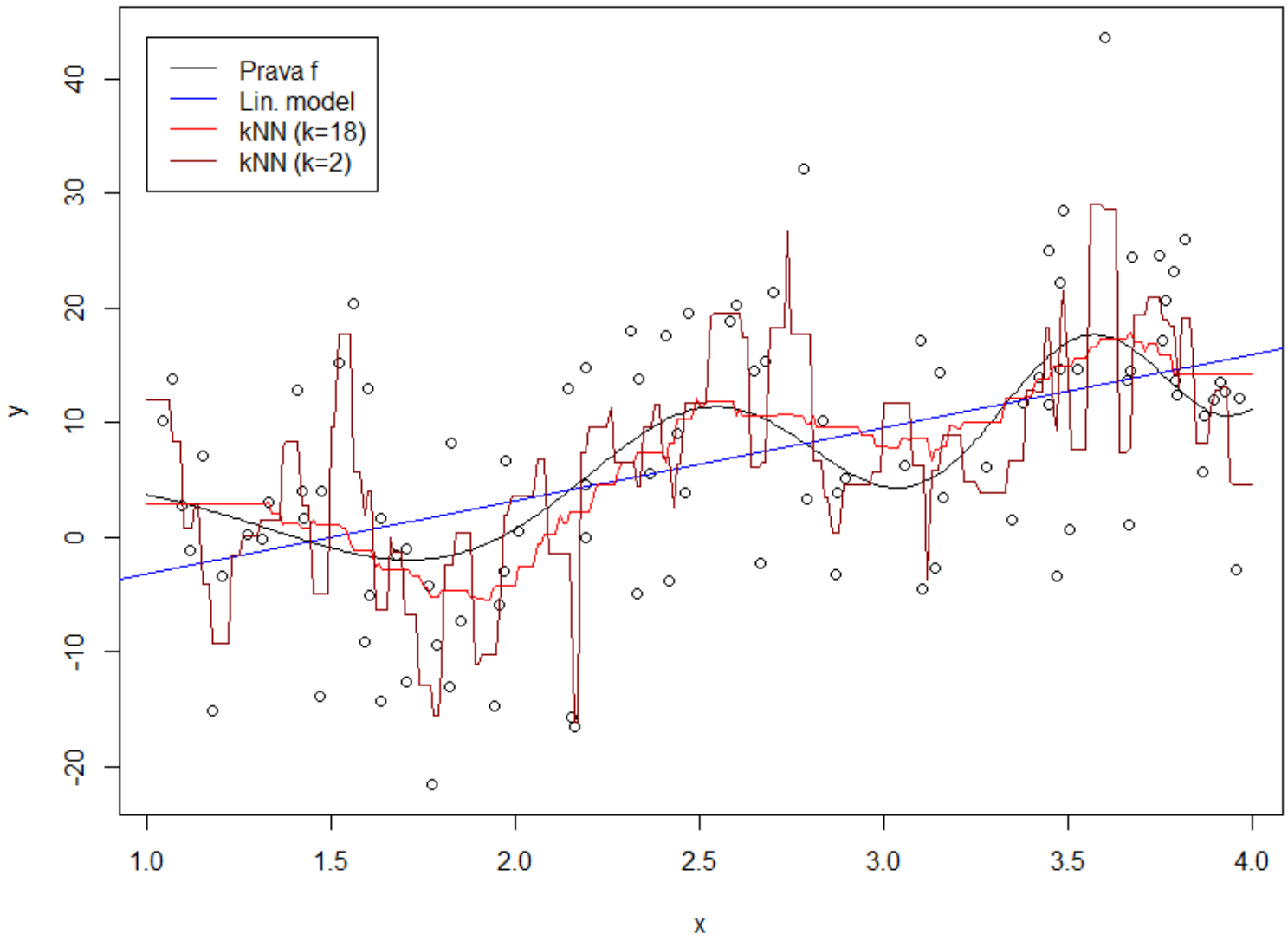


Slika: Jako fleksibilan kNN model daje najbolje rezultate.

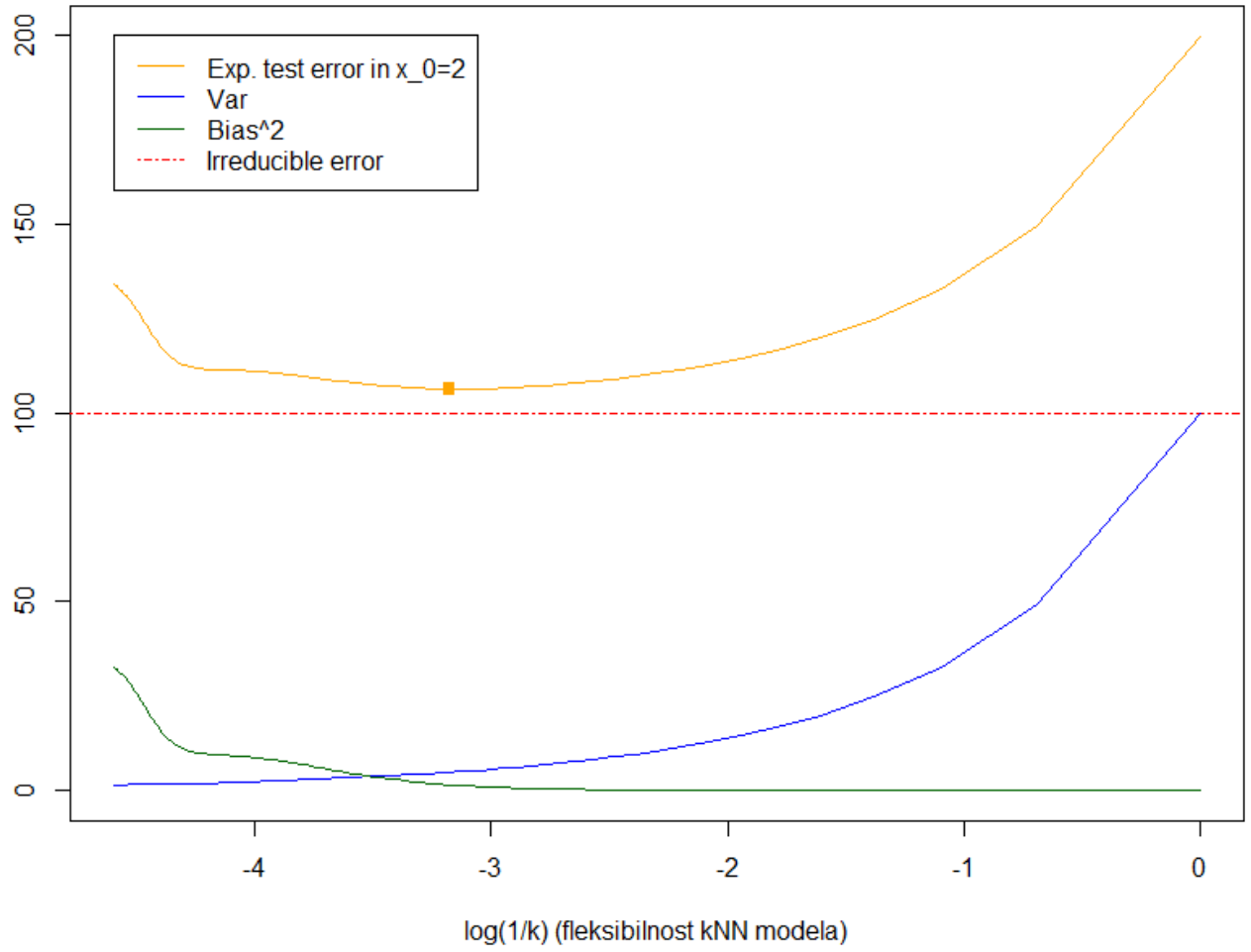
Napomena – "curse of dimensionality"

Kako se povećava broj kovarijata p, kNN metoda brzo postaje inferiorna u usporedbi sa jednostavnim linearnim modelom i u slučajevima kada je stvarna regresijska funkcija f daleko od linearne $\rightsquigarrow$ vidi Figure 3.20 u ISLR.

Odnos između pristranosti i varijance



Odnos između pristranosti i varijance (kNN)



Slika: Mala fleksibilnost daje "veliku pristranost i malu varijancu", a velika fleksibilnost "malu pristranost i veliku varijancu".