Statističko učenje

Uvodno predavanje

listopad, 2022

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

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

kNN vs linearna regresija

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

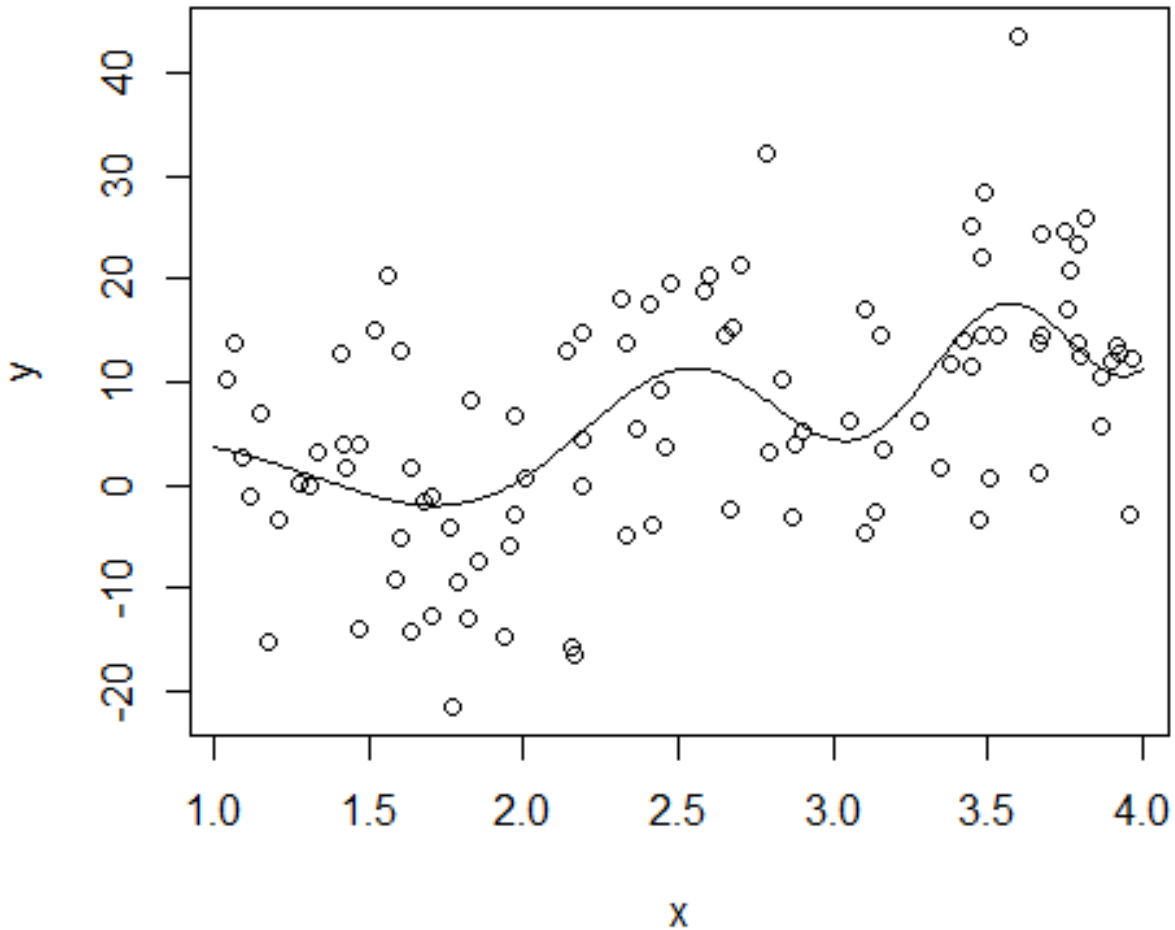
$$L_\tau(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x^{(i)}))^2 \,,$$

i testna greška

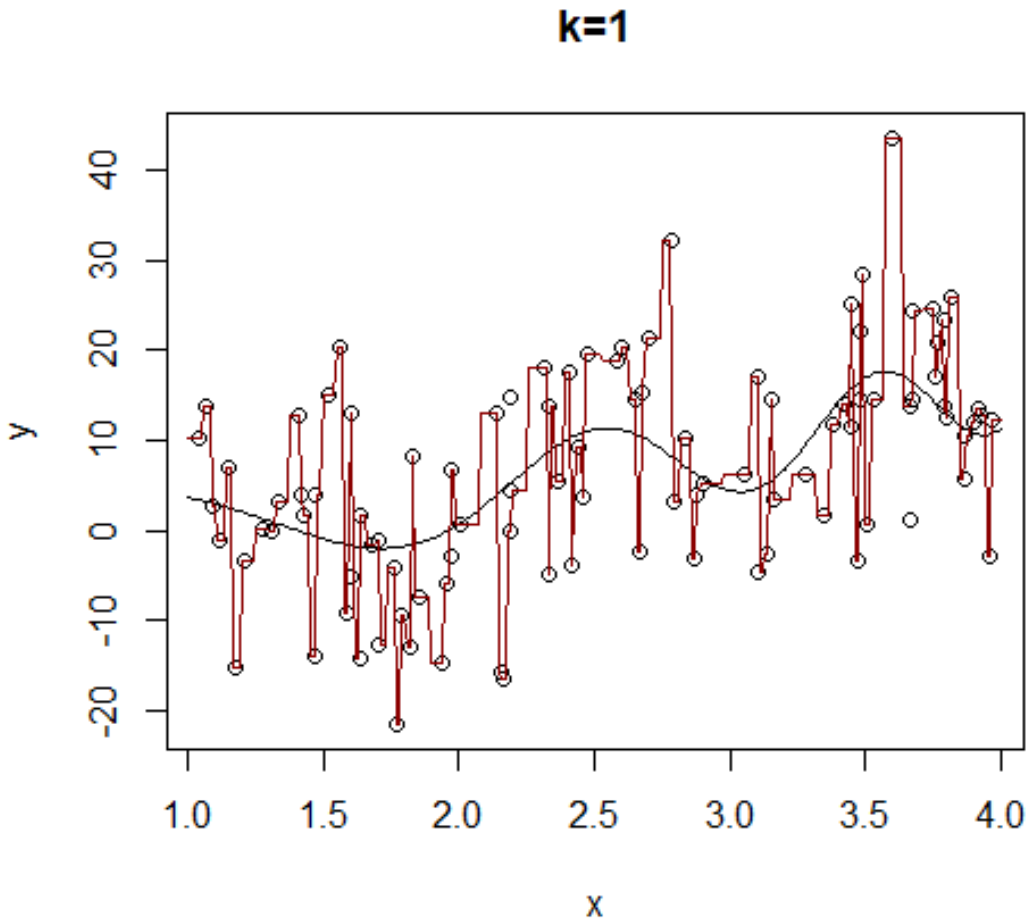
$$L(\hat{f}) = \mathbb{E}[(Y - \hat{f}(X))^2] \,,$$

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

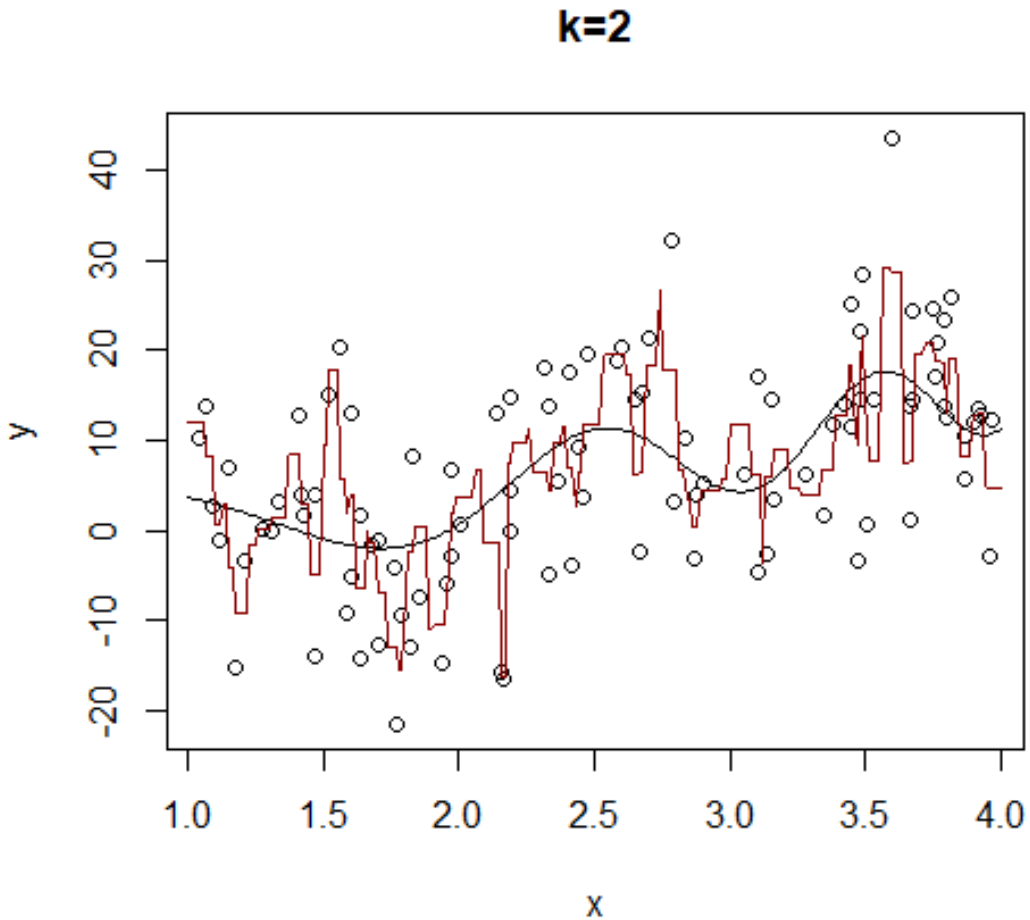
Primjer 1 – f nije linearna



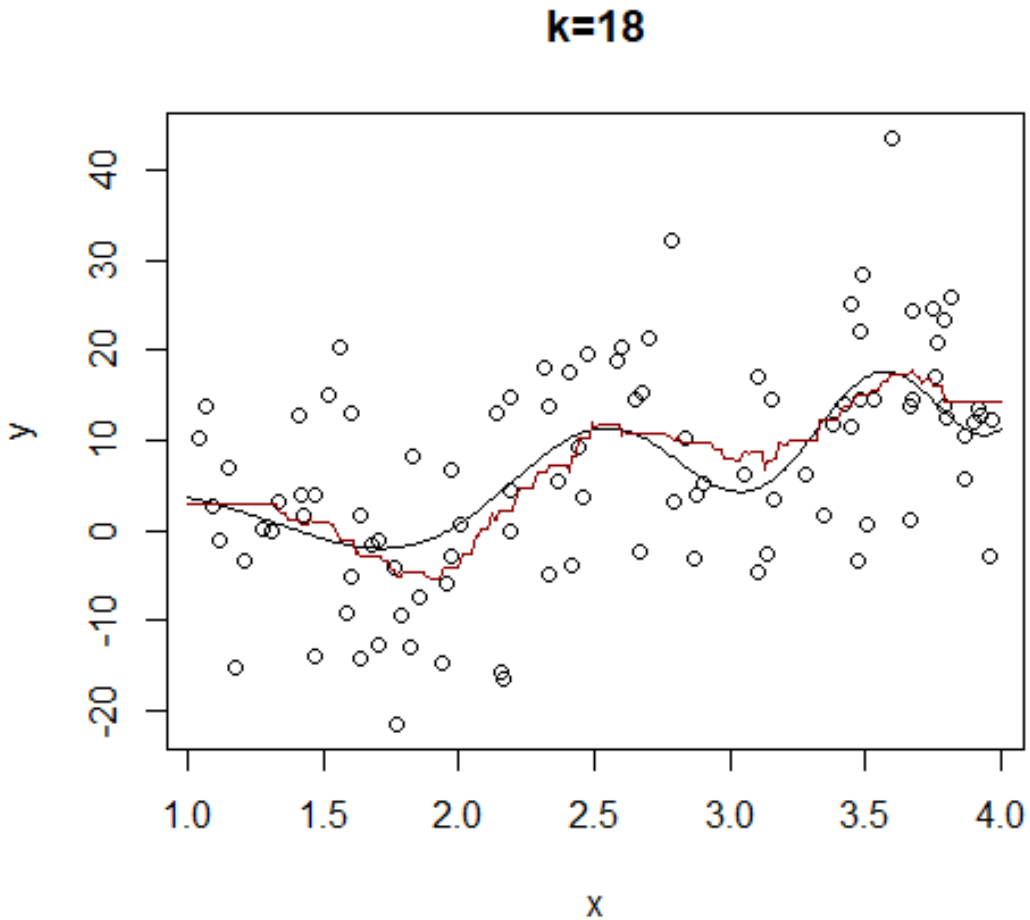
kNN metoda



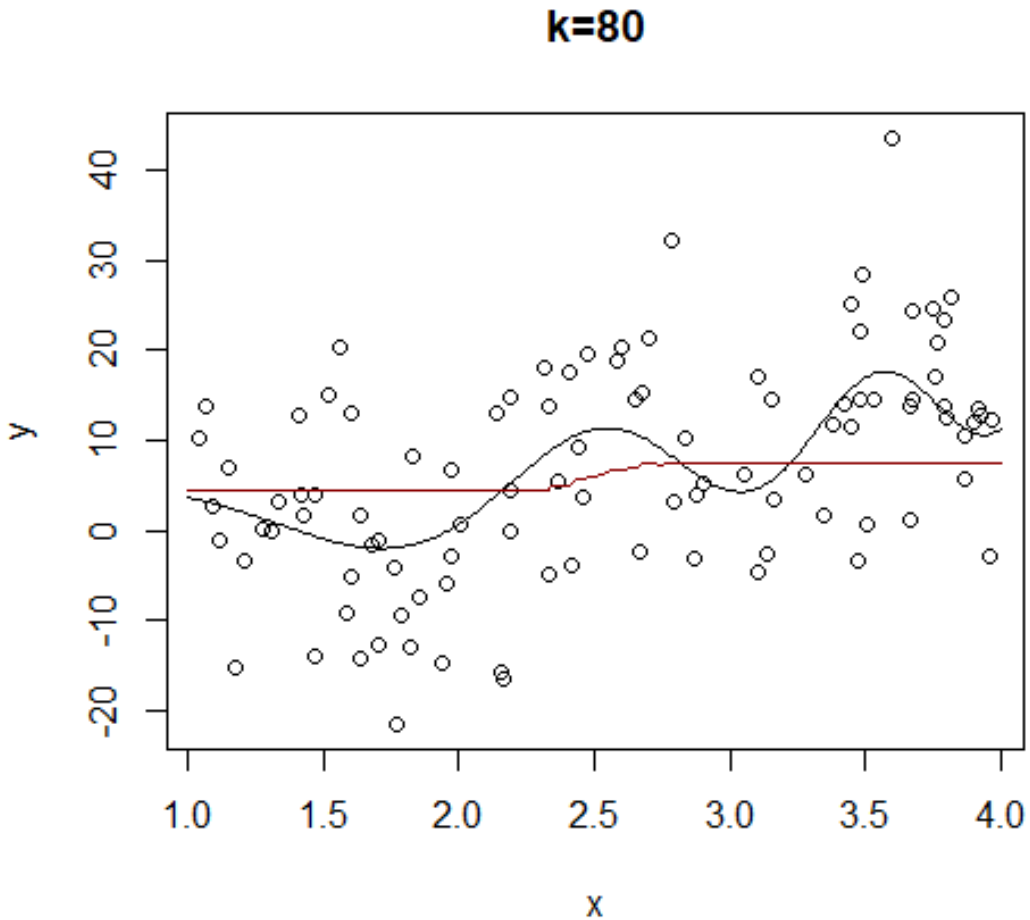
kNN metoda



kNN metoda

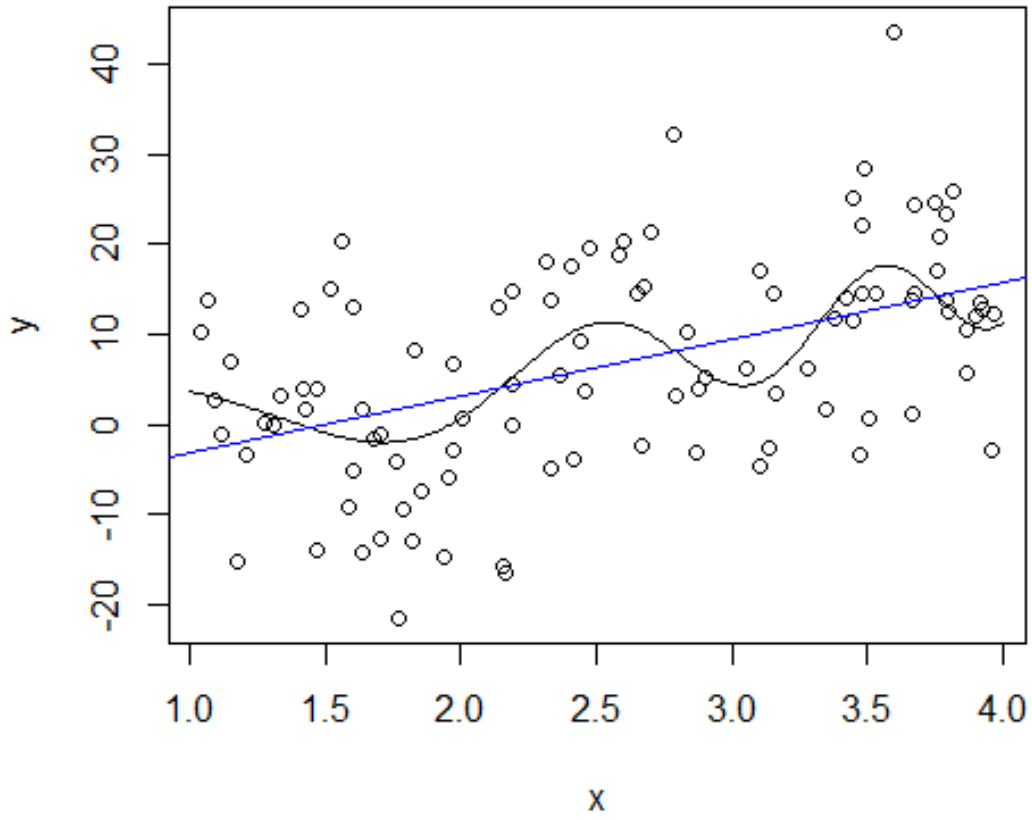


kNN metoda



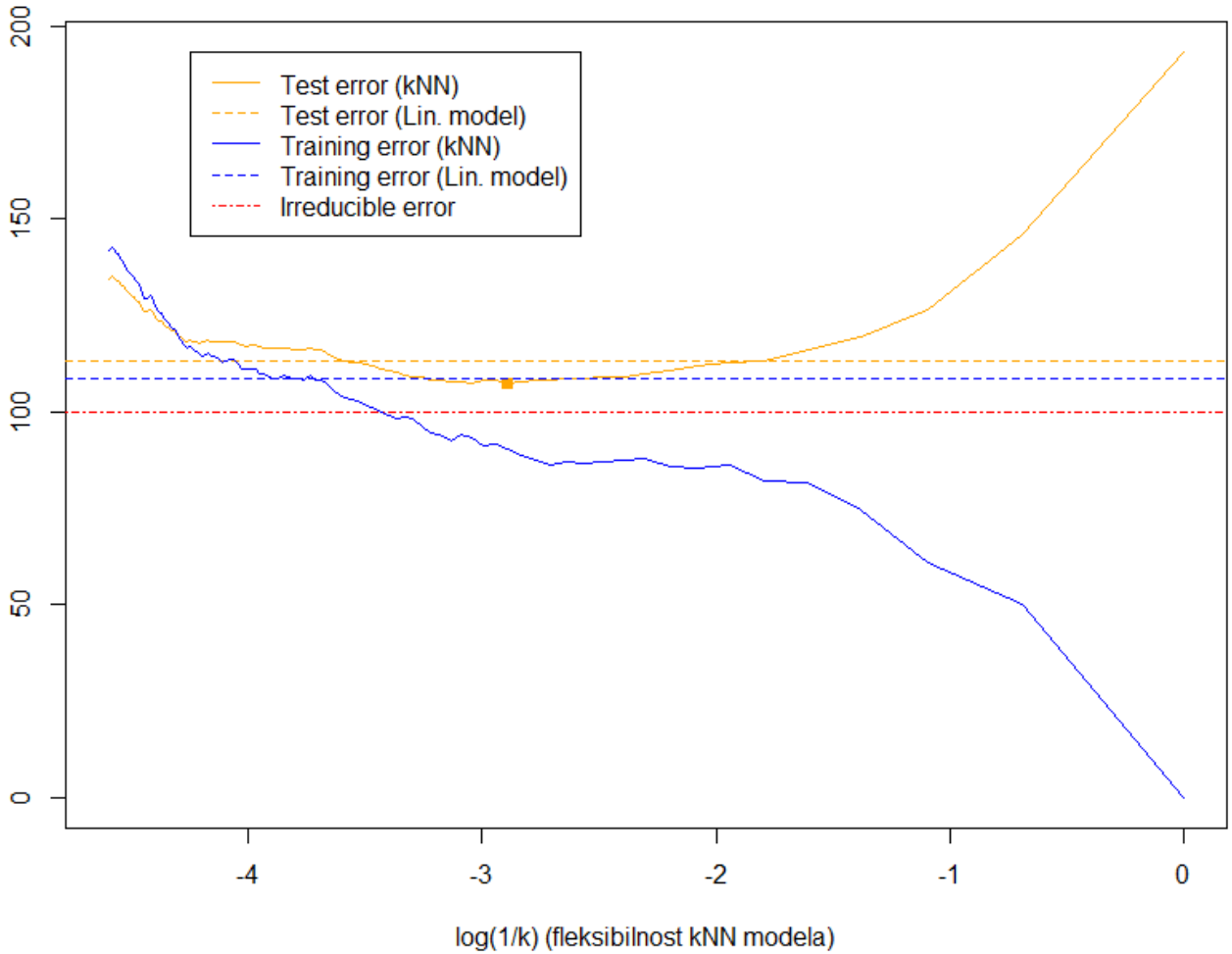
Parametar k kontrolira kompleksnost (fleksibilnost) metode – povećavanjem k **smanjuje** se kompleksnost. Za premali k kNN procjena se "previše" prilagođava podacima – tzv. overfitting, dok se suprotno događa za preveliki k – tzv. underfitting.

Linearna regresija pomoću najmanjih kvadrata



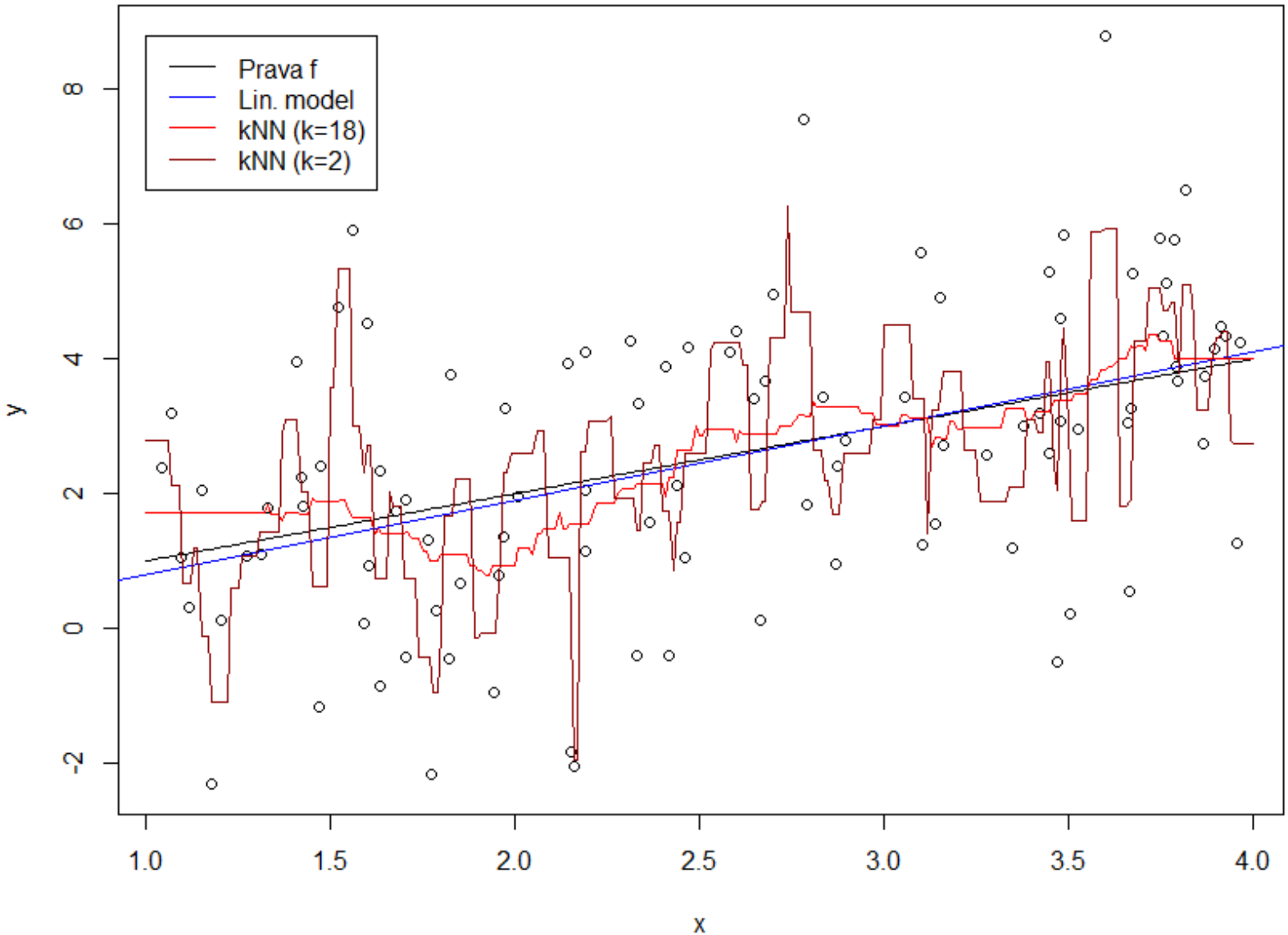
Slika: Lin. model je nedovoljno fleksibilan da bi dobro procijenio f. Ipak, budući da je ireducibilna greška relativno velika, relativno na nju **greška** linearnog modela neće biti loša.

Primjer 1 – testna vs trening greška

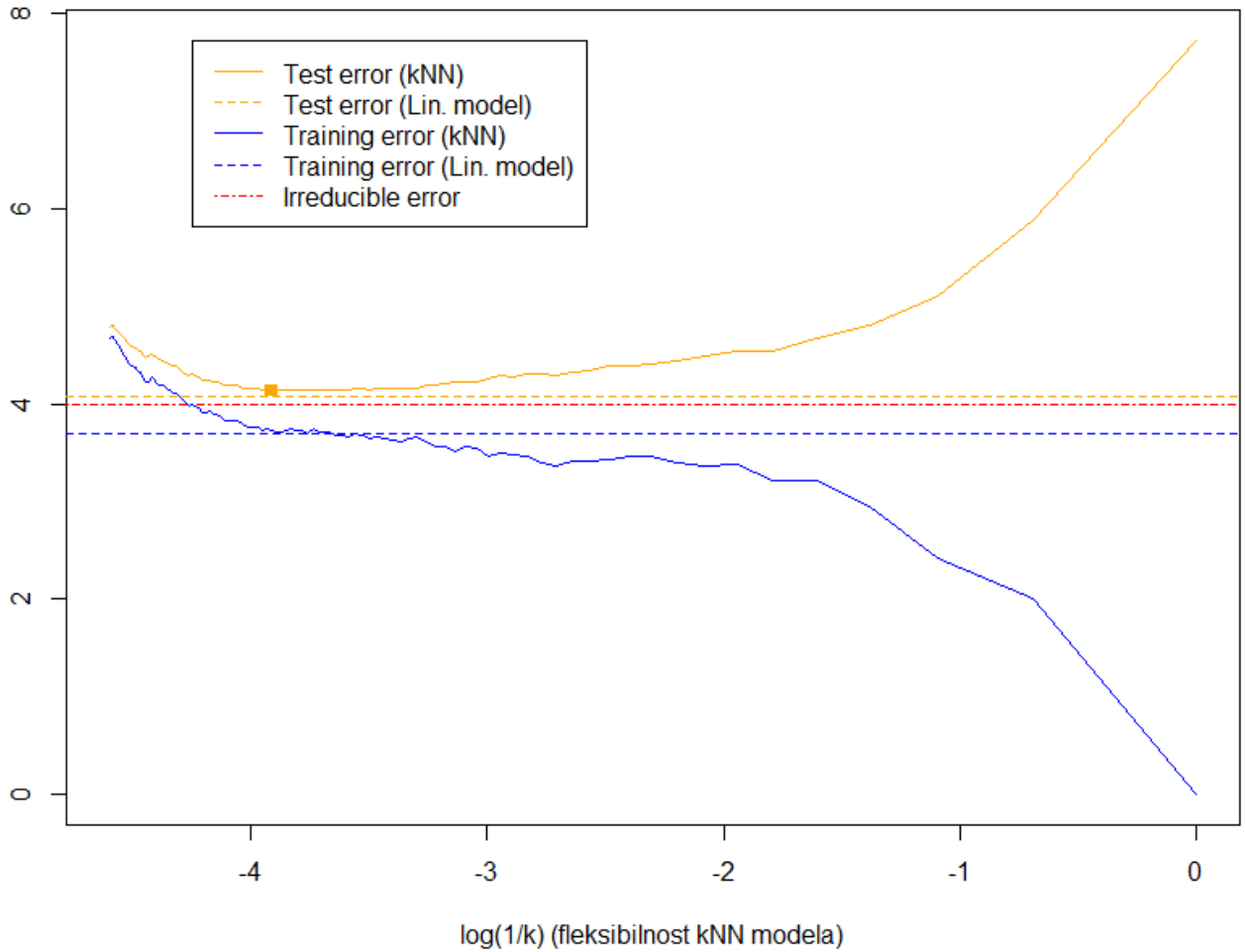


Slika: Greška na skupu za trening τ uglavnom neće biti dobra procjena za stvarnu grešku jer $\hat{f}$ konstruiramo na temelju τ.

Primjer 2 – f je linearna

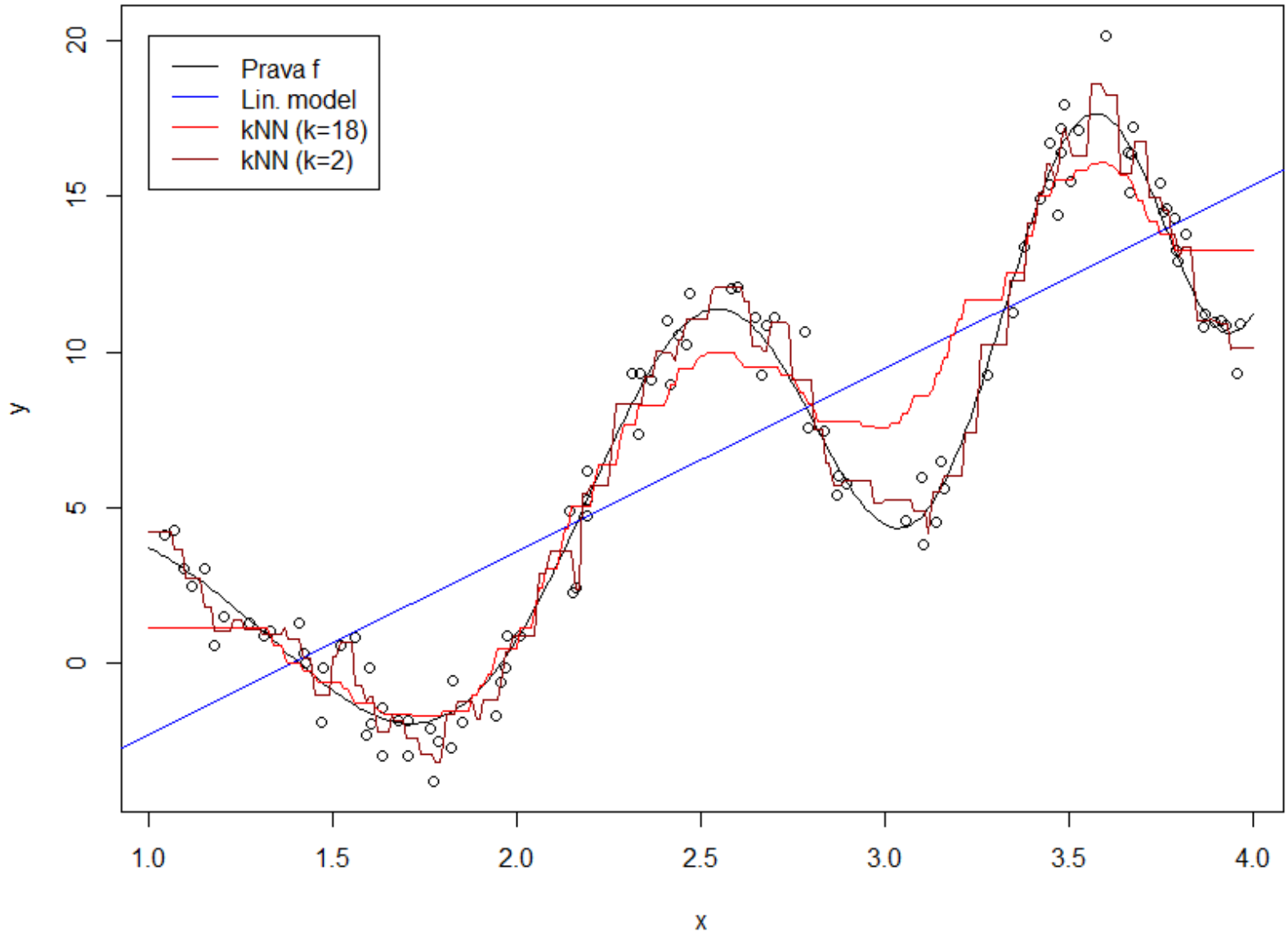


Primjer 2

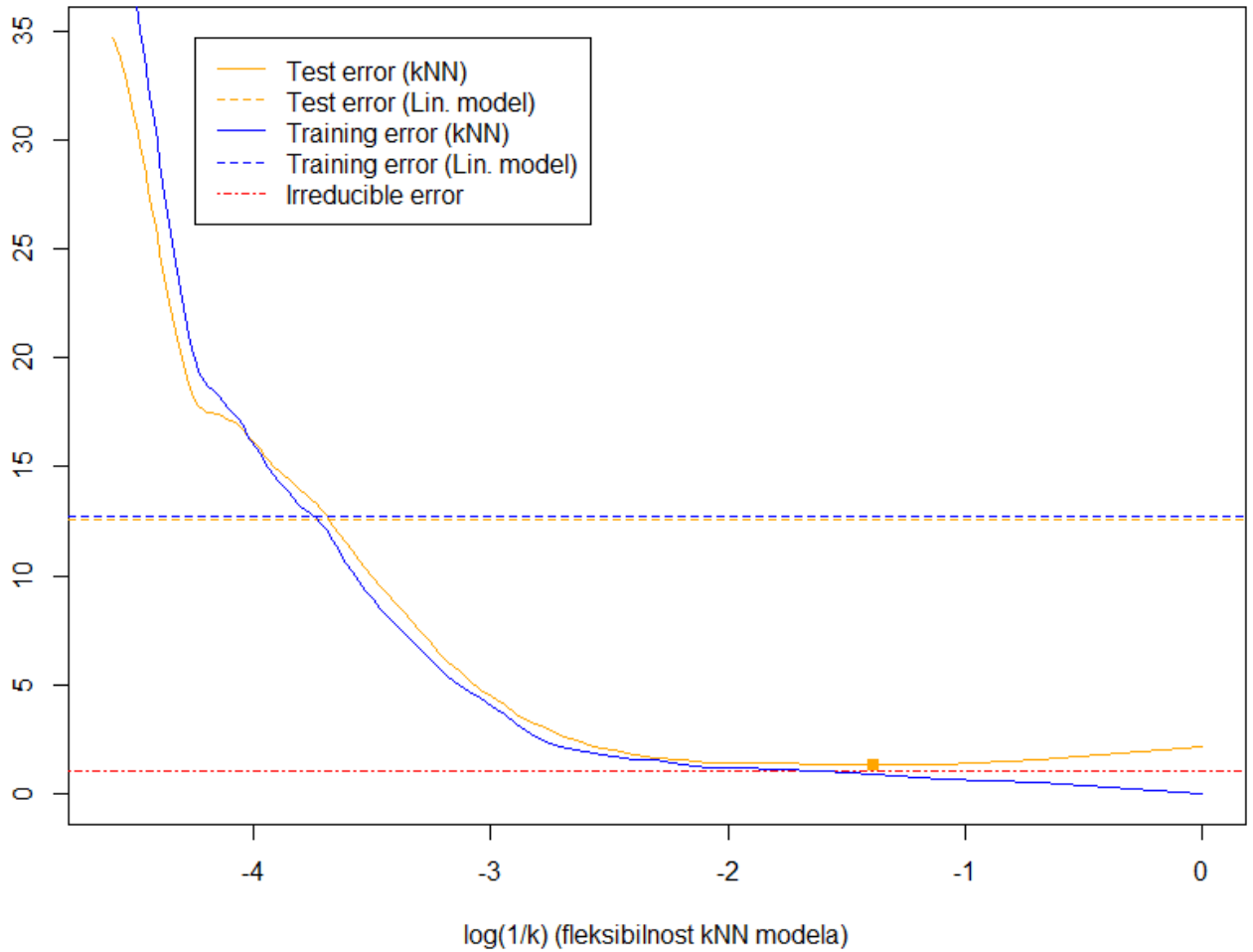


Slika: Budući da je f linearna, linearni model i manje fleksibilan kNN model daju skoro pa optimalan rezultat.

Primjer 3 – f nije linearna i varijanca je mala

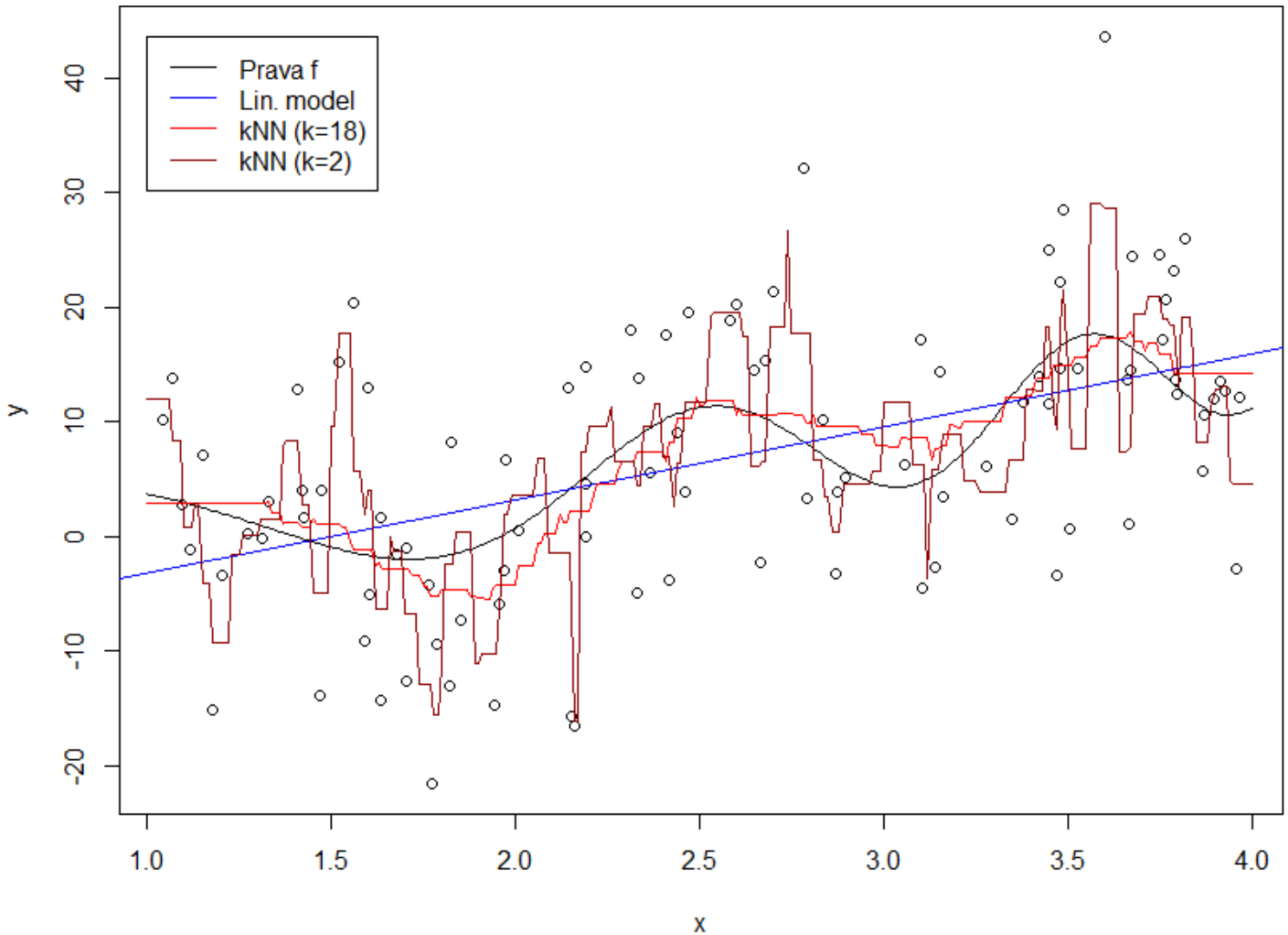


Primjer 3

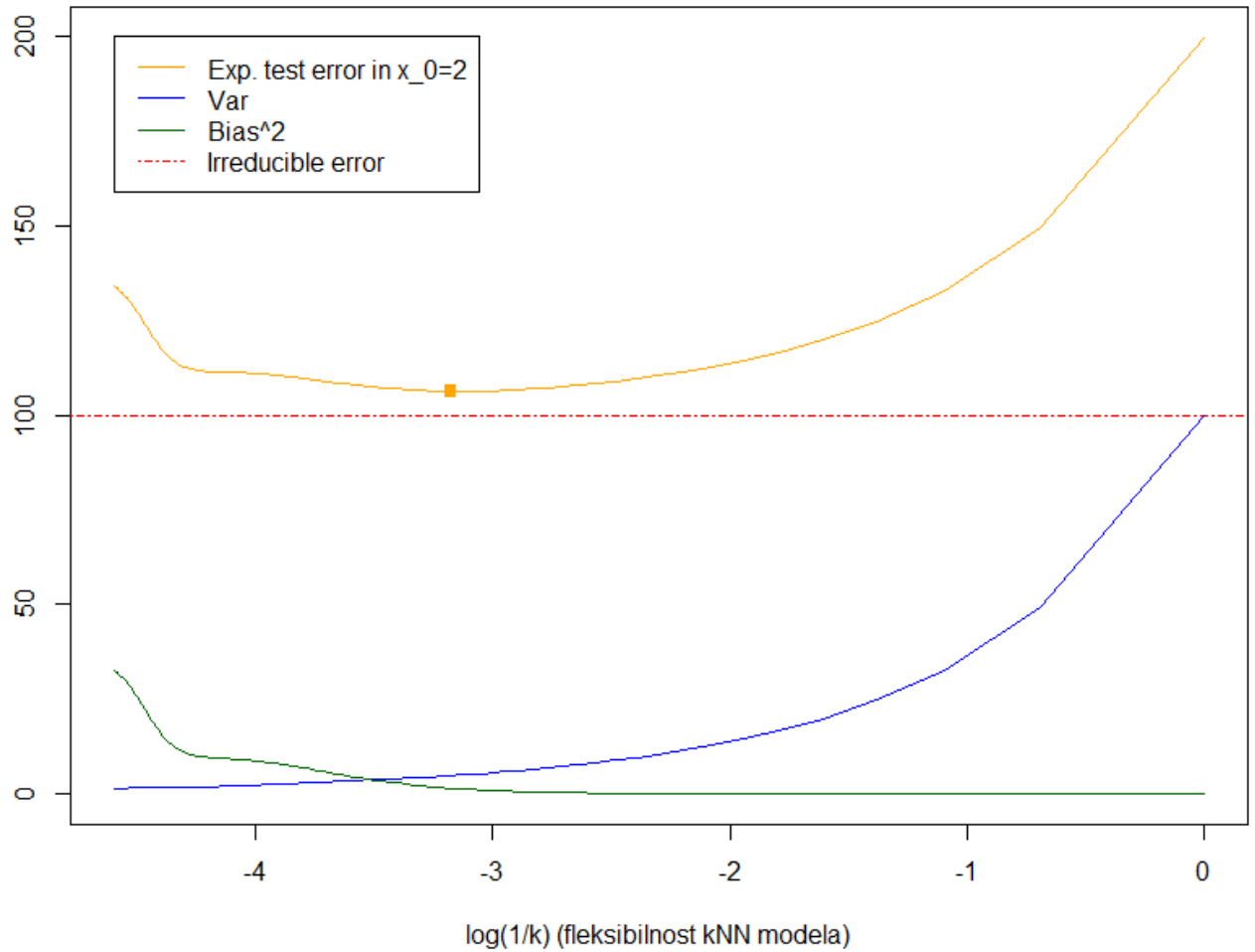


Slika: Jako fleksibilan kNN model daje najbolje rezultate, a lin. model je nedovoljno fleksibilan.

Odnos između pristranosti i varijance


Prava f
Lin. model
kNN (k=18)
kNN (k=2)
y
x
40
30
20
10
0
-10
-20
1.0
1.5
2.0
2.5
3.0
3.5
4.0

Odnos između pristranosti i varijance (kNN)



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

Napomene

Uz uvjet da znamo odabrati dobar k, *kNN* metoda u sva tri slučaja daje skoro pa optimalan rezultat (s obzirom na testnu grešku). Pitanje je zašto bi onda uopće koristili parametarske metode poput linearne regresije?

Problem za *kNN* (i slične potpuno neparametarske) metode nastaje kada imamo veći broj kovarijata p (u praksi, već za p veće od 3 ili 4).

Problem dimenzionalnosti (*curse of dimensionality*)

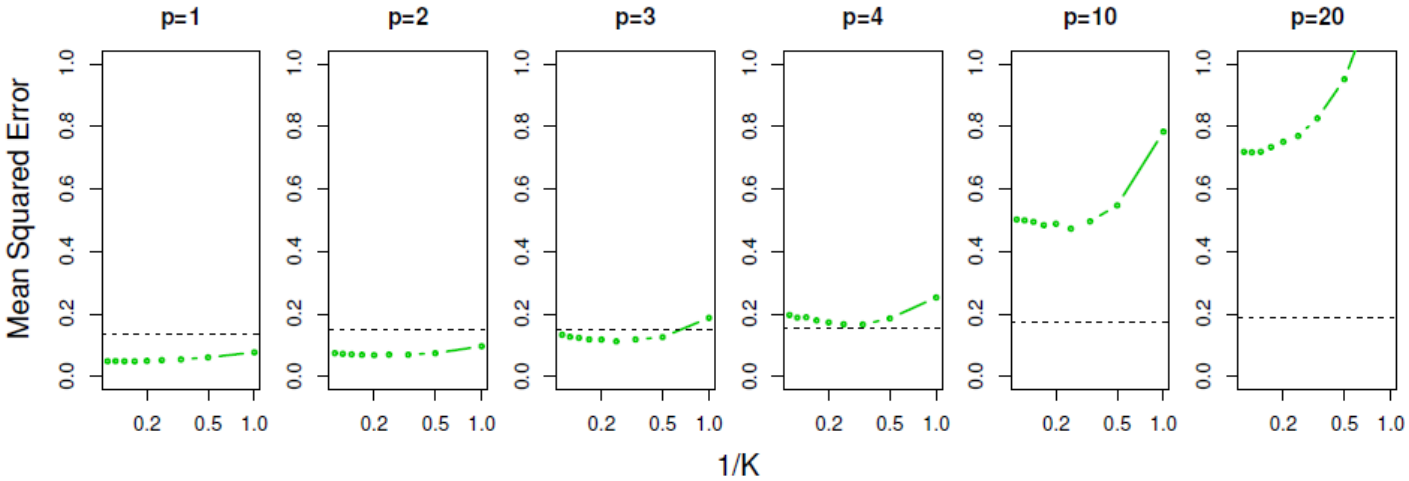


FIGURE 3.20. *Test MSE for linear regression (black dashed lines) and KNN (green curves) as the number of variables p increases. The true function is non–linear in the first variable, as in the lower panel in Figure 3.19, and does not depend on the additional variables. The performance of linear regression deteriorates slowly in the presence of these additional noise variables, whereas KNN's performance degrades much more quickly as p increases.*

Intuitivno, za fiksan n, kada se p povećava k najbližih susjeda od x u T su u prosjeku sve udaljeniji od x pa pristranost (jako brzo) raste. Drugim riječima, za veće p potrebno je puno više podataka kako bismo postigli jednako dobru procjenu. Vidi ESL, Poglavlje 2.5.

Strojno vs statističko učenje?

– U strojnom učenju, fokus je uglavnom na problemu predikcije – naći $\hat{f}$ sa što manjom testnom greškom $L(\hat{f})$ (dakle, ne mora nužno vrijediti $\hat{f} \approx f^*$). Zato često imamo tzv. **black box** metode.

– Iako ćemo se u ovom kolegiju primarno baviti problemom predikcije, statističko učenje (i statistika) pokušava na temelju uzorka τ zaključiti nešto i o cijeloj distribuciji vektora (X, Y) – npr. naći dobar model za (X, Y) kako bi mogli zaključiti koje kovarijate X_j značajno utječu na odziv Y. $\rightsquigarrow$ statističko zaključivanje (*statistical inference*)

– Ipak, postoji značajan presjek i suradnja između ova dva područja.