Outlier

Tačka „različita“ od ostalih opservacija

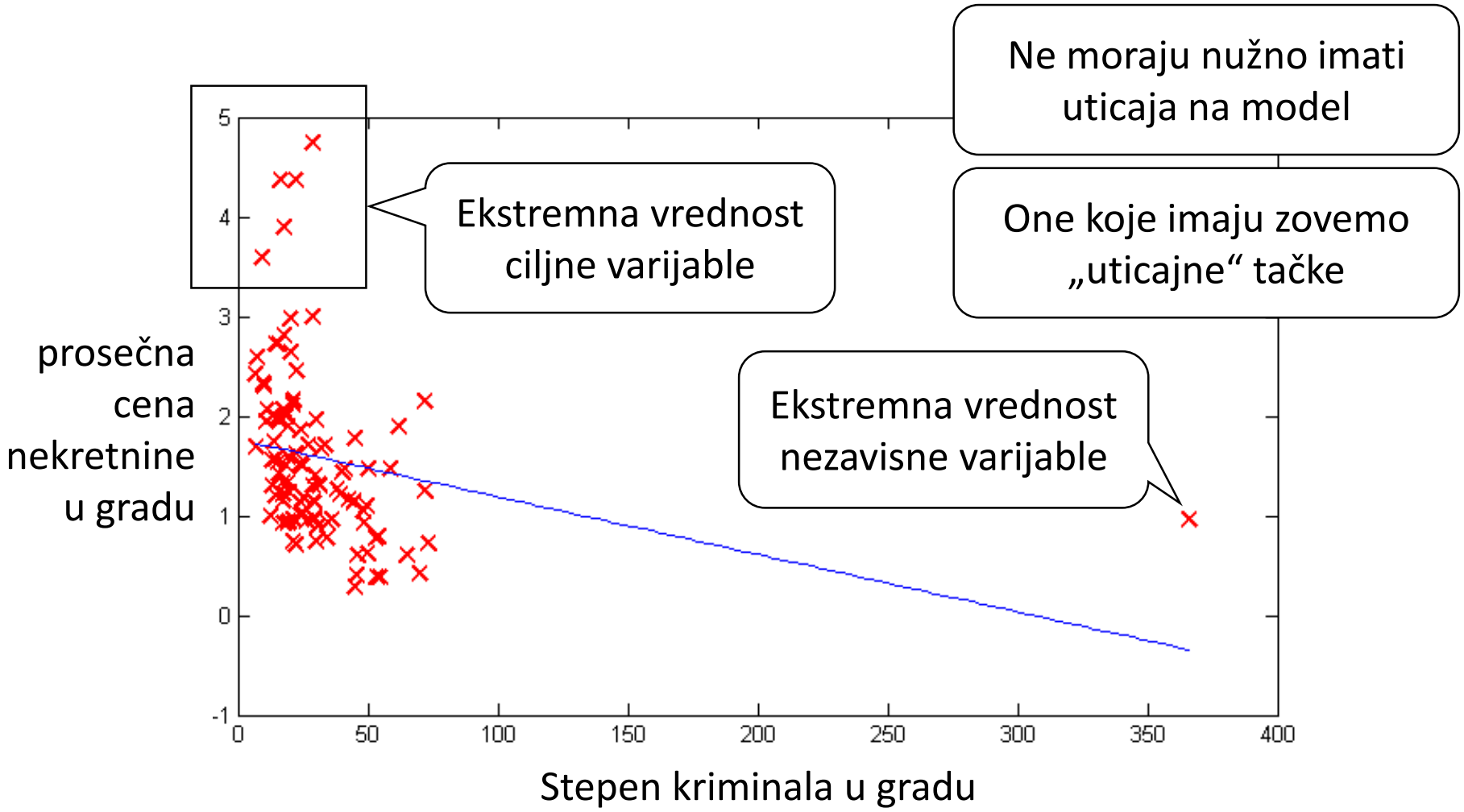
Promene u ponašanju sistema, ljudska greška, greška u merenju, prirodna devijacija

Outlier

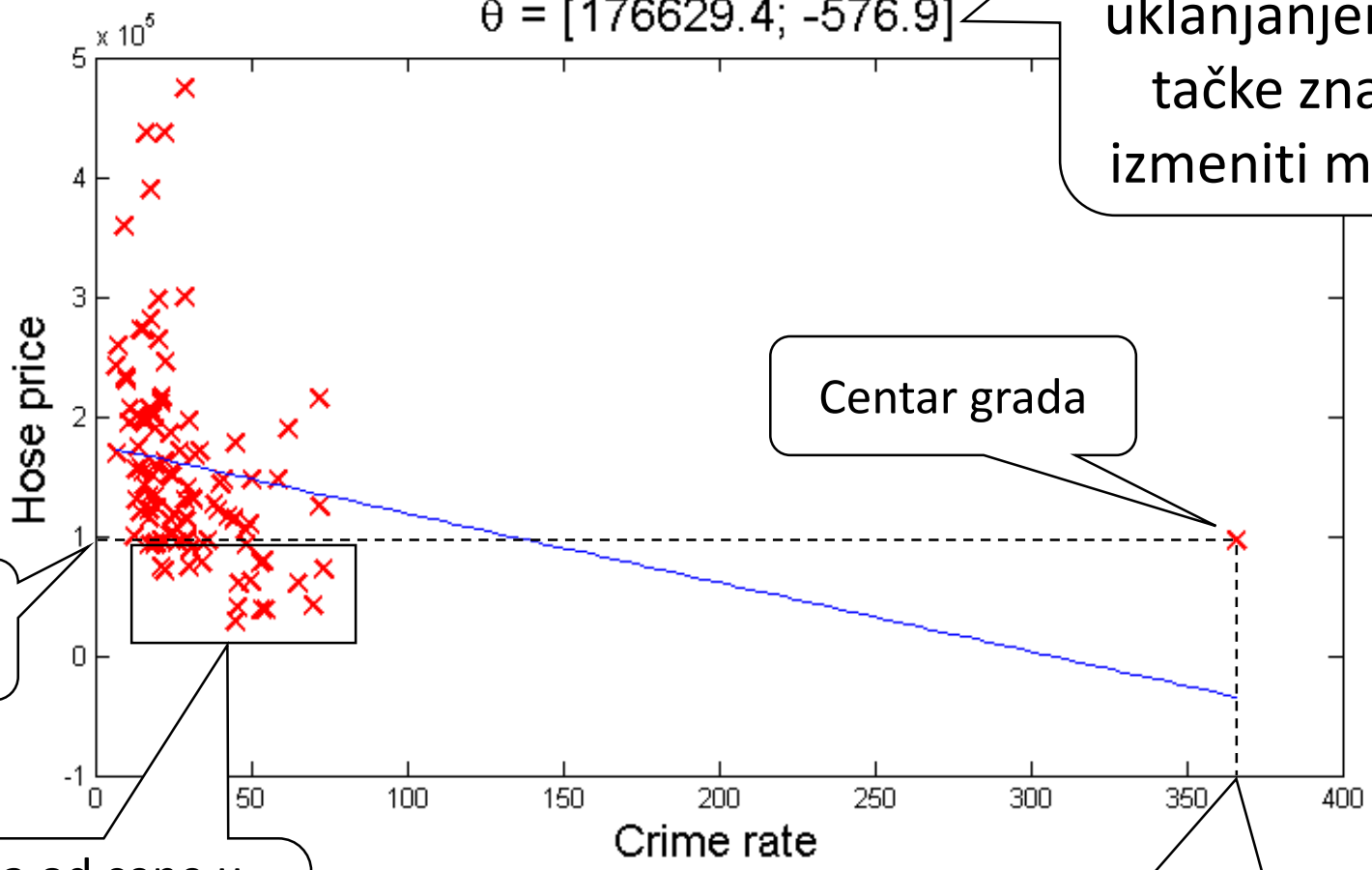
- Mogu imati veliki uticaj na rezultujući model

- Ovo može biti problematično ako:
  - je *outlier* zapravo greška
  - želimo da model dobro generalizuje
    (nije nam stalo do ekstremnih slučajeva)

Outlier – primer: gradovi Filadelfije



Ekstremi po x



Da li će se uklanjanjem ove tačke znatno izmeniti model?

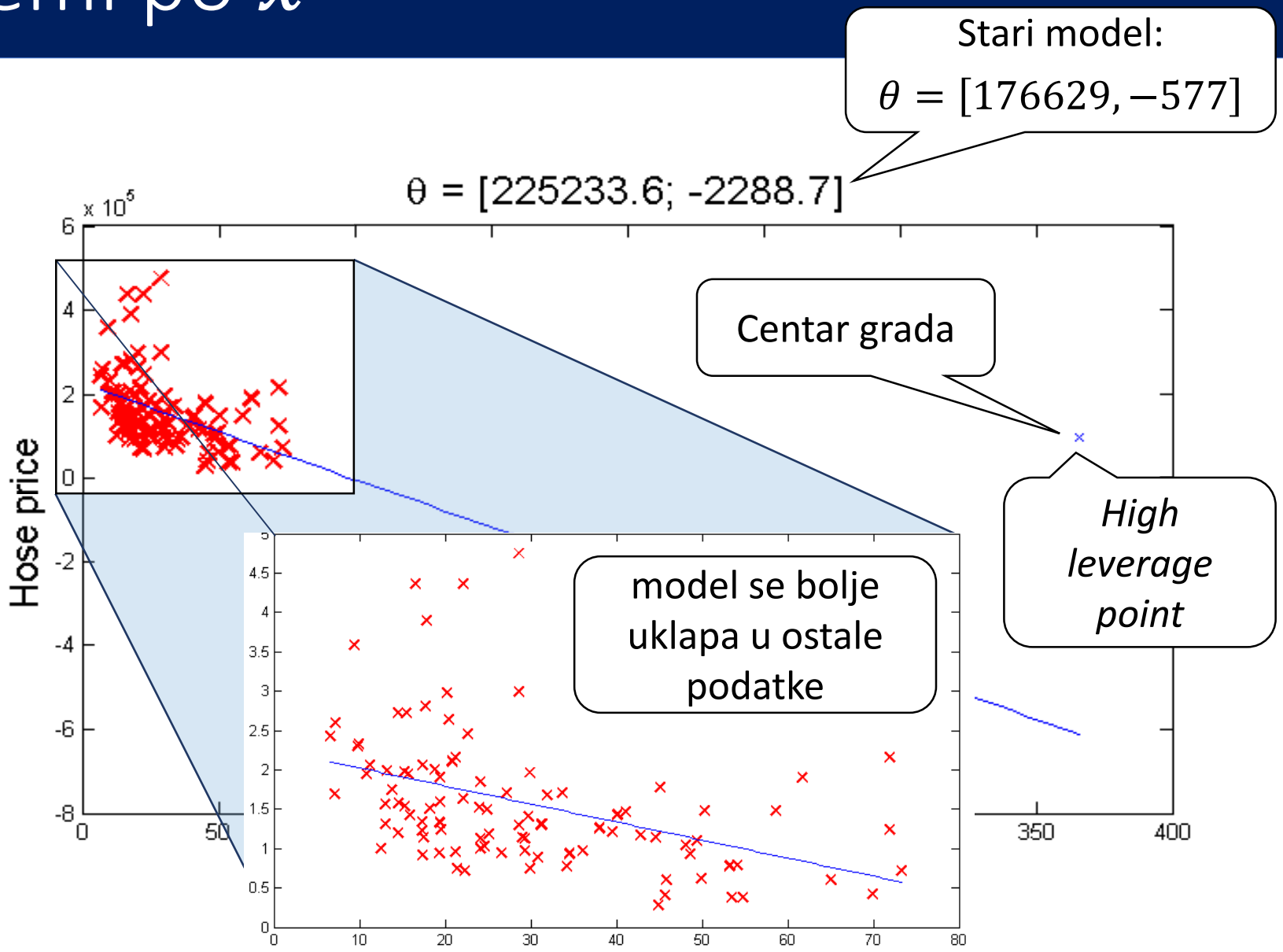
Centar grada

Cena je niska

Ali viša od cene u gradovima znatno niže stope kriminala

Ekstremno visok stepen kriminala

Ekstremi po x



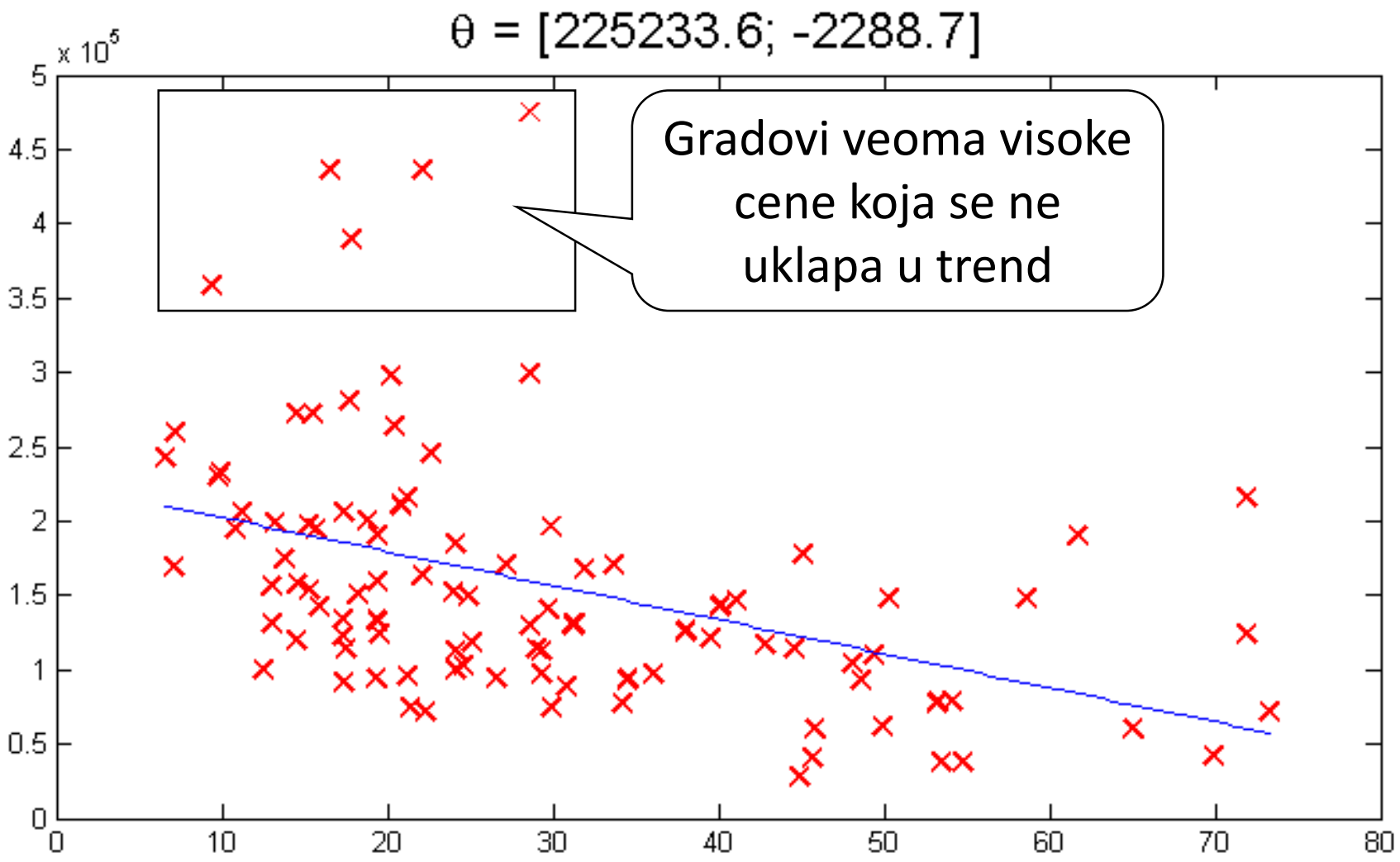
Ekstremi po x

- Potencijal za *high leverage point:*
  - Ekstremno malo/veliko x
  - Nema drugih tačaka u blizini

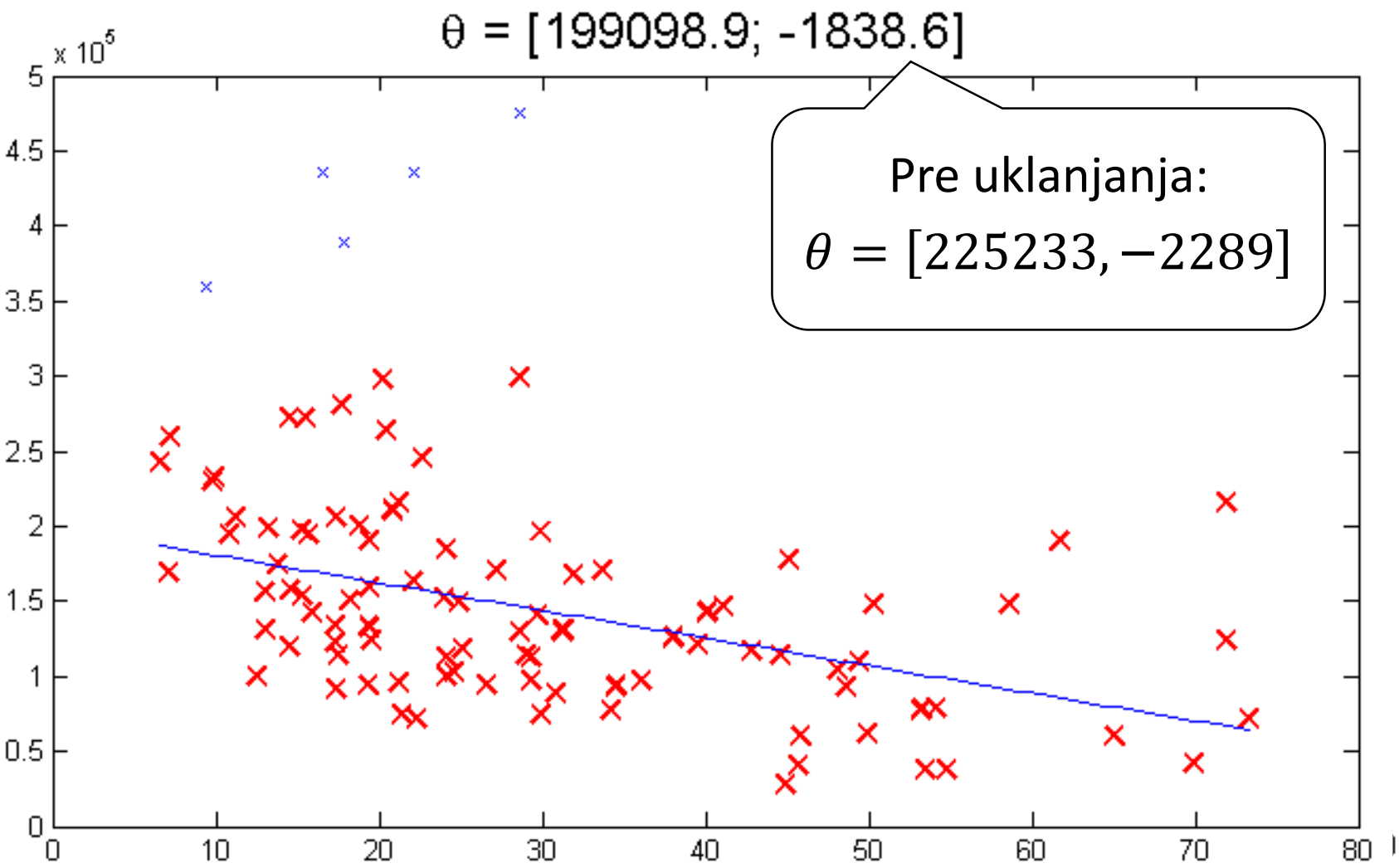
- Ovakva tačka je *high levarage point* ako ne prati trend ostalih podataka

Ekstremi po y

Outliers



Ekstremi po y

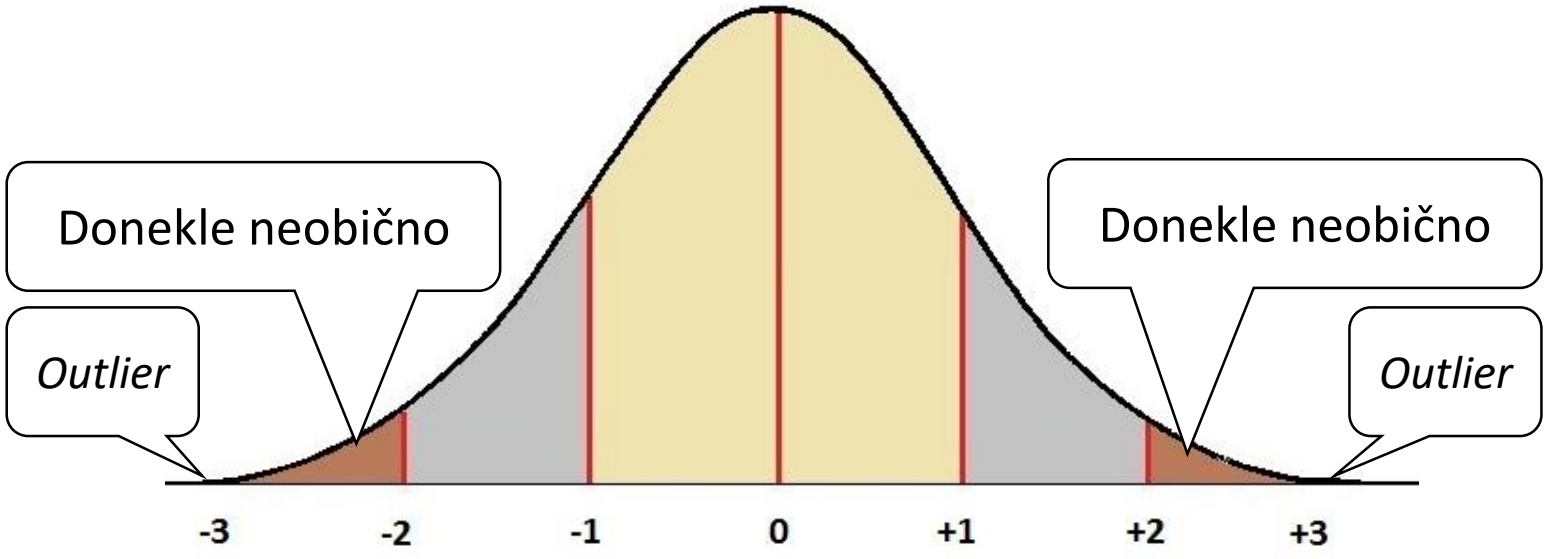


Outlier – ukloniti ili ne?

- Da, ako smo sigurni da je greška (uglavnom nismo)
- Da, ako imamo mnogo podataka, a malo *outlier*-a
- Postoje alternative uklanjanju
  - *Winsorizing, log-scale transformation, binning, imputing values*
- Treniranje posebnog modela ako ima mnogo *outlier*-a
- Umesto manipulacije podacima:
  - Odabrati robustan model
  - Promeniti meru performansi

Tehnike za detekciju *outlier*-a

- U slučaju jednog prediktora možemo da pretpostavimo da podaci dolaze iz poznate distribucije (npr., Gausove)



$$z\text{-}score\text{: } z = \frac{x-\mu}{\sigma}$$

Udaljenost opservacije od srednje vrednosti izražena u broju standardnih devijacija

Zaključak

- Eksplorativna analiza i razumevanje podataka su važni

- U ovom slučaju smo uočili jedan *outlier*