

VIŠESTRUKA LINEARNA REGRESIJA

Definicija višestruke linearne regresije

- Alat za modelovanje veze **zavisne promenljive** Y sa više **nezavisnih promenljivih** X_1, \dots, X_n .
- Veza se modeluje kao linearna kombinacija zavisnih promenljivih i parametara (težina).
- Parametri se određuju iz podataka.

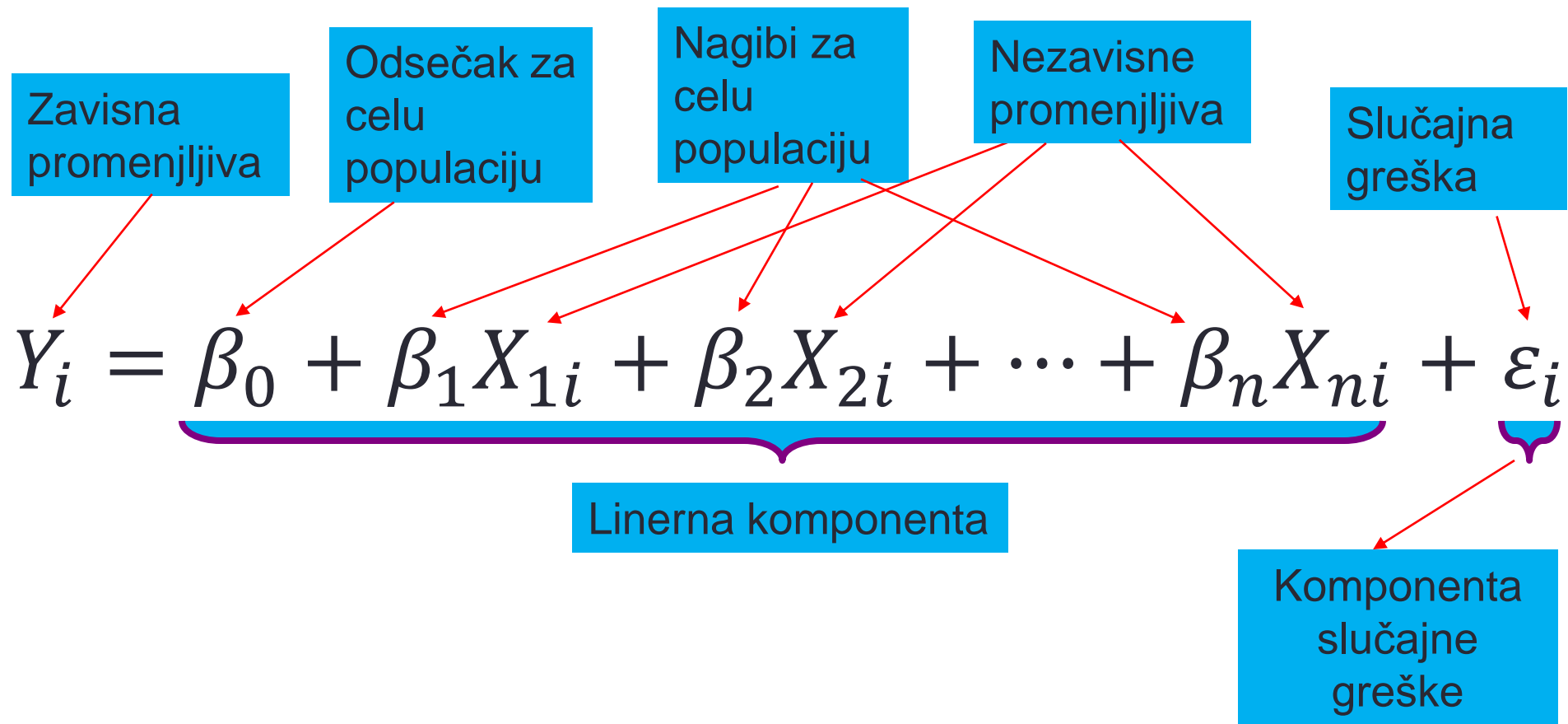
Višestruka linearna regresija - primer

- Koristimo primer sa prošlog predavanja da objasnimo razliku između jednostruke i višestruke regresije.
- Sada koristimo **sve** nezavisne promenljive.
- Skup podataka o 280 kuća iz kanadskog grada Windsor (1987):
 1. **price**: sale price of a house
 2. **lotsize**: the lot size of a property in square feet;
 3. **bedrooms**: number of bedrooms
 4. **bathrms**: number of full bathrooms
 5. **stories**: number of stories excluding basement
 6. **driveway**: dummy, 1 if the house has a driveway
 7. **recroom**: dummy, 1 if the house has a recreational room
 8. **fullbase**: dummy, 1 if the house has a full finished basement
 9. **gashw**: dummy, 1 if the house uses gas for hot water heating
 10. **airco**: dummy, 1 if there is central air conditioning
 11. **garagepl**: number of garage places
 12. **prefarea**: dummy, 1 if located in the preferred neighbourhood of the city

Primer – deo skupa podataka

price	lotsize(m^2)	bedrooms	bathrms	stories	driveway	recroom	fullbase	gashw	airco	garagepl	prefarea
74700.0	658.5	3.0	1.0	1.0	1.0	1.0	1.0	.0	.0	2.0	1.0
85000.0	652.4	3.0	1.0	1.0	1.0	.0	1.0	.0	1.0	2.0	1.0
68500.0	650.6	3.0	1.0	2.0	1.0	.0	1.0	.0	.0	.0	.0
82900.0	650.6	3.0	1.0	1.0	1.0	.0	1.0	.0	.0	2.0	1.0
86000.0	641.3	3.0	2.0	1.0	1.0	1.0	1.0	.0	.0	.0	1.0
78900.0	641.3	3.0	1.0	1.0	1.0	1.0	1.0	.0	.0	.0	1.0
69000.0	637.8	3.0	1.0	2.0	1.0	.0	.0	.0	1.0	2.0	1.0
77500.0	634.3	3.0	1.0	1.0	1.0	1.0	1.0	.0	1.0	.0	1.0
86000.0	632.0	2.0	1.0	1.0	1.0	1.0	1.0	.0	.0	2.0	.0
91700.0	627.3	2.0	1.0	1.0	1.0	1.0	1.0	.0	.0	2.0	1.0
70000.0	624.6	3.0	1.0	1.0	1.0	.0	.0	.0	.0	.0	.0
77000.0	623.6	3.0	2.0	2.0	1.0	1.0	1.0	.0	.0	1.0	1.0
93000.0	619.9	3.0	1.0	3.0	1.0	.0	1.0	.0	.0	.0	1.0
80750.0	619.0	4.0	2.0	2.0	1.0	1.0	1.0	.0	.0	1.0	1.0
87000.0	614.8	4.0	2.0	2.0	1.0	1.0	.0	1.0	.0	1.0	.0
89900.0	613.4	3.0	2.0	3.0	1.0	.0	.0	.0	1.0	.0	1.0
89000.0	613.4	3.0	2.0	1.0	1.0	.0	1.0	.0	1.0	.0	1.0
87000.0	613.4	3.0	1.0	1.0	1.0	1.0	1.0	.0	.0	2.0	1.0
72000.0	613.4	3.0	1.0	1.0	1.0	1.0	1.0	.0	.0	.0	1.0
80000.0	613.4	4.0	2.0	1.0	1.0	.0	1.0	.0	.0	.0	1.0
78000.0	613.4	4.0	2.0	2.0	1.0	1.0	1.0	.0	.0	.0	1.0
85000.0	607.8	3.0	1.0	1.0	1.0	1.0	1.0	.0	.0	2.0	1.0
75000.0	607.8	4.0	2.0	2.0	.0	.0	.0	.0	1.0	.0	.0
85000.0	606.4	3.0	2.0	4.0	1.0	.0	.0	.0	.0	1.0	.0
84000.0	604.1	3.0	2.0	3.0	1.0	.0	.0	.0	1.0	.0	.0
62000.0	599.5	4.0	1.0	2.0	1.0	.0	.0	.0	.0	.0	.0
76900.0	599.5	3.0	2.0	1.0	1.0	1.0	1.0	1.0	.0	.0	.0
67900.0	598.5	2.0	1.0	1.0	1.0	.0	.0	.0	1.0	3.0	.0
87500.0	596.7	3.0	1.0	3.0	1.0	.0	1.0	.0	.0	.0	1.0
85000.0	596.7	3.0	1.0	1.0	1.0	.0	1.0	.0	1.0	.0	1.0
90000.0	594.8	3.0	1.0	1.0	1.0	1.0	1.0	.0	1.0	1.0	1.0
63900.0	591.1	2.0	1.0	1.0	1.0	.0	1.0	.0	1.0	1.0	.0
82000.0	591.1	3.0	1.0	1.0	1.0	1.0	1.0	.0	1.0	2.0	1.0
80000.0	591.1	3.0	1.0	3.0	1.0	.0	.0	.0	.0	.0	1.0
88500.0	590.2	3.0	2.0	3.0	1.0	1.0	.0	.0	1.0	.0	.0
68000.0	587.5	3.0	1.0	2.0	1.0	.0	1.0	.0	1.0	1.0	.0
70000.0	585.5	3.0	1.0	1.0	1.0	.0	.0	.0	1.0	2.0	.0
85000.0	581.2	4.0	2.0	1.0	1.0	.0	1.0	.0	.0	1.0	1.0
78000.0	577.2	4.0	1.0	4.0	1.0	1.0	.0	.0	1.0	.0	.0
79000.0	566.9	3.0	2.0	1.0	1.0	.0	1.0	.0	.0	2.0	1.0
78000.0	566.9	3.0	1.0	3.0	1.0	1.0	.0	.0	1.0	.0	1.0
73000.0	566.9	3.0	1.0	1.0	1.0	.0	1.0	.0	1.0	.0	1.0
74900.0	562.3	3.0	1.0	1.0	1.0	.0	1.0	.0	.0	.0	1.0
69000.0	561.4	3.0	1.0	1.0	1.0	.0	.0	.0	.0	2.0	1.0
82500.0	557.6	3.0	2.0	4.0	1.0	.0	.0	.0	1.0	.0	.0
67000.0	557.6	2.0	1.0	1.0	1.0	.0	1.0	.0	1.0	1.0	.0
59000.0	556.2	3.0	1.0	1.0	1.0	.0	1.0	.0	.0	.0	.0

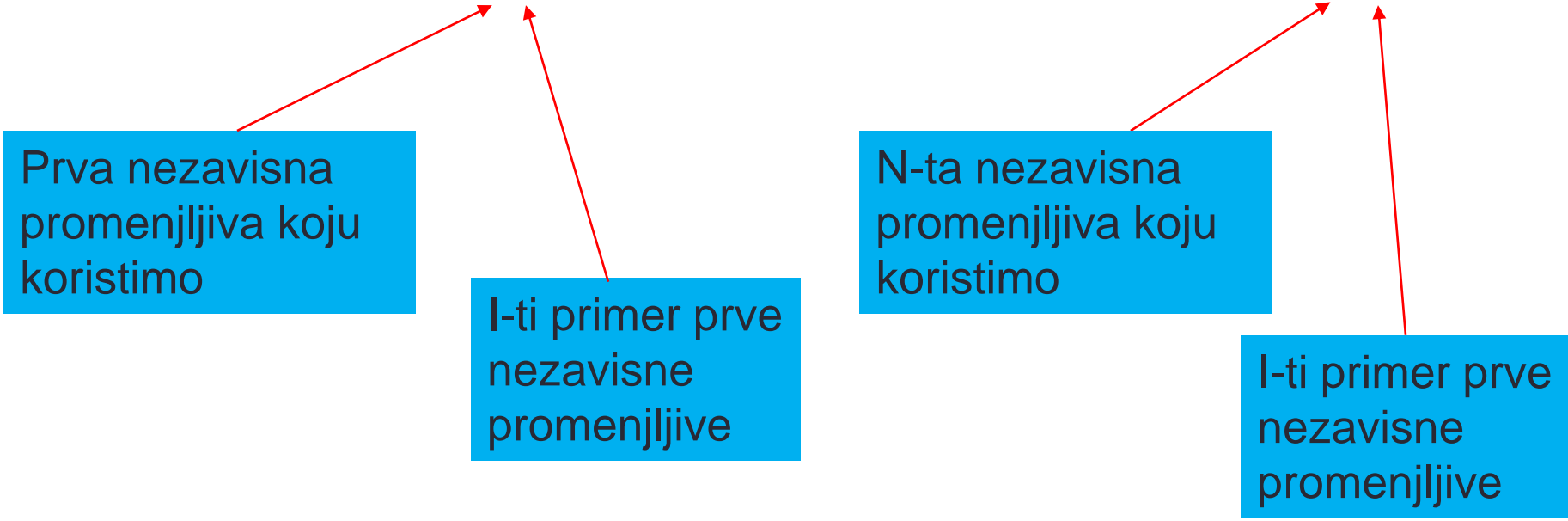
Populacioni model višestruke linearne regresije



Populacioni model višestruke linearne regresije

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + \varepsilon_i$$

Prva nezavisna
promenljiva koju
koristimo



I-ti primer prve
nezavisne
promenljive

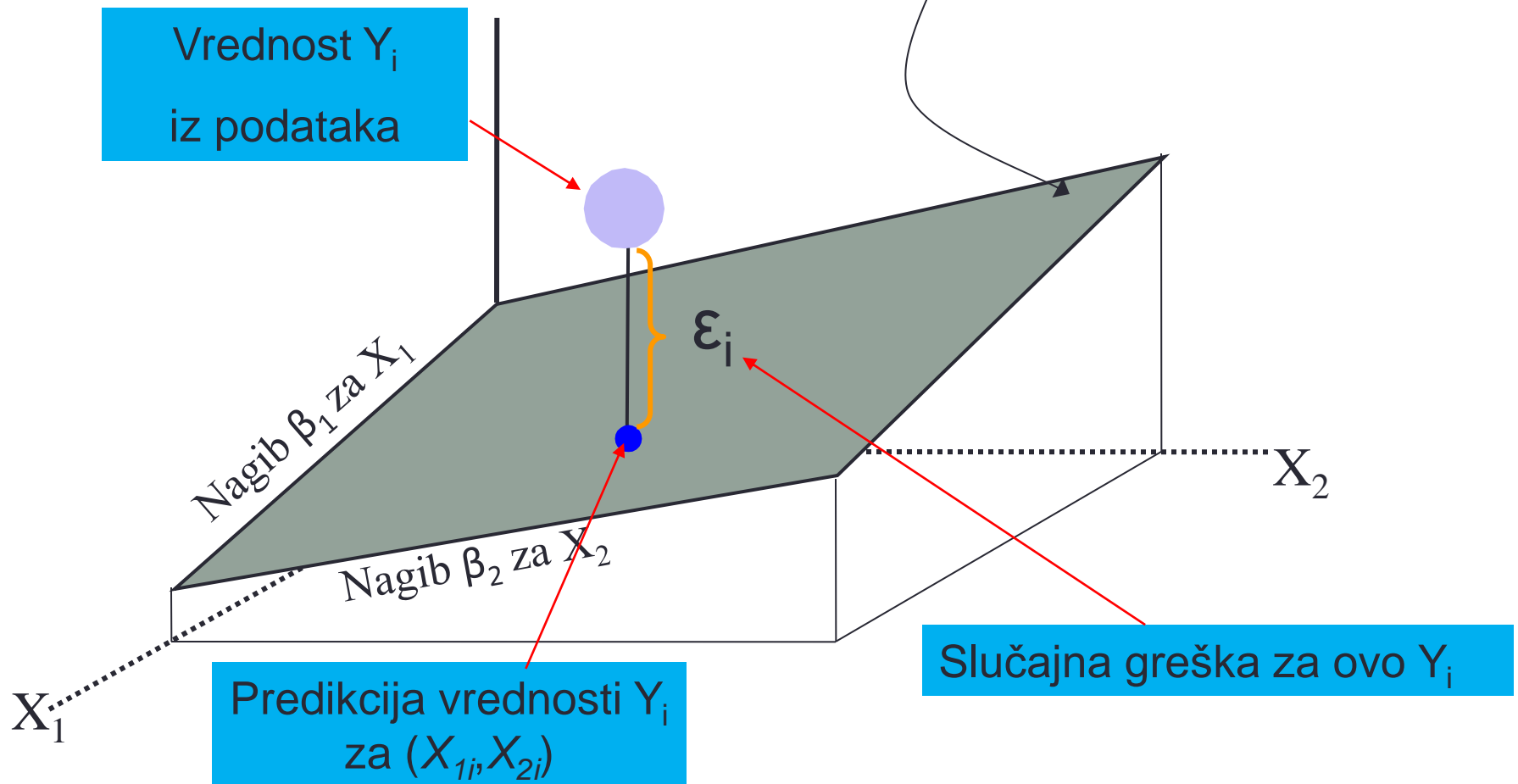
N-ta nezavisna
promenljiva koju
koristimo

I-ti primer prve
nezavisne
promenljive

Na primer, X_{13} je površina placa treće kuće u skup podataka, dok je X_{43} broj kupatila treće kuće.

Populacioni model višestruke linearne regresije

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$



Populacioni model višestruke linearne regresije

- Podsetimo se:
- Prikazani model je **statistički model** koji modeluje populaciju (**populacioni model**).
- **Populacija** je skup svih mogućih primera predmeta (pojave) koji se analizira.
- Na primer, ako predviđamo cenu stana na osnovu kvadrature u Srbiji, populacioni model obuhvatio bi sve stanove koji postoje u Srbiji.
- Podaci koje koristimo za regresionu analizu su jedan **uzorak** (**semp**) populacije.

Model – procene parametara iz uzorka podataka

Jednačina višestruke linearne regresije je **procena** parametara populacionog modela dobijena iz uzorka podataka.

Predikcija i-te
vrednosti
promenljive Y

Procenjena vrednost
nagiba iz uzorka

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni}$$

Procenjena
vrednost odsečka iz
uzorka

Metod Najmanjih Kvadrata

- Parametri $b_0, b_1, \dots, b_{n_{su}}$ dobijeni iz podataka (uzorka populacije) optimizacijom sume kvadrata grešaka, odnosno razlika između Y i \hat{Y} :

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots + b_n X_{ni}))^2$$

Interpretacija parametara

- b_0 je procenjena (iz podataka) srednja vrednost Y kada je $X_i=0$ za $i=1, \dots, n$.
- b_i je procenjena promena srednje vrednosti Y za povećanje X_i za jednu jedinicu **kada su vrednosti svih ostalih X_i za $i \neq j$ konstantne (fiksirane)**.
- Pogledajmo primer na sledećem slajdu

Interpretacija parametara

- Procenjujemo cenu kuće na osnovu površine placa i broja kupatila.
- Pomoću Python biblioteke *statsmodels*:

$$\widehat{cena_kuce}_i = 22840 + 78.01 \cdot površina_{placa_i} + 2546.38 \cdot br_spavacih_soba_i$$

	coef

const	2.284e+04
lotsize(m^2)	78.0178
bedrooms	2546.3854

Interpretacija parametara

- b_2 je procenjena promena srednje vrednosti cene kuća za povećanje broja spavaćih soba za jednu, pod uslovom da je površina placa fiksirna.
 - Na primer, ako fiksiramo površinu placa na 100m² imamo:

$$\widehat{cena_kuce}_i = 22840 + 78.01 \cdot površina_placa_i + 2546.38 \cdot br_spavacih_soba_i$$

$$\widehat{cena_kuce} = 22840 + 78.01 \cdot 100 + 2546.38 \cdot 1 = 33187.38$$

$$\widehat{cena_kuce} = 22840 + 78.01 \cdot 100 + 2546.38 \cdot 2 = 35733.76$$

$$\underline{35733.76 - 33187.38 = 2546.38}$$

- Cena kuće površine placa 100m² poveća se za 2546.38 dolara ako se doda još jedna spavaća soba.

Interpretacija parametara

- Zašto naglašavamo konstantnost (fiksiranost) svih nezavisnih promenljivih osim one čiji koeficijent interpretiramo?
- Na koji način bi interpretirali negativan b_2 kao u primeru ispod?

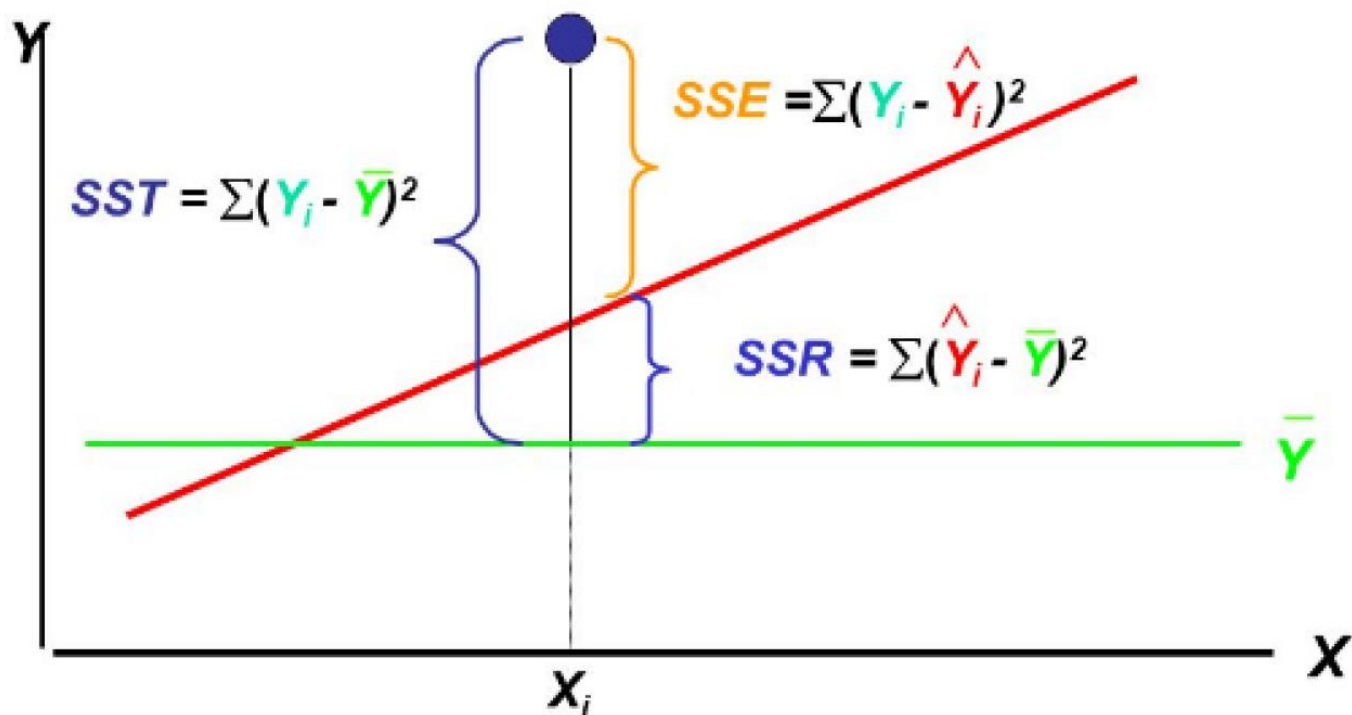
$$\widehat{cena_kuce}_i = 22840 + 78.01 \cdot površina_placa_i - 2546.38 \cdot br_spavacih_soba_i$$

- Interpretacija da povećanje broja spavaćih soba utiče negativno na cenu bi bilo pogrešno.
 - Ova interpretacija takođe i nema puno smisla jer znamo da su kuće sa više soba obično veće i samim tim skuplje.
- Ispravna interpretacija: povećanje broja spavaćih soba u kući, a da površina placa ostane ista, negativno utiče na cenu.
 - Što može da ima smisla jer ljudi mogu više da vole kuću sa malo velikih soba, u odnosu na onu sa puno malih soba.

r^2 podsećenje

- Podsetimo se prvo koeficijenta determinacije r^2 – odnos varijabilnosti zavisne promenljive (SST) koja je objašnjenja pomoću nezavisnih promenljivih (SSR).

$$r^2 = \frac{SSR}{SST}$$



Evaluacija modela – prilagođeni r^2

- Pored koeficijenta determinacije (r^2) alati za linearnu regresiju kao rezultat vraćaju i prilagođeni koeficijent determinacije (*adjusted r^2*).
- Uporedićemo vrednosti r^2 i prilagođenog r^2 za slučaj jednostruke i višestruke regresije na primeru cena kuća.

R-squared:

0.608

Adj. R-squared:

0.607

coef

const	2.942e+04
lotsize(m^2)	80.0664

R-squared:

0.742

Adj. R-squared:

0.731

coef

const	1.912e+04
lotsize(m^2)	63.1780
bedrooms	41.8319
bathrms	4348.5315
stories	3903.9671
driveway	1371.5360
recroom	3985.9209
fullbase	2492.1516
gashw	148.2952
airco	3045.3796
garagepl	671.0931
prefarea	4289.9797

- Vidimo da se vrednosti r^2 i prilagođenog r^2 značajno razlikuju u slučaju višestruke regresije.
- To nam je indikator da je vrednost prilagođenog r^2 informativna u slučaju višestruke regresije.
- U nastavku ćemo objasniti zašto.

Dodavanje novih nezavisnih promenljivih i r^2

- r^2 nije adekvatna mera kod višestruke regresije jer dodavanjem svake nove nezavisne promenljive vrednost r^2 ili raste ili ostaje ista, odnosno nikada se ne smanjuje.
- Nezavisna promenljiva koju dodajemo može biti totalno neinformativna (nepovezana) sa problemom koji rešavamo.
- Lako je ručno utvrditi da boja očiju vlasnika nema veze sa cenom kuće (ili ima...), ali je teško proceniti doprinos nove informativne promenljive u kombinaciji sa postojećim.
 - Na primer, za naš slučaj cena kuća, promenljiva *gashw* (gas za grejanje vode) ne doprinosi kvalitetu modela koji koristi samo *lotsize* (površinu placa).

Dodavanje novih nezavisnih promenljivih i r^2

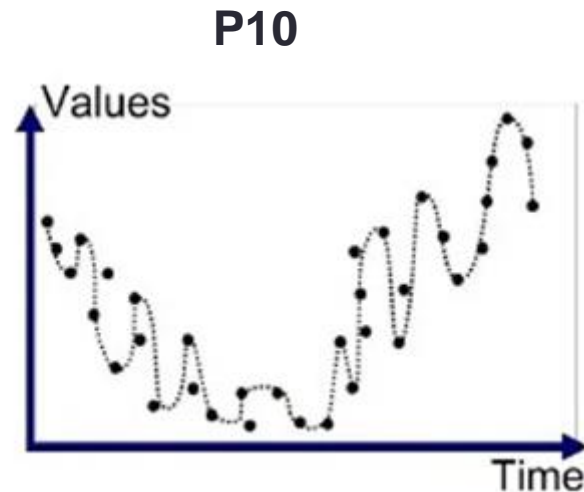
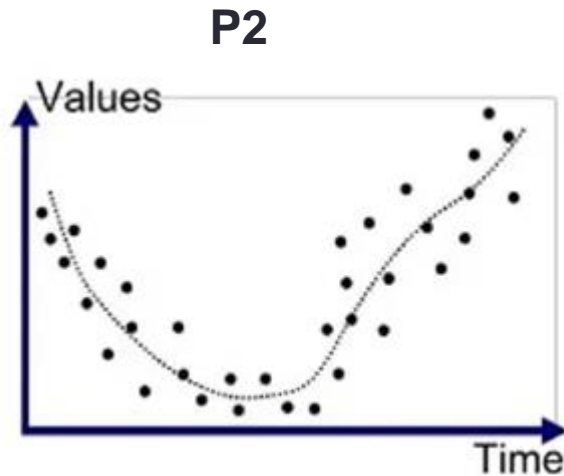
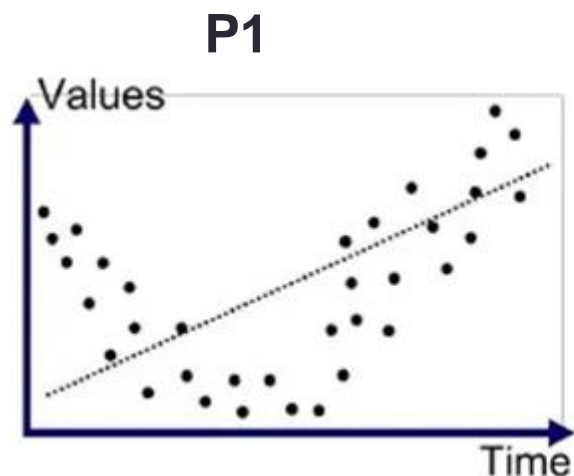
- Problem sa konstantnim povećanjem r^2 dodavanjem svake nove nezavisne promenljive u model ukazuje na to da je potrebno:
 1. Definirati alternativni način merenja kvaliteta modela.
 2. Odlučiti kako postupiti u slučaju da smo utvrdili da nova nezavisna promenljiva ne doprinosi kvalitetu modela.
- Slučaj 1. rešavamo uvođenjem prilagođenog r^2 koji definišemo u nastavku.
- Slučaj 2. je komplikovaniji i njega takođe obrađujemo u nastavku iz dva aspekta.

Nova nezavisna promenljiva ne doprinosi kvalitetu modela

- Iskoristićemo primer promenljive *gashw* koja ne doprinosi kvalitetu modela koji koristi samo *lotsize*. Naravno imamo dve mogućnosti da (1) izbacimo promenljivu *gashw* iz modela ili (2) da je zadržimo u modelu.
- Odluka zavisi od cilja našeg istraživanja (modelovanja pomoću višestruke linearne regresije).
- Ako nam je cilj isključivo predikcija onda možemo izbaciti promenljivu i time dobiti jednostavniji model istog kvaliteta – o tome diskutujemo u narednim slajdovima.
- Ako nam je cilj analiza uticaja različitih faktora na nezavisnu promenljivu, onda bi trebalo da zadržimo promenljivu u modelu – o tome diskutujemo kasnije u prezentaciji kada se bavimo korelacijama između nezavisnih promenljivih.

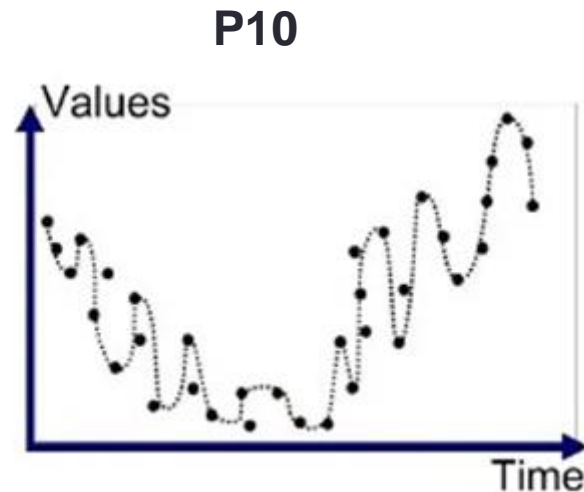
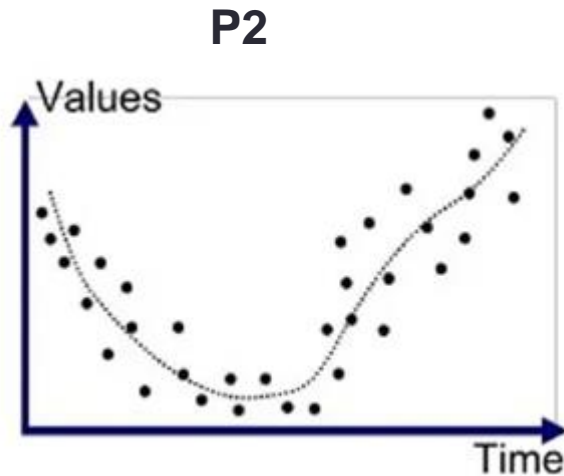
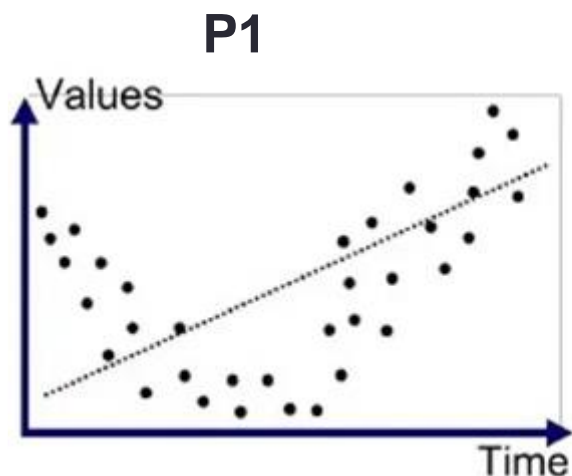
Dodavanje novih nezavisnih promenljivih i r^2

- Zašto nam je uopšte važno koliko nezavisnih promenljivih imamo i da su sve stvarno korisne?
- Recimo da modelujemo prodaju zimske odeće tokom jedne godine.
- Modeli su polinomi, redom sa leva na desno: prvog (P1), drugog (P2) i desetog stepena (P10).



Dodavanje novih nezavisnih promenljivih i r^2

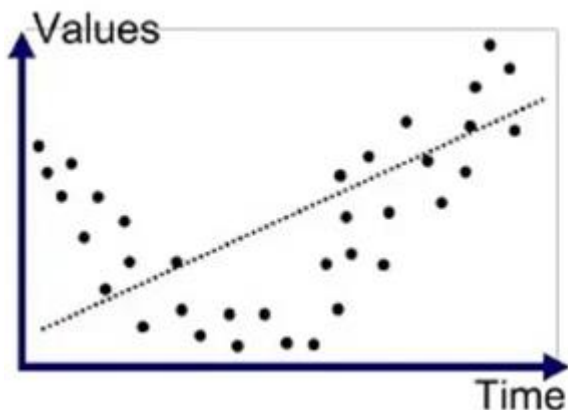
- Modeli su polinomi, redom sa leva na desno: prvog (P1), drugog (P2) i desetog stepena (P10).
- Tehnički mi nismo dodali novu nezavisnu promenljivu, već smo transformisali postojeću (*Vreme*) stepenovanjem i tako dodali nove sabirke u model: b_2x^2 , b_3x^3, \dots
- Za problem koji želimo da ilustrujemo potpuno je isto da li koristimo polinome ili kompletno nove nezavisne promenljive.



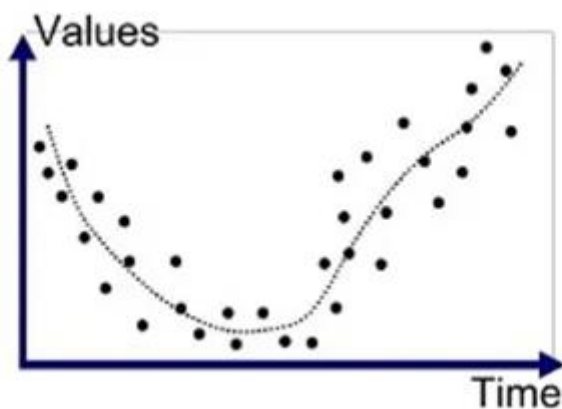
Dodavanje novih nezavisnih promenljivih i r^2

- P10 se očigledno najbolje uklapa u podatke (ima tačne predikcije za većinu tačaka). To je posledica kompleksnosti (fleksibilnosti) P10 u odnosu na P1 i P2.
- Međutim, posledica kompleksnosti su i velike oscilacije P10. Za male promene X imamo velike promene Y.
- Tako nagle promene ne modeluju dobro naš problem (prodaju odeće).

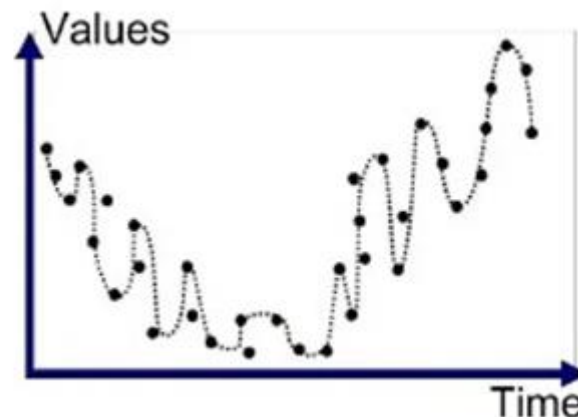
P1



P2



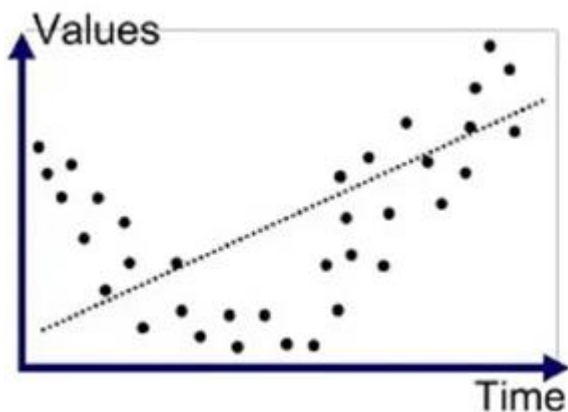
P10



Preprilagođavanje (*Overfitting*)

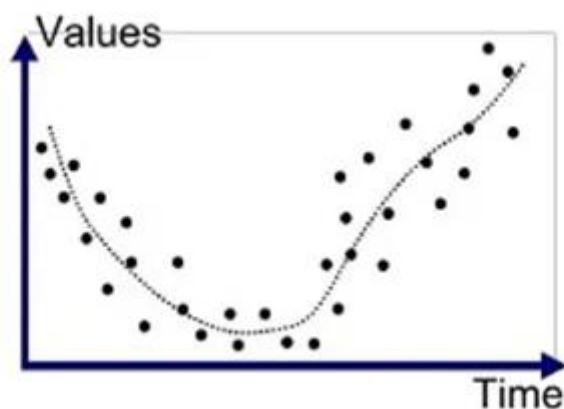
- Tako nagle promene ne modeluju dobro naš problem (prodaju odeće).
- Dakle P10 skoro savršeno modeluje date podatke, ali ne modeluje dobro generalni trend koji podaci prate pa zato **ne očekujemo da će dobro raditi na novim nepoznatim podacima**.
- Takva situacija naziva se preprilagođavanje modela ili *overfitting*.

P1



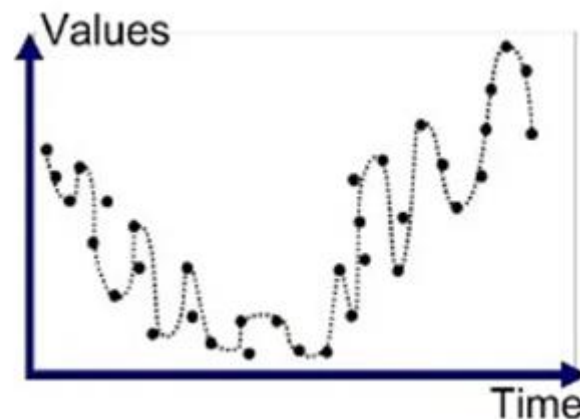
Underfitted

P2



Good Fit/Robust

P10

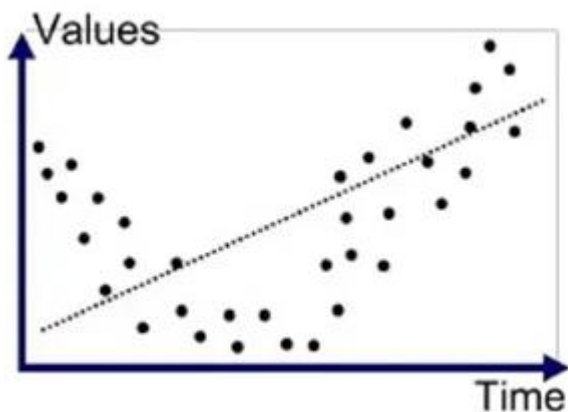


Overfitted

Nedovoljno prilagođavanje (*Underfitting*)

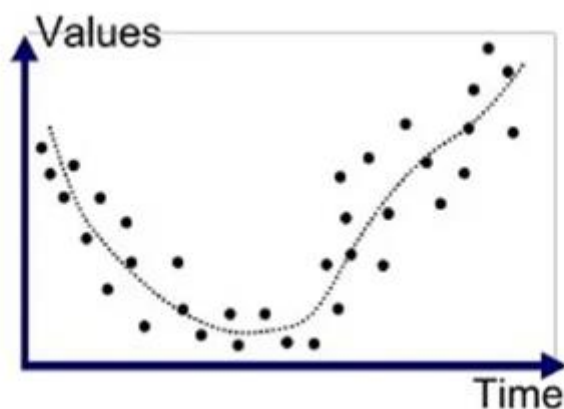
- *Overfitting* je karakteristika **kompleksnih** modela, a to su modeli sa **velikim brojem nezavisnih promenljivih**.
- Suprotna situacija kod koje model nije dovoljno kompleksan da modeluje generalni trend u podacima zove se nedovoljno prilagođavanje ili *underfitting*.

P1



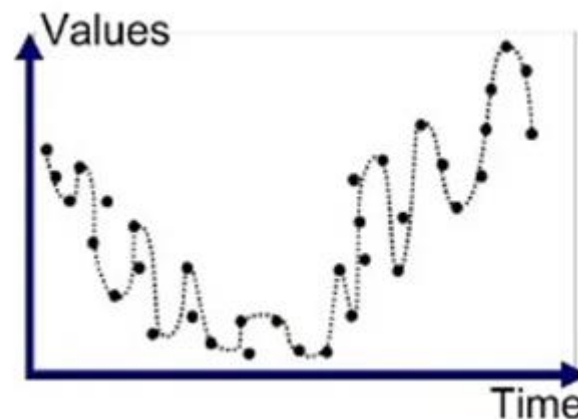
Underfitted

P2



Good Fit/Robust

P10

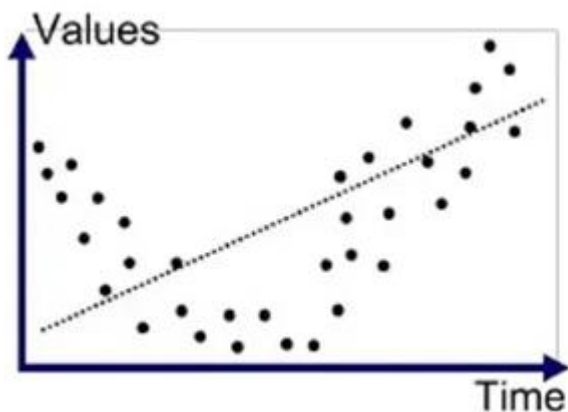


Overfitted

Kako postići balans između *over-* i *underfittinga*?

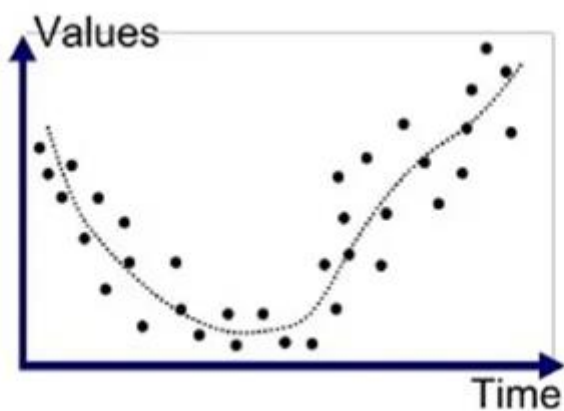
- Ozbiljan i otvoren problem u prediktivnom modelovanju.
- U našem primeru je P2 balans, ali dodavanjem novih nezavisnih promenljivih (pored vremena) povećali bi dimenzionalnost prostora i posle 3D ne bismo vizualno mogli da uočimo trend.
- Iz tog razloga koristimo **mere performansi koje mogu da ukažu na *over-* ili *underfitting***. Jedna od tih mera je **Prilagođeni r^2** .

P1



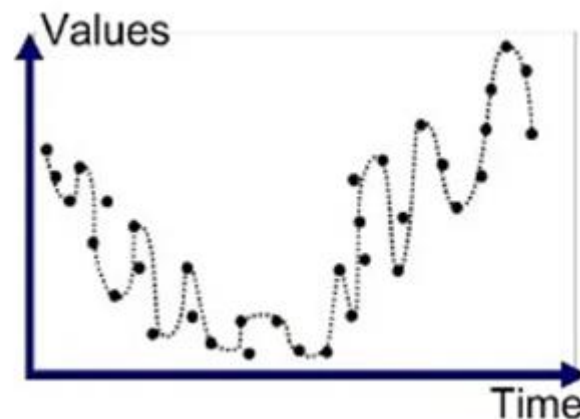
Underfitted

P2



Good Fit/Robust

P10



Overfitted

Dodavanje novih nezavisnih promenljivih i r^2

- Rekli smo da r^2 nije adekvatna mera kod višestruke regresije jer **dodavanjem svake nove nezavisne promenljive vrednost r^2 ili raste ili ostaje ista**, odnosno, nikada se ne smanjuje*.
- Pokazali smo u širem smislu zašto je to problem.
- Sada ostaje da pokažemo kako prilagođavamo r^2 sa obzirom na broj nezavisnih promenljivih.

*Na sledećem slajdu data je šira ideja dokaza ove tvrdnje

Dodavanje novih nezavisnih promenljivih i r^2

- Recimo da imamo jednačinu modela jednostruke regresije:

$$\hat{Y}_i = b_0 + b_1 X_{1i}$$

- Parametre smo odredili minimizacijom SSE.
- Dakle, u prostoru svih modela (pravih linija) našli smo onaj koji ima minimalnu SSE.
- Ako dodamo još jednu nezavisnu promenljivu u model imamo:

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i}$$

- Sada pretražujemo prostor modela koji kao podskup ima modele jednostruke regresije jer su to modeli za koje je $b_2=0$.
- Nadskup modela može sadržati samo bolji ili u najgorem slučaju isti model kao podskup, odnosno ne možemo povećati SSE (pa ni r^2) već samo smanjiti ili ostati pri istom.

Prilagođeni r^2

- Formula za prilagođeni (adjusted) r^2 je:

$$\bar{r}^2 = 1 - (1 - r^2) \frac{n - 1}{n - p - 1}$$

- Vrednost prilagođenog r^2 zavisi od:
 - samog r^2 ,
 - broj primera u uzorku podataka – n ,
 - broj nezavisnih promenljivih – p ,

Tumačenje vrednosti prilagođenog r^2

$$\bar{r}^2 = 1 - (1 - r^2) \frac{n - 1}{n - p - 1}$$

- Povećanjem broja nezavisnih promenljivih p , vrednost opada, odnosno velika vrednost originalnog r^2 je „kažnjena“ sa porastom kompleksnosti modela.
- Sa porastom broja primera vrednost raste.
- Za kompleksne modele generalno važi da: što je veći uzorak podataka iz koje model uči to je mogućnost za *overfitting* manja, a moć generalizacije postaje veća.
- Generalno u AI disciplinama: što više podataka to bolje.

Prilagođeni r^2 – Primer

- Poredimo dva modela za predikciju cene kuća, model sa svih 11 nezavisnih promenljivih i model bez promenljive *gashw* (grejanje vode na gas).

R-squared:

0.742

R-squared:

0.742

Adj. R-squared:

0.731

Adj. R-squared:

0.732

coet

coef

const	1.912e+04
lotsize(m^2)	63.1780
bedrooms	41.8319
bathrms	4348.5315
stories	3903.9671
driveway	1371.5360
recroom	3985.9209
fullbase	2492.1516
gashw	148.2952
airco	3045.3796
garagepl	671.0931
prefarea	4289.9797

- Vidimo da je vrednost prilagođenog r^2 veća za model bez promenljive *gashw*
- To znači da nam ta promenljiva ne donosi nove informacije o ceni, ako već imamo ostalih 10 promenljivih.
- Dodavanjem bi samo nepotrebno povećali kompleksnost.

const	1.912e+04
lotsize(m^2)	63.1767
bedrooms	41.4861
bathrms	4356.4592
stories	3905.6483
driveway	1366.8607
recroom	3991.1393
fullbase	2486.9367
airco	3038.5901
garagepl	673.0433
prefarea	4285.3678

Pretpostavke MNK linearne regresije - Podsećanje

- Linearity - Linearnost
 - Odnos između X i Y je linearan
- Independence of Errors – Nezavisnost grešaka ε_i
 - Greške ε_i su statistički nezavisne
 - Greška za neko X_i ne zavisi od greške za neko X_j
 - Naročito značajno za podatke koji se prikupljaju kroz vreme
- Normality of Error – Normalnost grešaka
 - Greške ε_i su normalno distribuirane oko srednje vrednosti 0 za svako dato X
- Equal Variance – Jednaka varijansa
 - Distribucija grešaka ε_i oko srednje vrednosti 0 ima jednaku varijansu za svako dato X

Dodatna pretpostavka za višestruku regresiju

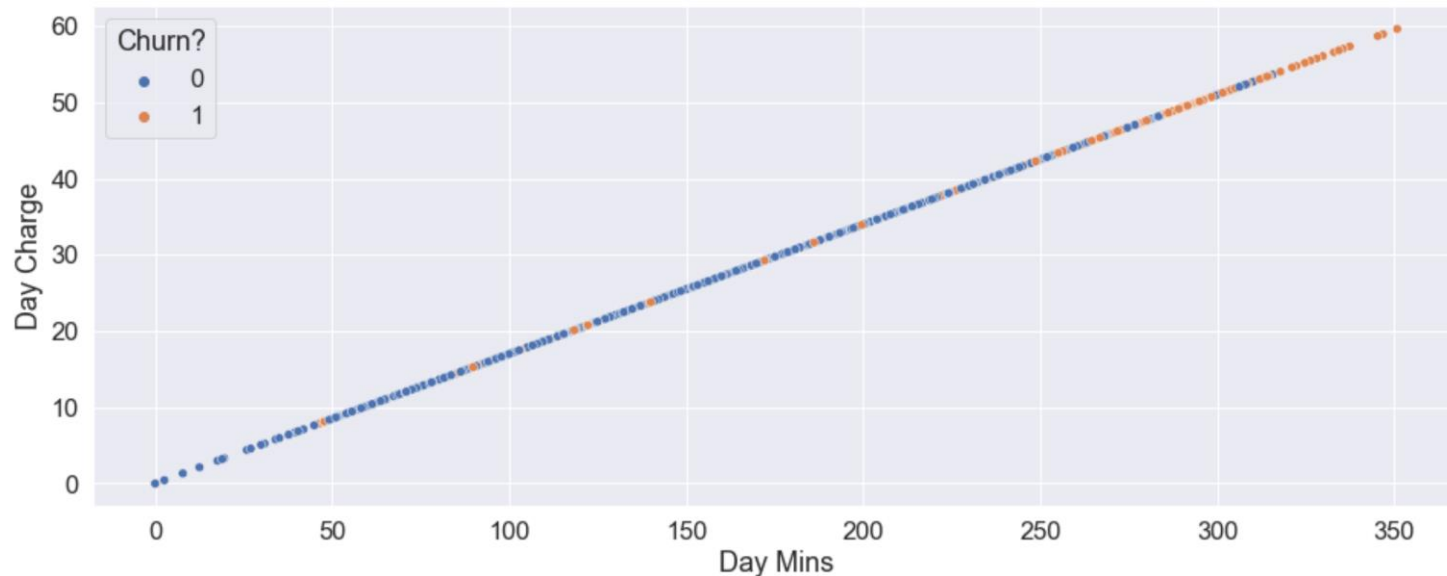
- Za višestruku linearnu regresiju postoji dodatna pretpostavaka.
- Ne postoji savršena kolinearnost između bilo koje dve ili više nezavisnih promenljivih.
- Savršena kolinearnost između dve promenljive X_1 i X_2 postoji kada su povezane linearnom funkcijom:
$$X_2 = aX_1 + b$$
- Na sledećem slajdu dat je realan primer savršene kolinearnosti.

Savršena kolinearnost - Primer

- Recimo da modelujemo mušterije telefonske kompanije i da imamo sledeće dve nezavisne promenljive: cenu poziva u toku dana (*cena*) i broj minuta poziva u toku dana (*minuti*).
- Cena se naravno obračunava po nekoj formuli, recimo da je ta formula:

$$cena = 5 \cdot minuti + 100$$

- Onda možemo da konstatujemo da postoji savršena kolinearnost između cenu poziva u toku dana i broj minuta poziva u toku dana.



Savršena kolinearnost i MNK

- Ozbiljan problem za linearnu regresiju jer ne možemo odrediti vrednosti parametara (koeficijenata) regresionog modela.
- Ne postoji analitičko rešenje pomoću MKN jer matrica ATA nema inverznu.

$$(A^T A)a = A^T y$$
$$A = \begin{bmatrix} 1 & X_{11} & \cdots & X_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1N} & \cdots & X_{kN} \end{bmatrix}$$

- To je zato što je broj linearno nezavisnih kolona manji od broja kolona matrice.
 - Jedna kolona (nezavisna promenljiva) može se dobiti kao linearna kombinacija druge.

Savršena kolinearnost - detekcija

- Ako postoji savršena kolinearnost onda moramo da uklonimo po jednu promenljivu iz svakog para nezavisnih promenljivih koje su povezane na taj način.
- U primeru sa telefonskom kompanijom mogli smo, na primer, da uklonimo cenu dnevnih poziva iz modela.
- U nastavku ćemo pokazati
 - na koji način još možemo da detektujemo savršenu kolinearnost osim vizualno
 - da li je velika ali ne savršena kolinearnost problematična?

Kolinearnost i Korelacija

- Korelacija je termin koji se odnosi na linearnu vezu dve promenljive.
- Koeficijent korelacije je vrednost kojim se meri jačina linearne veze (korelacije).
- Kolinearnost je takođe termin koji se odnosi na linearnu vezu dve promenljive, ali se koristi u linearnoj regresiji i odnosi se na vezu između nezavisnih promenljivih.
 - Naravno, kolinearnost između zavisne i nezavisne promenljive je poželjna u modelu.

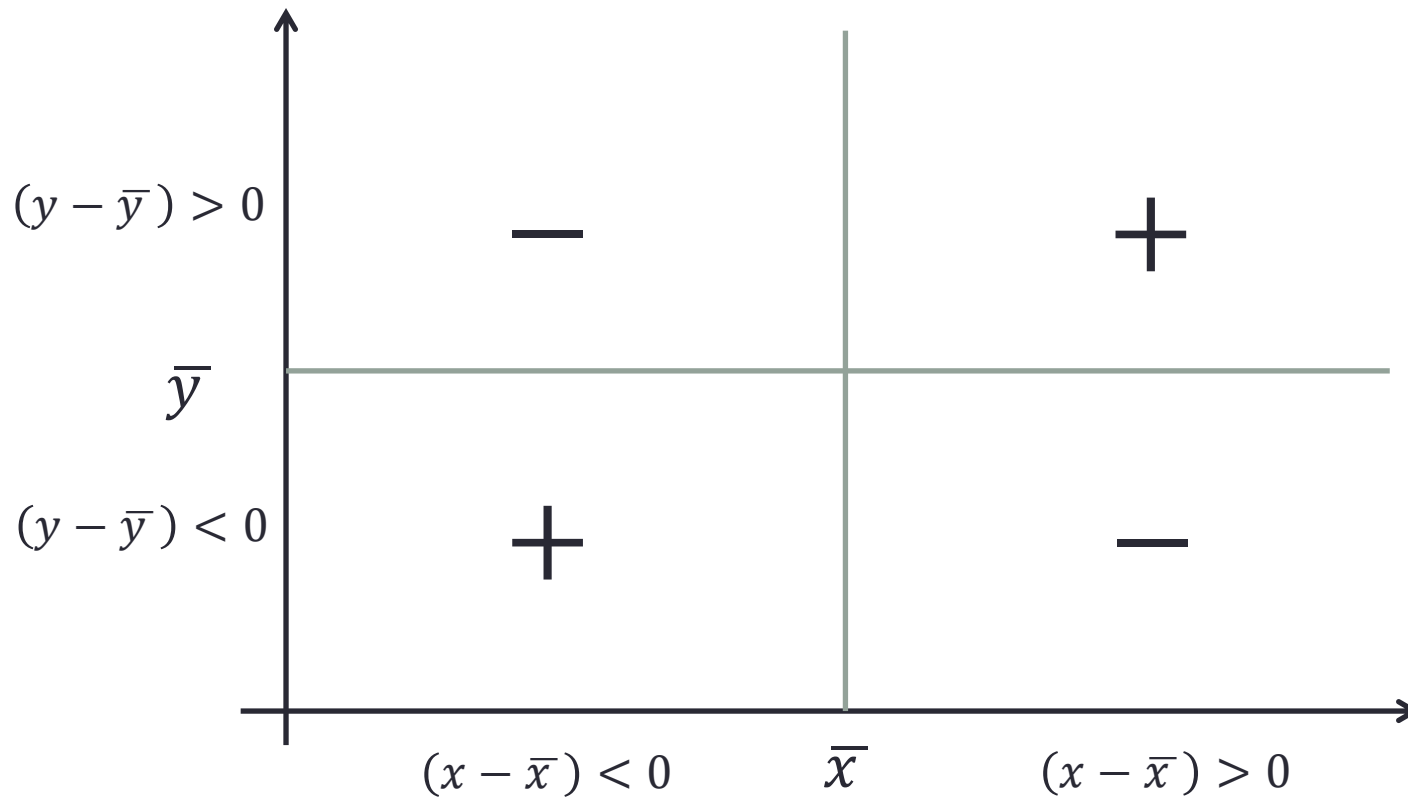
Koeficijent korelacije - Kovarijansa

- Vrednost kovarijanse je osnova za dobijanje vrednosti koeficijenta korelacije.
- Recimo da imamo dve promenljive X i Y i n vrednosti $x(1), \dots, x(n)$ i $y(1), \dots, y(n)$. Kovarijansa za X i Y je:

$$Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x(i) - \bar{x})(y(i) - \bar{y})$$

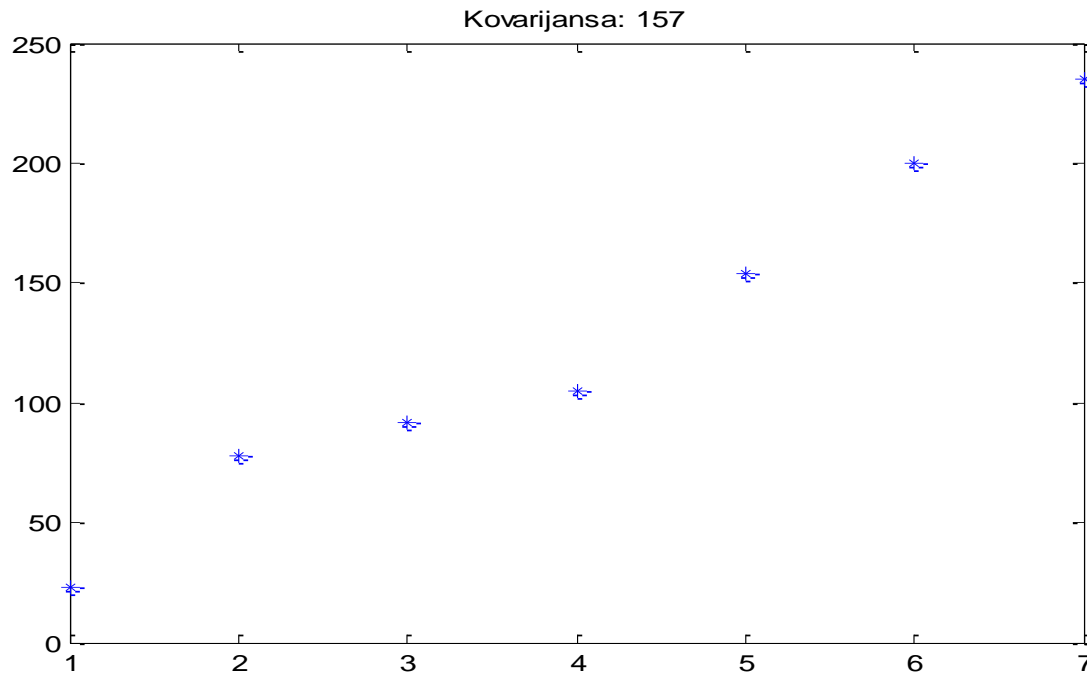
- Kovarijansa meri kako X i Y variraju zajedno:
 - ima pozitivnu vrednost ako velike vrednosti X odgovaraju velikim vrednostima Y i ako male vrednosti X odgovaraju malim vrednostima Y.
 - ima negativnu vrednost ako velike vrednosti X odgovaraju malim vrednostima Y i ako male vrednosti X odgovaraju velikim vrednostima Y.

Kovarijansa



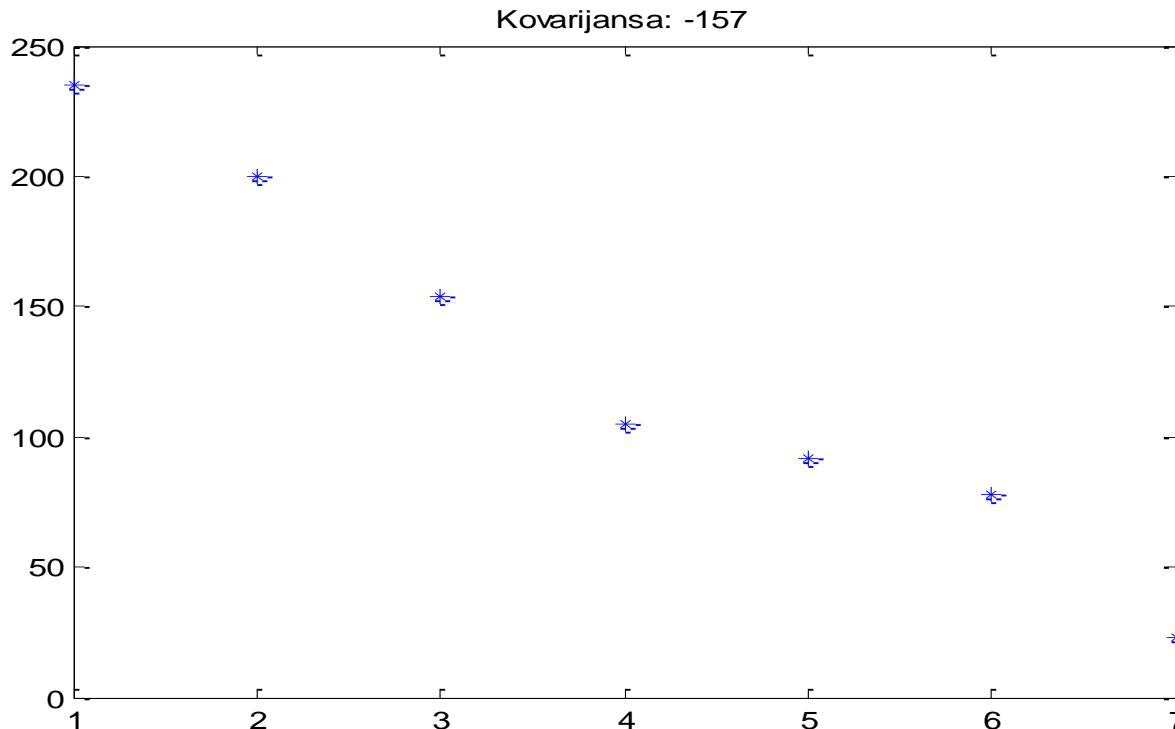
Kovarijansa

- Kovarijansa meri kako X i Y variraju zajedno:
 - ima pozitivnu vrednosti ako velike vrednosti X odgovaraju velikim vrednostima Y i ako male vrednosti X odgovaraju malim vrednostima Y.



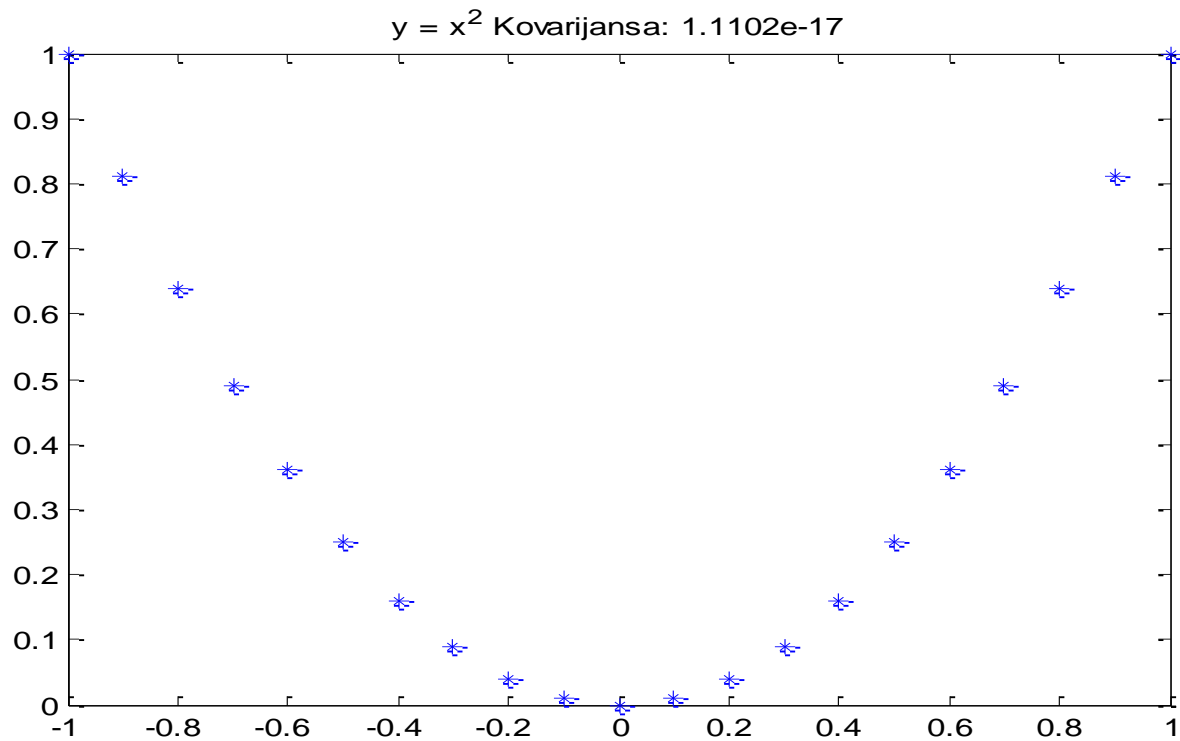
Kovarijansa

- Kovarijansa meri kako X i Y variraju zajedno:
 - ima negativnu vrednosti ako velike vrednosti X odgovaraju malim vrednostima Y i ako male vrednosti X odgovaraju velikim vrednostima Y.



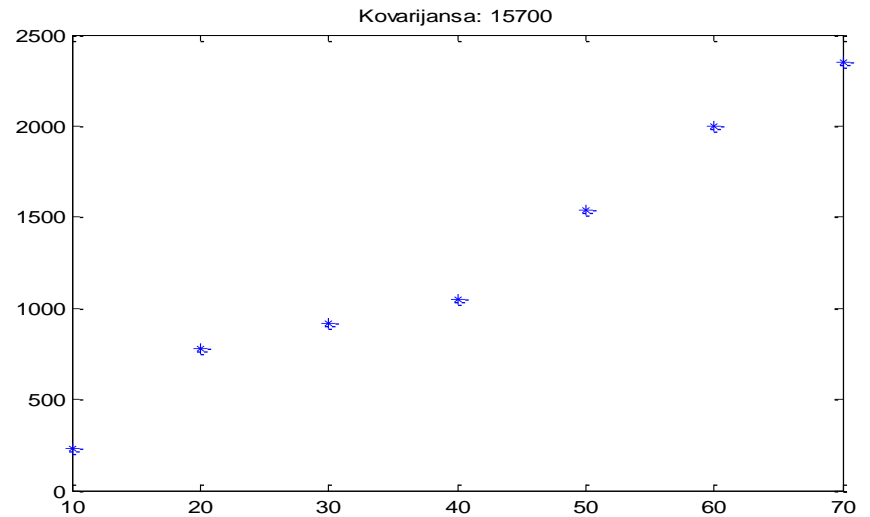
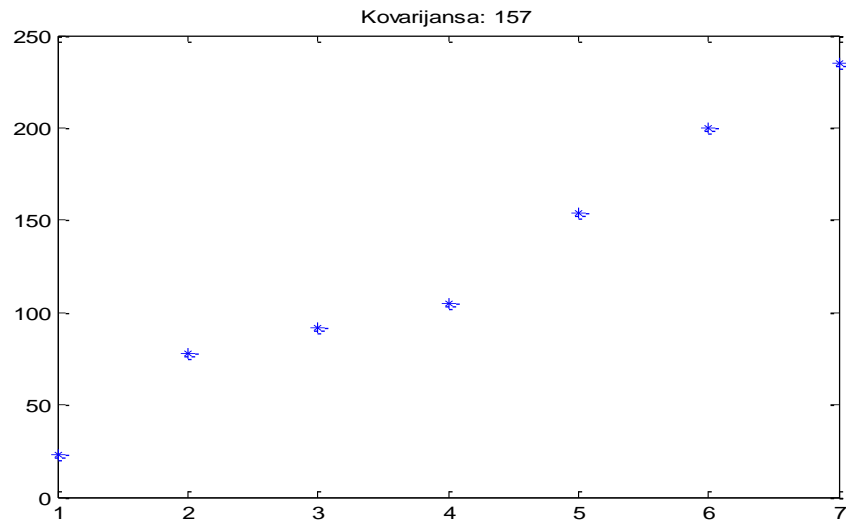
Kovarijansa

- Kovarijansa meri kako X i Y variraju zajedno:
 - Kovarijansa meri linearan odnos



Kovarijansa

- Vrednost kovarijanse zavisi od raspona vrednosti X i Y.



Koeficijent korelacije

- Vrednost kovarijanse zavisi od raspona vrednosti X i Y.
 - Normalizujemo je deljenjem sa standardnom devijacijom
 - Tako dobijamo koeficijent korelacije
 - Koeficijent korelacije definisan je sa:

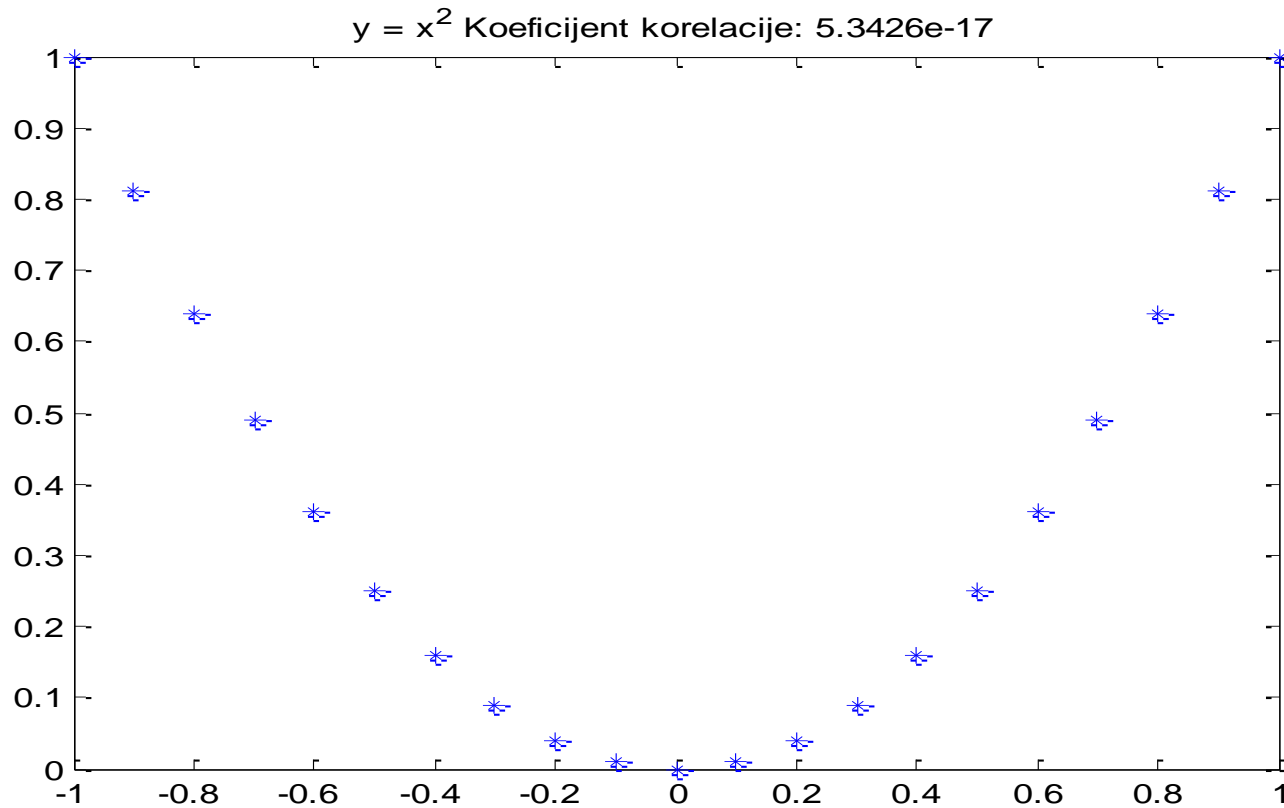
$$\rho(X, Y) = \frac{\sum_{i=1}^n (x(i) - \bar{x})(y(i) - \bar{y})}{\sqrt{\sum_{i=1}^n (x(i) - \bar{x})^2} \sqrt{\sum_{i=1}^n (y(i) - \bar{y})^2}}$$

Koeficijent korelacije

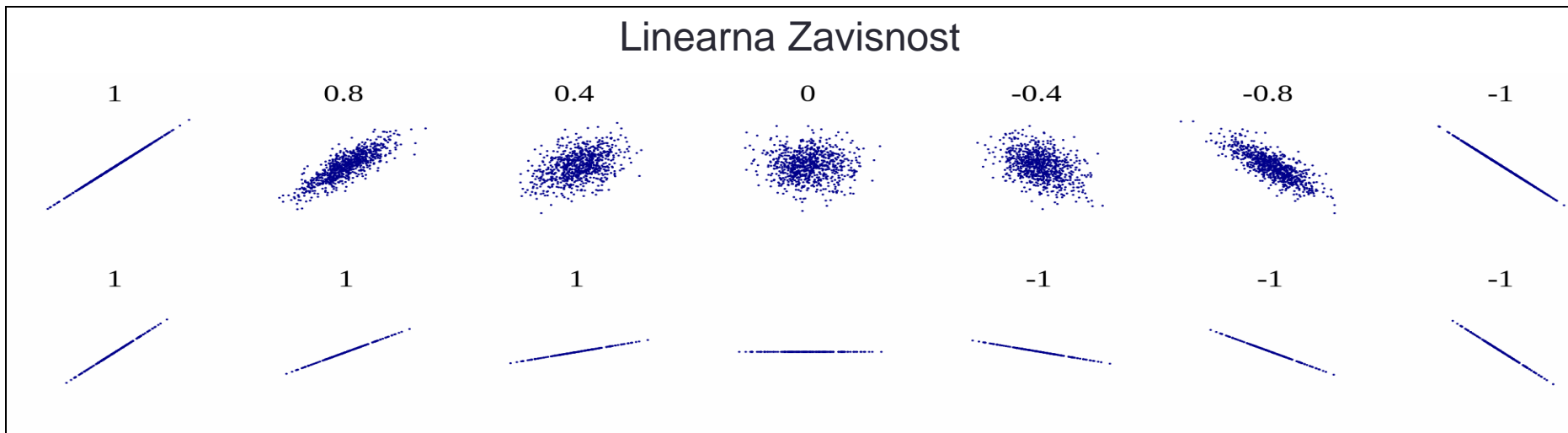
- Koeficijent korelacije meri linearan odnos
- Raspon vrednosti je $[-1, 1]$
- Vrednost 1 je indikator savršene pozitivne kolinearnosti
- Vrednost -1 je indikator savršene negativne kolinearnosti

Koeficijent korelacije - napomena

- Napomena da koeficijent korelacije meri linearan odnos

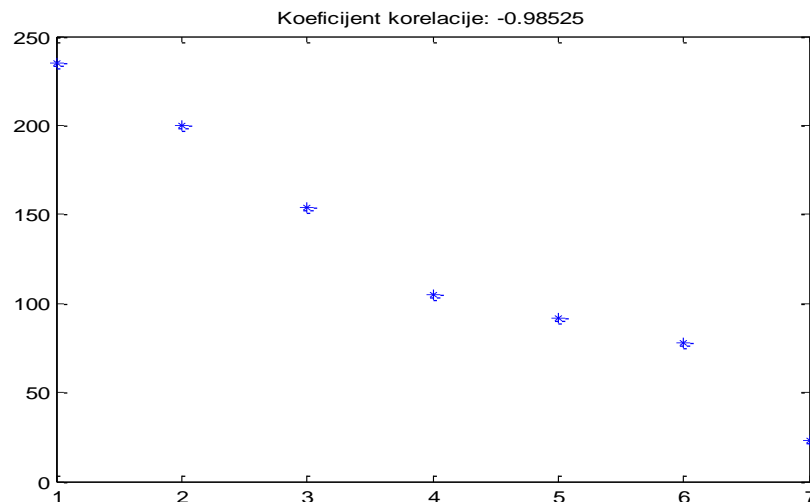
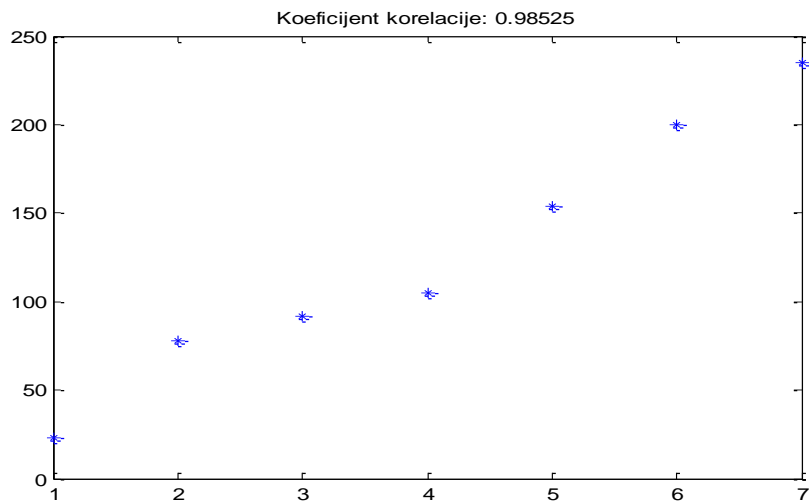


Koeficijent korelacije – primeri dijagrama rasipanja



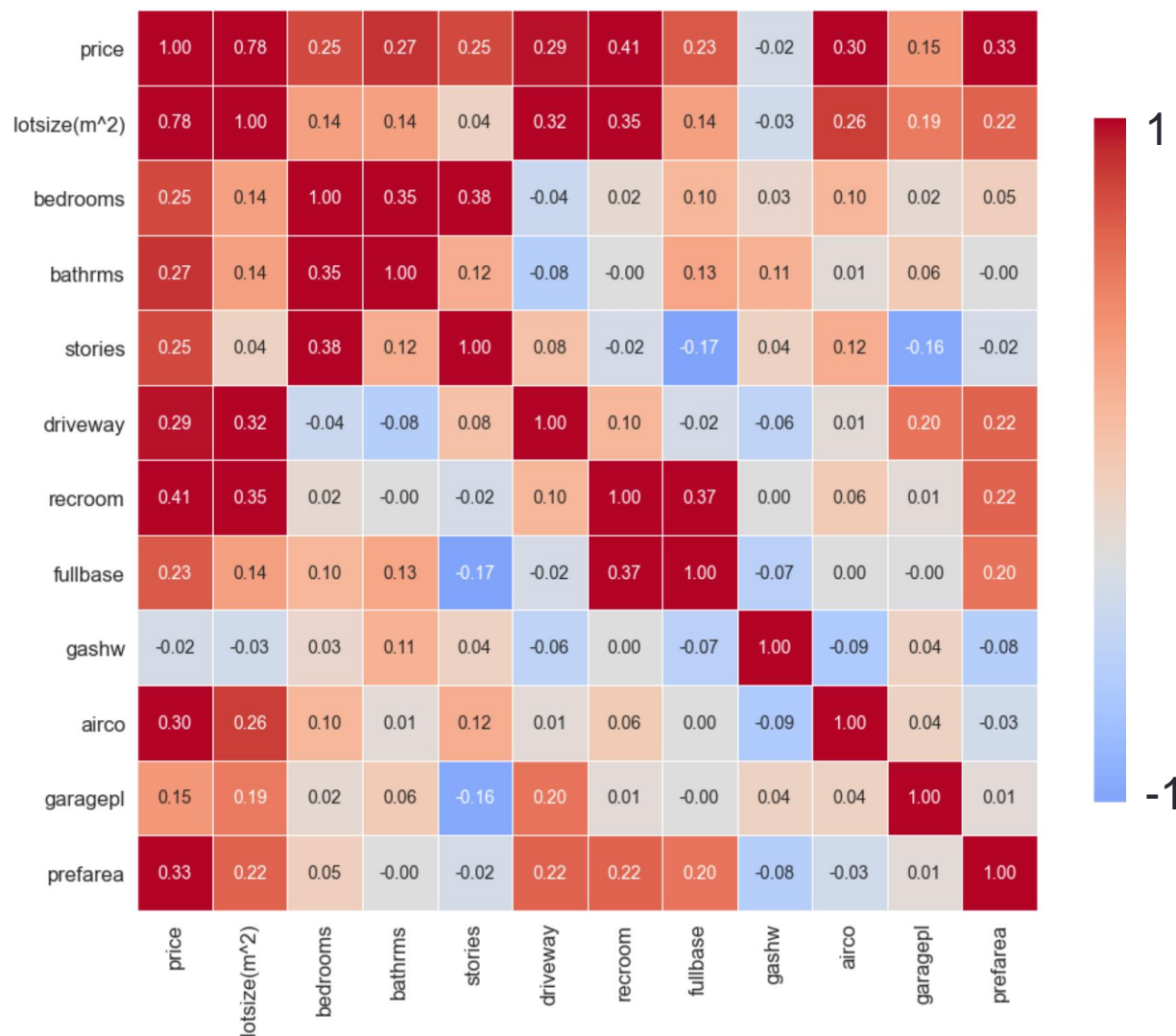
Koeficijent korelacije – velika ali ne savršena kolinearnost

- Vrednosti koeficijenta korelacije imaju mnogo tumačenja i zavise od toga šta je predmet proučavanja. Za prirodne nauke tipično važi:
 - $[0.8, 1]$ – jaka korelacija
 - $[0.6, 0.8)$ – srednja korelacija
- U nastavku se kratko bavimo jakom ali ne savršenom korelacijom u linearnoj regresiji.



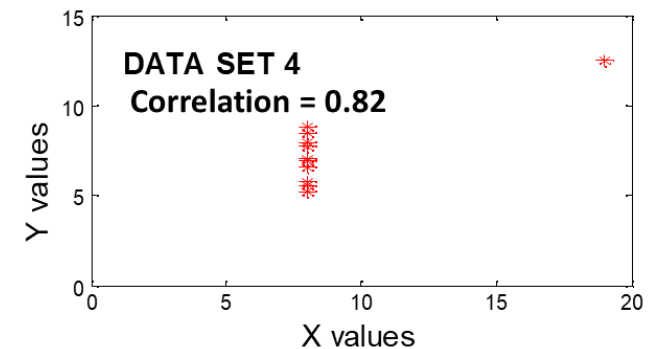
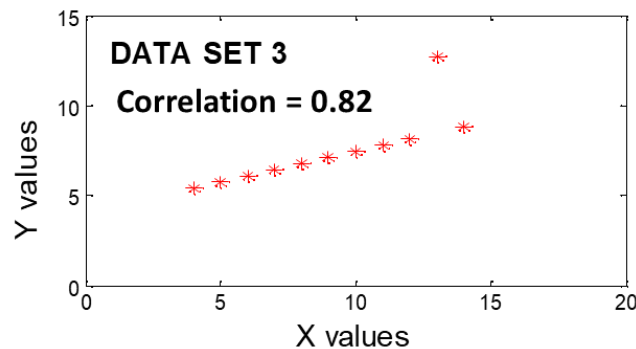
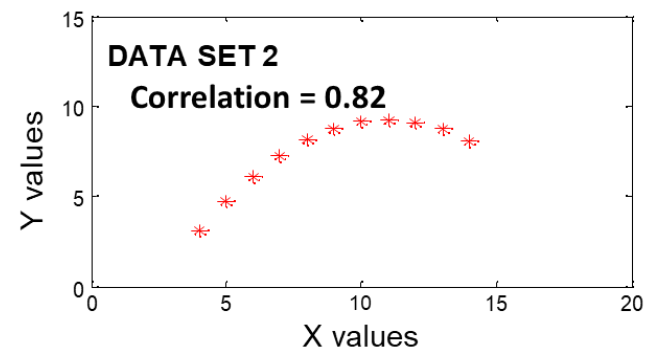
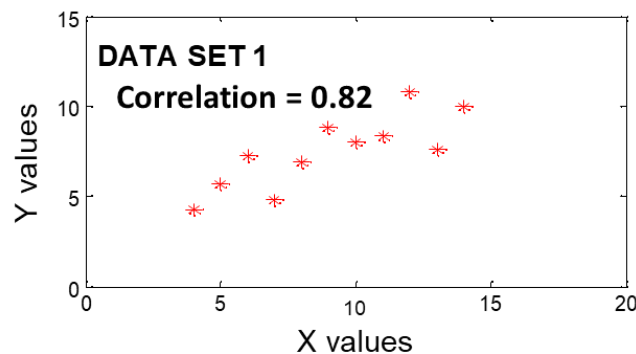
Velika ali ne savršena kolinearnost – Matrica korelacija

- Kod višestruke regresije jaku korelaciju tipično detektujemo pomoću **matrice korelacije**.
- Matrica korelacija pokazuje vrednosti koeficijenata korelacije za sve parove promenljivih.



Velika ali ne savršena kolinearnost – napomena

- Ako uočimo jaku korelaciju **važno je pogledati dijagram rasipanja uočenih promjenljivih.**
- Sama vrednost koeficijenta korelacije nije dovoljna. Na četiri grafika ispod prikazana su četiri skupa podataka* koji **svi imaju isti koeficijent korelacije.**



Velika ali ne savršena kolinearnost – posledice po rezultate regresije

- Velika kolinearnost nije problematična za MNK.
- Parametri koji su određeni pomoću MNK biće tačni i najbolji mogući u smislu sume kvadrata grešaka.
- Intuicija nam kaže da će možda uticaj jedne informacije sadržane u više promenljivih biti preveliki na model.
- U takvim slučajevima obično mislimo da je ispravno da izbacimo jednu od promenljivih.
- Međutim to može imati ozbiljne posledice na model i samo istraživanje iz koga je model proistekao.
- Na sledećem slajdu ćemo to pokazati kroz primer.

Velika ali ne savršena kolinearnost – posledice po rezultate regresije

- Recimo da modelujemo potrošnju jedne porodice na nivou godine i da model ima sledeći oblik:

$$potrosnja = \beta_2 \cdot prihodi + \beta_1 \cdot uštedjevina + \beta_0$$

- Recimo da imamo uzorak kod koga su prihodi i ušteđevina u jakoj korelaciji.
 - Jer tipično porodice koje imaju veliku ušteđevinu imaju i velike prihode.
- Ako bi iz modela izbacili *prihode* ili *ušteđevinu* napravili bi grešku u modelovanju.
- Model koji uključuje oba faktora je korektan jer potrošnja zavisi od oba ova faktora, moguće je samo da naš uzorak podataka ne sadrži dovoljno porodica koje imaju veliku ušteđevinu i male prihode i obrnuto.
- Ako želimo da naše istraživanje bude ispravno i da ne odbacujemo faktore iz modela bolje je **da ne radimo ništa** ili **ako je moguće pribavimo dodatne podatke**.

t-test kod višestruke regresije

- Podsetimo se t-test nam je kod jednostruke regresije služio da statističkim testom proverimo da li je $\beta_1=0$ odnosno da li postoji linearna veza između X i Y.
- Na isti način koristimo t-test u višestrukoj regresiji, samo što se radi za svaku nezavisnu promenljivu posebno.
- Takođe, ako je p-vrednost ≤ 0.05 smatra se da postoji linearna veza između te nezavisne promenljive i Y.
- Međutim, interpretacija t-testa je sada malo drugačija.

t-test kod višestruke regresije – interpretacija

- Kod višestruke regresije t-testom testiramo da li postoji statistički značajna linearna veza date nezavisne promenljive sa zavisnom, **ali u prisustvu svih drugih promenljivih u modelu.**
- Dakle, moguće je da neka promenljiva u prisustvu drugih više nema značaj, iako bi u slučaju jednostruke regresije (sa samo tom promenljivom) imala značaj.
- Pogledajmo primer na sledećem slajdu.

t-test kod višestruke regresije – Primer 1/2

- Vidimo da broj spavaćih soba (bedrooms) nije prošao t-test u višestrukoj regresiji, dok u jednostrunoj jeste.

	coef	std err	t	P> t	[0.025	0.975]
const	4.853e+04	3069.052	15.813	0.000	4.25e+04	5.46e+04
bedrooms	4399.0195	1024.645	4.293	0.000	2381.971	6416.068

	coef	std err	t	P> t	[0.025	0.975]
const	1.912e+04	2203.038	8.679	0.000	1.48e+04	2.35e+04
lotsize(m^2)	63.1780	3.862	16.359	0.000	55.574	70.782
bedrooms	41.8319	639.285	0.065	0.948	-1216.828	1300.492
bathrms	4348.5315	1033.176	4.209	0.000	2314.359	6382.705
stories	3903.9671	630.475	6.192	0.000	2662.652	5145.282
driveway	1371.5360	1234.256	1.111	0.267	-1058.535	3801.607
recroom	3985.9209	1244.683	3.202	0.002	1535.319	6436.522
fullbase	2492.1516	912.742	2.730	0.007	695.094	4289.209
gashw	148.2952	1988.473	0.075	0.941	-3766.721	4063.311
airco	3045.3796	938.757	3.244	0.001	1197.104	4893.655
garagepl	671.0931	526.604	1.274	0.204	-365.713	1707.899
prefarea	4289.9797	982.849	4.365	0.000	2354.892	6225.068

t-test kod višestruke regresije – Primer 2/2

- Razlog za veliku p-vrednost za *bedrooms* je najverovatnije u tome što je u korelaciji sa *bathrms* i *stories*.

	coef	std err	t	P> t	[0.025	0.975]
-----	-----	-----	-----	-----	-----	-----
const	4.255e+04	3232.381	13.164	0.000	3.62e+04	4.89e+04
bedrooms	1848.2657	1129.879	1.636	0.103	-376.009	4072.541
bathrms	6511.4425	1809.861	3.598	0.000	2948.557	1.01e+04
stories	3283.0483	1068.564	3.072	0.002	1179.477	5386.620

- Ako su pretpostavke regresije zadovoljene onda rezultate ovog t-testa kao istraživač koga zanima šta utiče na cenu kuće možemo da interpretiramo na sledeći način:
 - Informacija o broju spavaćih soba kuće nema statistički značajan linearni uticaj na cenu kuće ako nam je poznat broj spratova i kupatila.

F-test

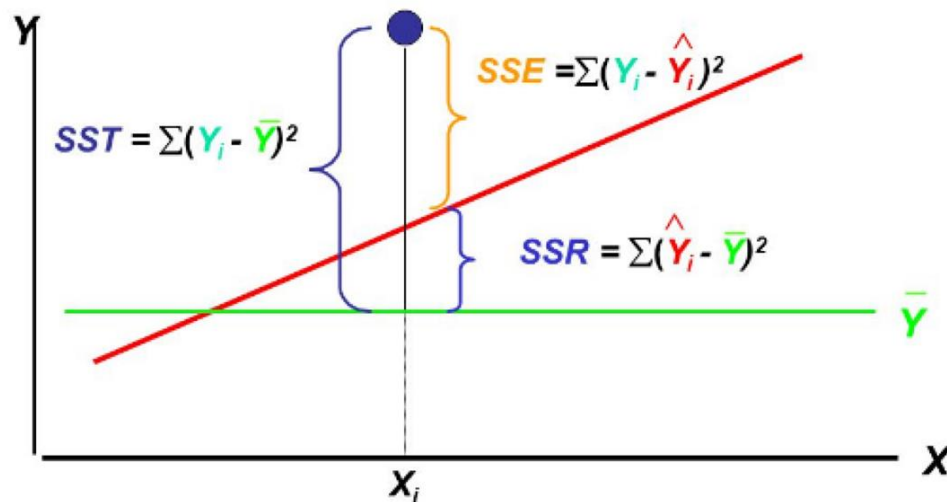
- Dok t-test služi za testiranje značaja pojedinačnih promenljivih, F-test se koristi za testiranje značaja celog modela.
- F-test daje odgovor na pitanje „Da li bar jedna nezavisna promenljiva u modelu ima statistički značajnu linearnu vezu sa zavisnom promenljivom?“.
- Hipoteza koje se testira pomoću F-testa je:
$$\beta_1 = \beta_2 = \dots = \beta_k = 0,$$
- gde je k broj nezavisnih promenljivih u modelu.
- Testiramo da li je bar jedna vrednost β_i različita od nule.

F-test - Informativno

- F-test je takođe razlomak dve komponente signala i buke (*noise*):

$$\frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n - k - 1}}$$

- gde je k broj nezavisnih promenljivih u modelu, a n broj primera u podacima.
- Signal je MSR odnosno **varijabilnost podataka koja je objašnjena pomoću modela**.
- Buka je MSE odnosno **varijabilnost podataka koju model nije mogao da objasni**.



F-test, MSR i MSE – Komentar - Informativno

- F-test je takođe razlomak dve komponente signala i buke (*noise*):

$$\frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n - k - 1}}$$

- Imenioci u formuli za MSR i MSE su *stepeni slobode*.
- Kompletno teorijsko objašnjenje stepena slobode je van opsega ovog kursa, ali će u nastavku biti dato intuitivno objašnjenje.

F-vrednosti i F-distribucija - Informativno

- Odnos signala i buke u F-testu je F-vrednost:

$$F - \text{vrednost} = \frac{MSR}{MSE}$$

- Istraživači su simulirali veliki broj uzoraka podataka kod kojih važi:

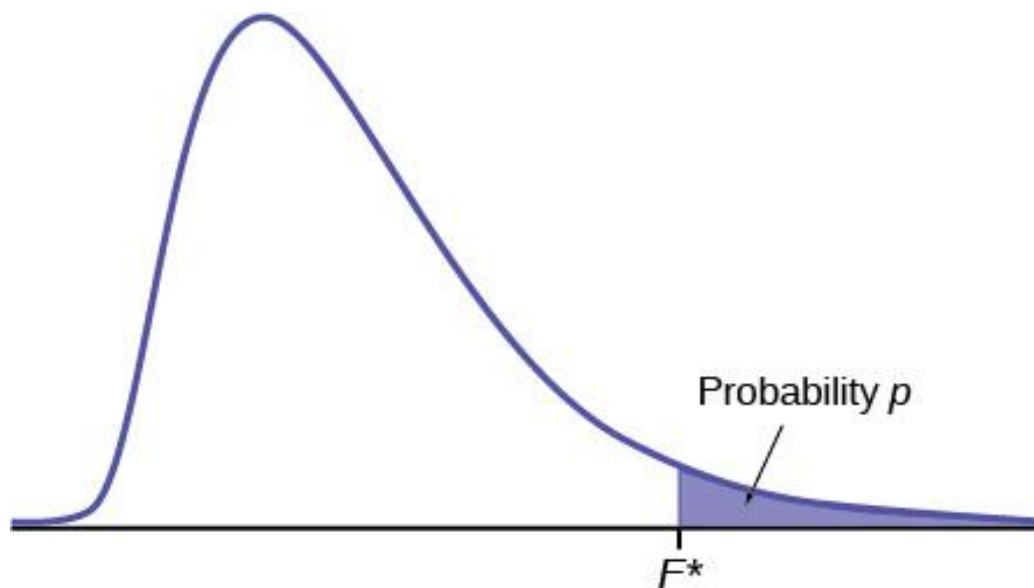
$$\beta_1 = \beta_2 = \dots = \beta_k = 0$$

- i na taj način shvatili kako su distribuirane F-vrednosti kad hipoteza $\beta_1 = \beta_2 = \dots = \beta_k = 0$ važi.

- Ta distribucija zove se F-distribucija.

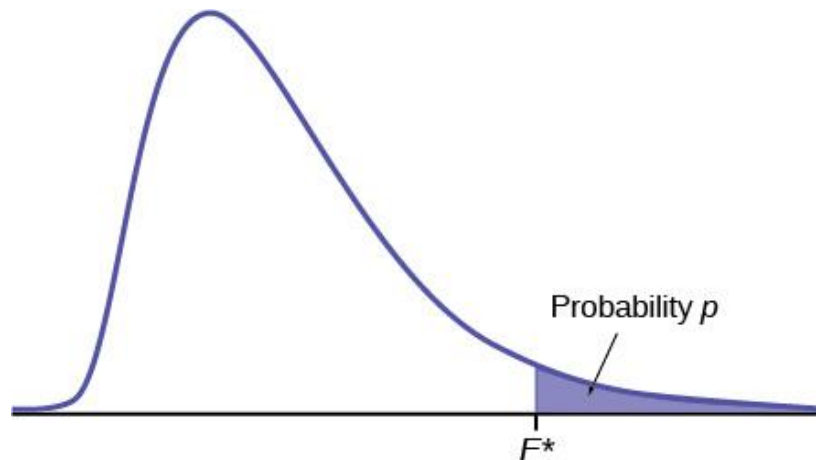
F-distribucija, p-vrednost - Informativno

- Kada imamo F-vrednost i F-distribuciju onda određujemo **p-vrednost** (*p-value*).
- P-vrednost je ukupna verovatnoća da ćemo iz F-distribucije izvući našu F-vrednost ili neku još manje verovatnu (obojen deo distribucije na slici).



F-distribucija, p-vrednost – Primer – Informativno

- Želimo što manju p-vrednost. Prag koji se koristi u praksi je **0.05**.
 - Ako je $p\text{-vrednost} \leq 0.05$ onda zaključujemo da signal postoji, odnosno da je bar jedno $\beta_i \neq 0, i=1\dots k$.
- Veći deo distribucije zauzimaju F-vrednosti koje bi dobili da važi $\beta_1 = \beta_2 = \dots = \beta_k = 0$ jer je tako F-distribucija formirana.
- Ako je $p\text{-vrednost} \leq 0.05$ to znači da postoji $\leq 5\%$ verovatnoće da ćemo dobiti našu F-vrednost ako važi $\beta_1 = \beta_2 = \dots = \beta_k = 0$.
 - Odnosno možemo sa 95% sigurnosti da zaključimo da je $\beta_i \neq 0, i=1\dots k$.



F-distribucija, p-vrednost - Informativno

- Pomoću Python biblioteke *statsmodels* dobijamo tabelu ispod.
- Takva tabela naziva se **ANOVA** (*Analysys of Variance*) tabela.

Dep. Variable:	price	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.731
Method:	Least Squares	F-statistic:	70.01
Date:	Thu, 21 Sep 2023	Prob (F-statistic):	3.21e-72
Time:	14:35:08	Log-Likelihood:	-2854.0
No. Observations:	280	AIC:	5732.
Df Residuals:	268	BIC:	5776.
Df Model:	11		

Broj stepeni slobode SSE $n-k-1=280-11-1=268$

Broj stepeni slobode SSR $n-1-(n-k-1)=280-1-(280-11-1)=11$

Broj primera podataka n

F-vrednost

p-vrednost

F-distribucija, p-vrednost - Informativno

- Iz tabele se vidi da je p-vrednost reda veličine 10^{-72} što je značajno manje od 0.05.
- Dakle, odbacujemo hipotezu $\beta_1 = \beta_2 = \dots = \beta_k = 0$ i **zaključujemo da bar jedna nezavisna promenljiva ima statistički značajnu linearnu vezu sa cenom kuća** (zavisnom promenljivom).

Dep. Variable:	price	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.731
Method:	Least Squares	F-statistic:	70.01
Date:	Thu, 21 Sep 2023	Prob (F-statistic):	3.21e-72
Time:	14:35:08	Log-Likelihood:	-2854.0
No. Observations:	280	AIC:	5732.
Df Residuals:	268	BIC:	5776.
Df Model:	11		

F-vrednost

p-vrednost

Broj primera podataka n

Broj stepeni slobode SSE $n-k-1=280-11-1=268$

Broj stepeni slobode SSR $n-1-(n-k-1)=280-1-(280-11-1)=11$