LINEARNA REGRESIJA

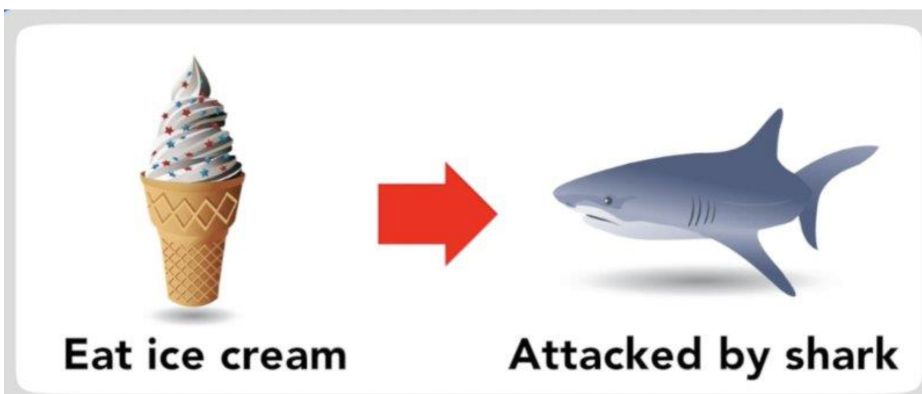
Pretpostavke linearne regresije

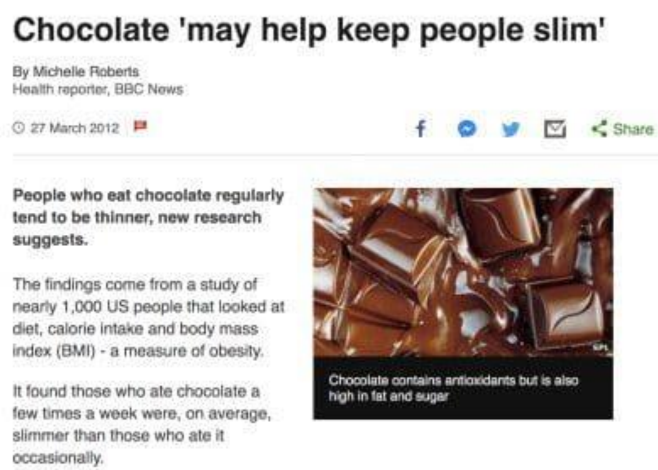
Definicija

- Alat za modelovanje veze zavisne promenljive *Y* sa jednom ili više nezavisnih promenljivih *X*.
- Veza se modeluje kao linearna kombinacija zavisnih promenljivih i parametara (težina).
- Parametri se određuju iz podataka.

Korelacija i Regresija

- Grafik rasipanja (*scatter plot*) je prvi alat za analizu odnosa dve promenljive.
- Analiza korelacije može da se koristi da se izmeri jačina linearne veze između dve promenljive.
  - Korelacija nam samo pruža informaciju o jačini veze.
  - Obe promenljive se tretiraju jednako.
  - Ne analizira se uticaj jedne promenljive ne drugu.





Chocolate 'may help keep people slim'

By Michelle Roberts
Health reporter, BBC News

27 March 2012

Share

People who eat chocolate regularly tend to be thinner, new research suggests.

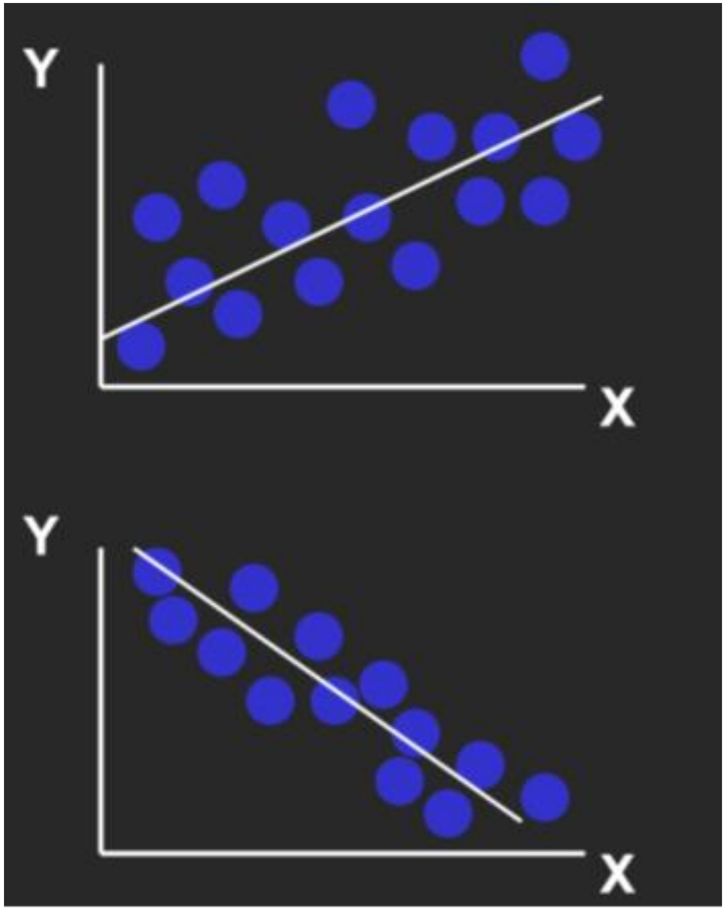
The findings come from a study of nearly 1,000 US people that looked at diet, calorie intake and body mass index (BMI) - a measure of obesity.

It found those who ate chocolate a few times a week were, on average, slimmer than those who ate it occasionally.

Chocolate contains antioxidants but is also high in fat and sugar

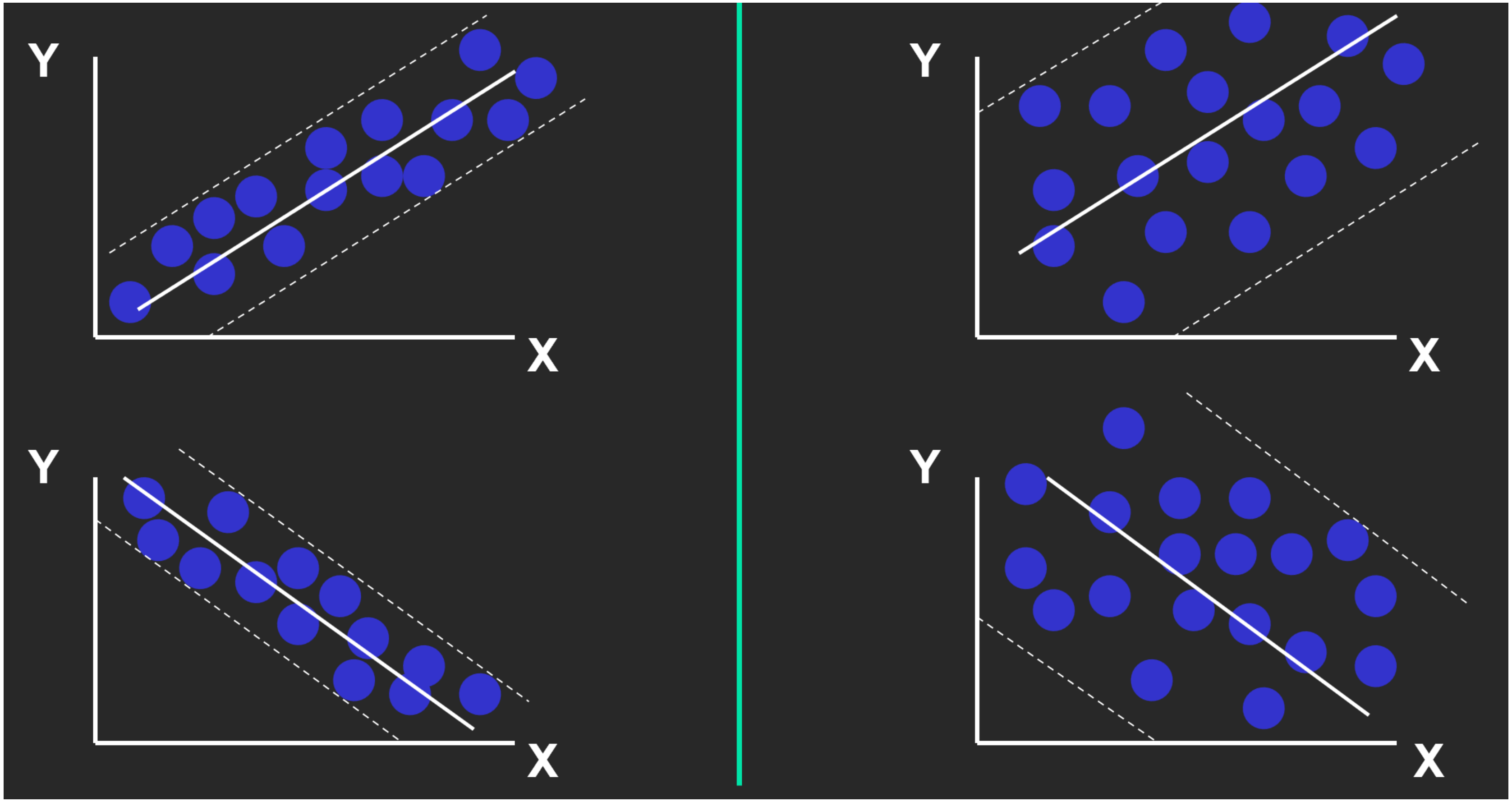
Primeri grafika rasipanja

Linearna veza



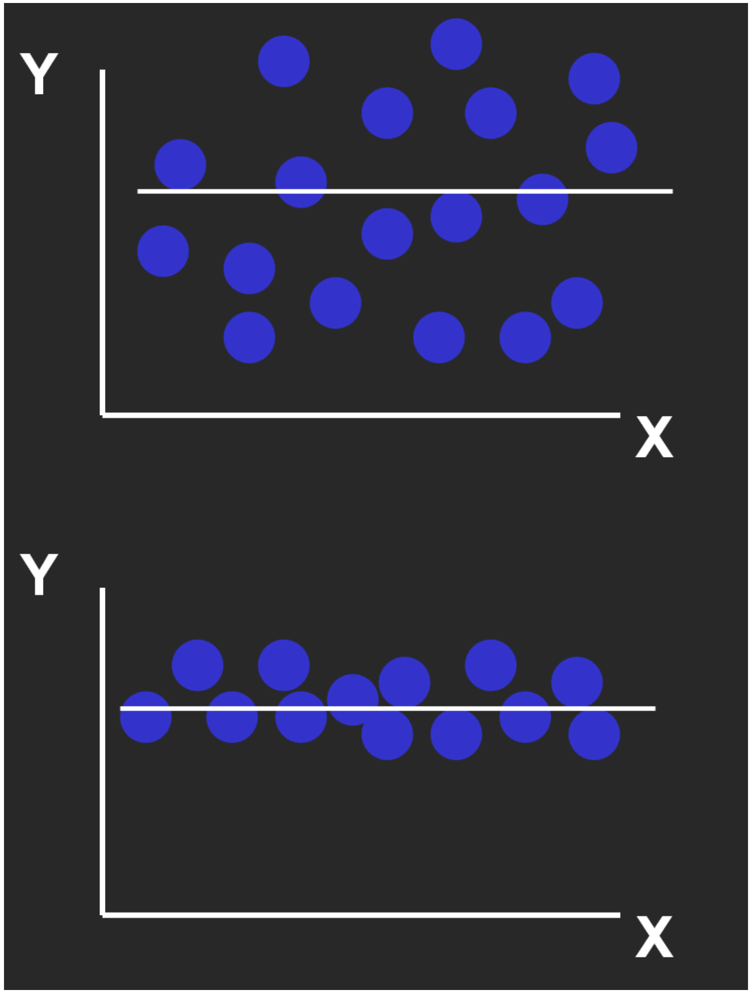
Primeri grafika rasipanja

Jake veze | **Slabe veze**



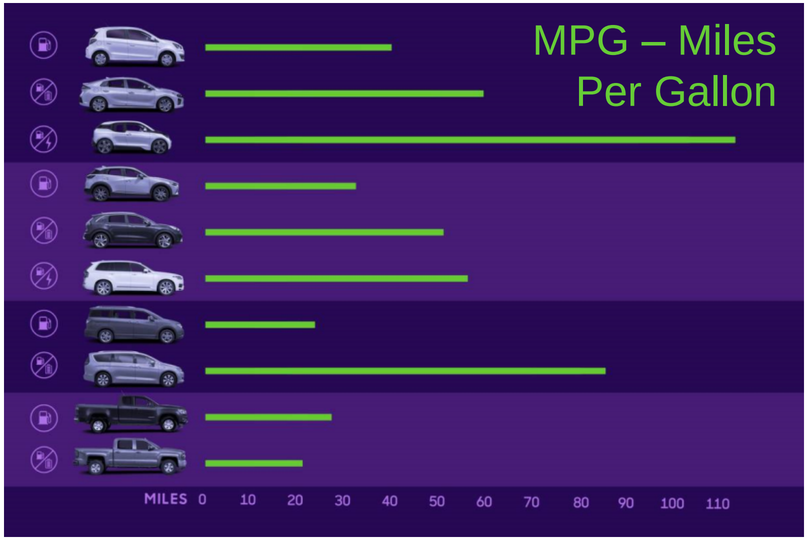
Primeri grafika rasipanja

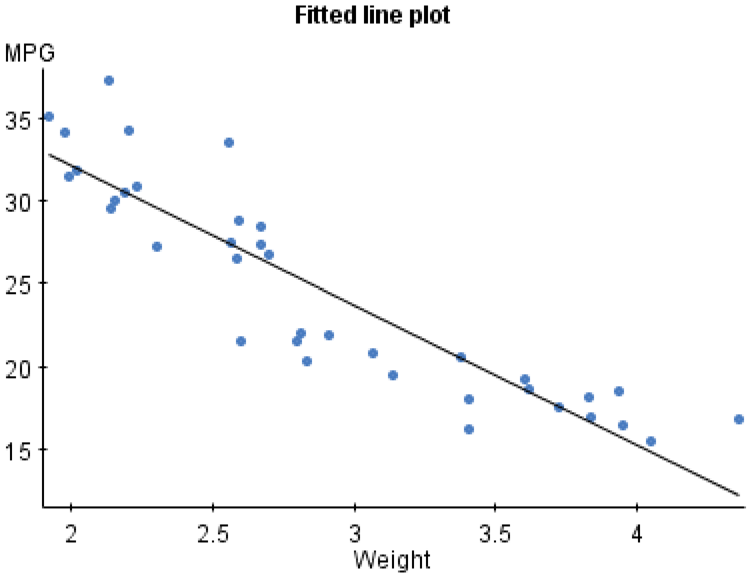
Nema povezanosti



Regresiona analiza

- Regresiona analiza se koristi za:
  - Predikciju zavisne promenljive na osnovu bar jedne nezavisne promenljive.
  - Da se objasni uticaj promene nezavisne promenljive na zavisnu promenljivu.
- Zavisna promenljiva: promenljiva koju želimo da predvidimo ili objasnimo.
- Nezavisna promenljiva: promenljiva koju koristimo za predikciju ili objašnjenje.



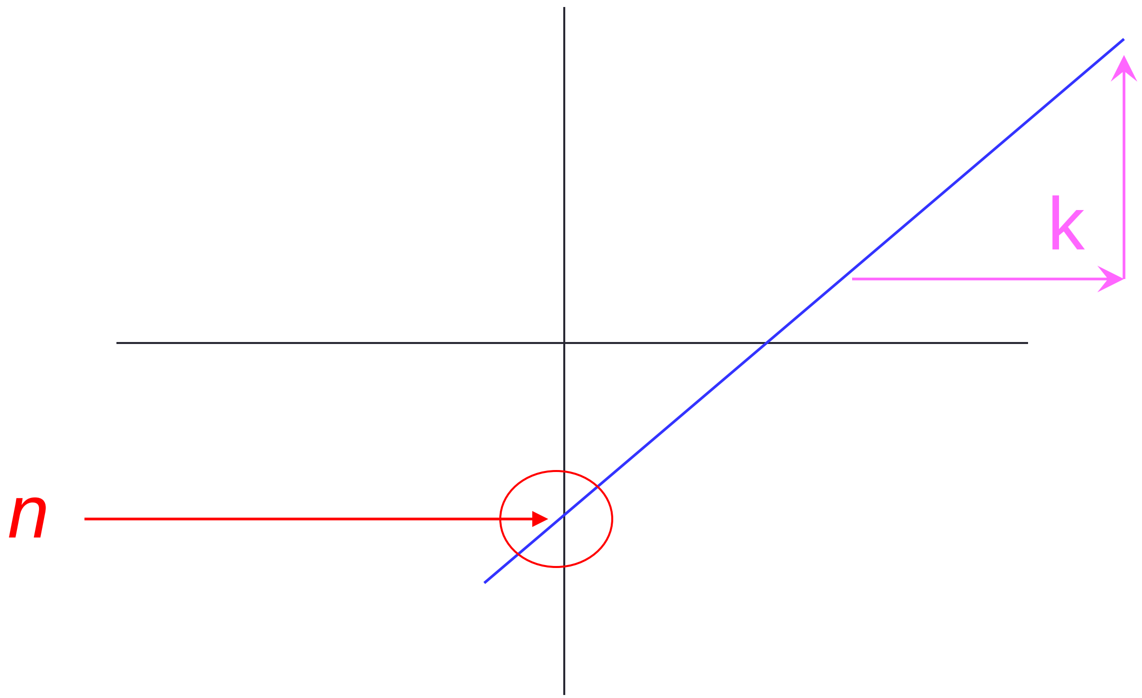


Jednostruka regesija

- Samo **jedna** nezavisna promenljiva, X.
- Veza između X i Y modeluje se kao linearna kombinacija dva parametra i vrednosti X.
- Parametri su nagib i odsečak prave.
- Parametri se određuju iz podataka.

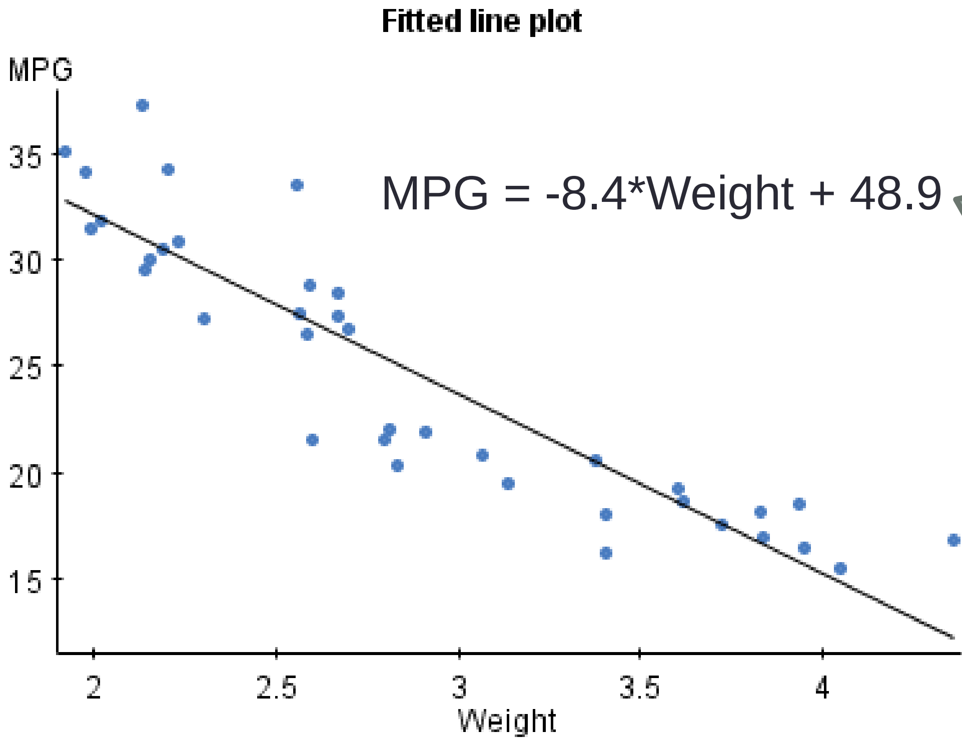
Linearna veza

$Y = kX + n$



Nagib u linearnoj vezi

- Veza između jedinične promene *X* i jedinične promene *Y*.
- Vrednost nagiba od 2: jedna jedinična promena u X daje dve jedinice promene u Y.



U proseku, uvećanje težine kola za 1000 funti smanjuje broj milja koji kola pređu sa jednim galonom goriva za 8.37

Da su težine vozila izražene u funtama umesto u hiljadama funti, nagib bi bio -0.00837

Model jednostruke linearne regresije

Odsečak za celu populaciju

Nagib za celu populaciju

Nezavisna promenljiva

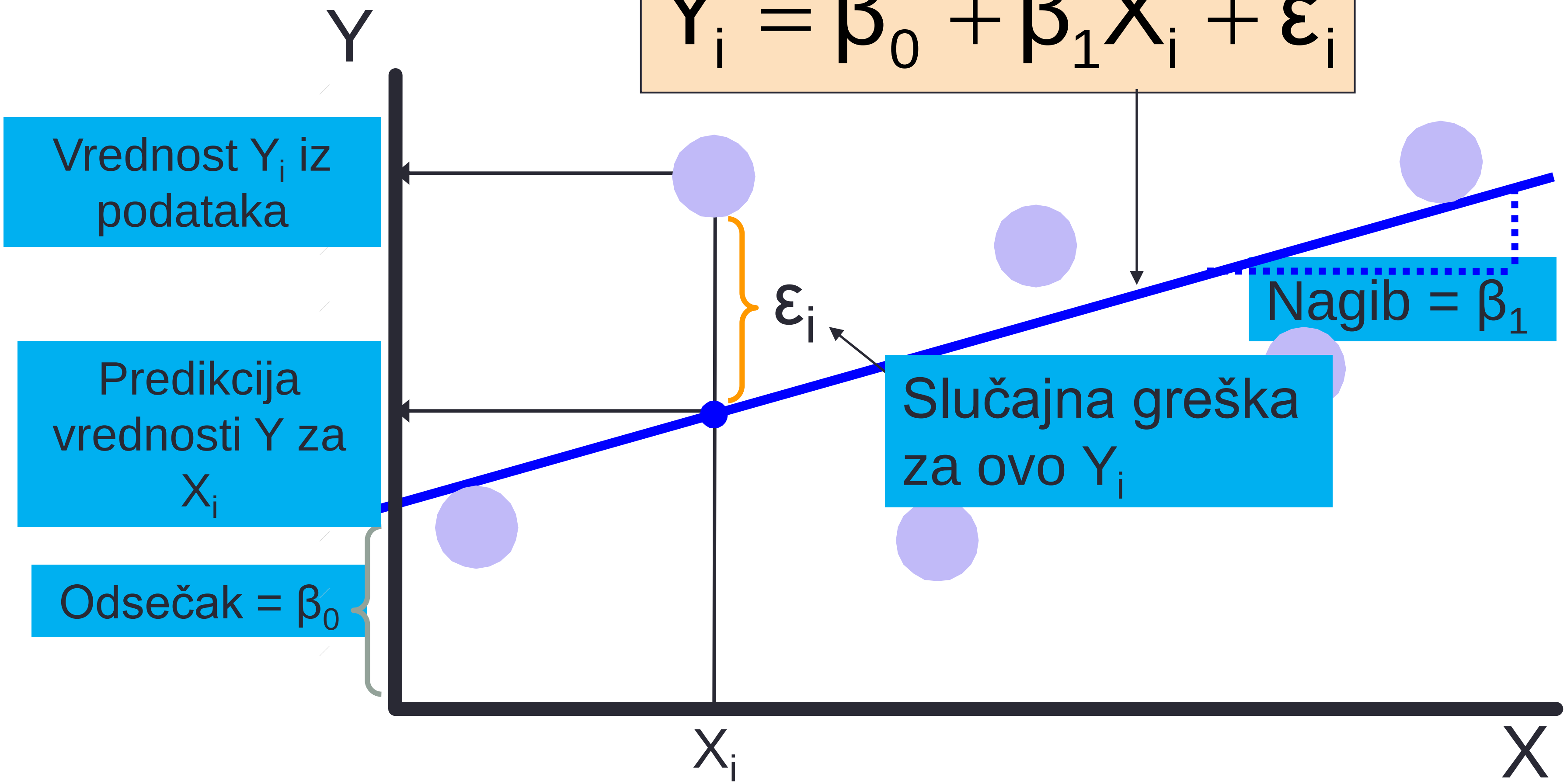
Slučajna greška

Zavisna promenljiva

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linerna komponenta}} + \underbrace{\varepsilon_i}_{\text{Komponenta slučajne greške}}$$

Model jednostruke linearne regresije

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Model jednostruke linearne regresije

- 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 (sempl) populacije.

Model jednostruke linearne regresije – Pouplacija - komentar

- Kada radimo modelovanje pomoću linearne regresije sami biramo šta je populacija.
- U većini slučajeva nije realno da populacija obuhvata sve primere na Zemlji, već je biramo shodno potrebama modelovanja (istraživanja).
- Shodno tome moramo da se pobrinemo i da podaci koje imamo adekvatno reprezentuju odabranu populaciju.
- Na primer, ako nas interesuje da predvidimo cene kuća u Srbiji onda su nam populacija sve kuće u Srbiji (ne na Zemlji).
  - Podaci iz kojih formiramo model bi trebalo da obuhvate različite vrste kuća iz različitih delova Srbije da bi bili reprezentativni.

Model – procene parametara iz uzorka podataka

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

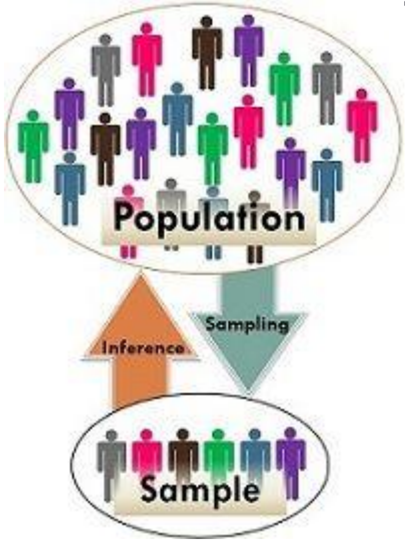
Predikcija i-te vrednosti promenljive Y

Procenjena vrednost odsečka iz uzorka

Procenjena vrednost nagiba iz uzorka

i-ta vrednost promenljive X

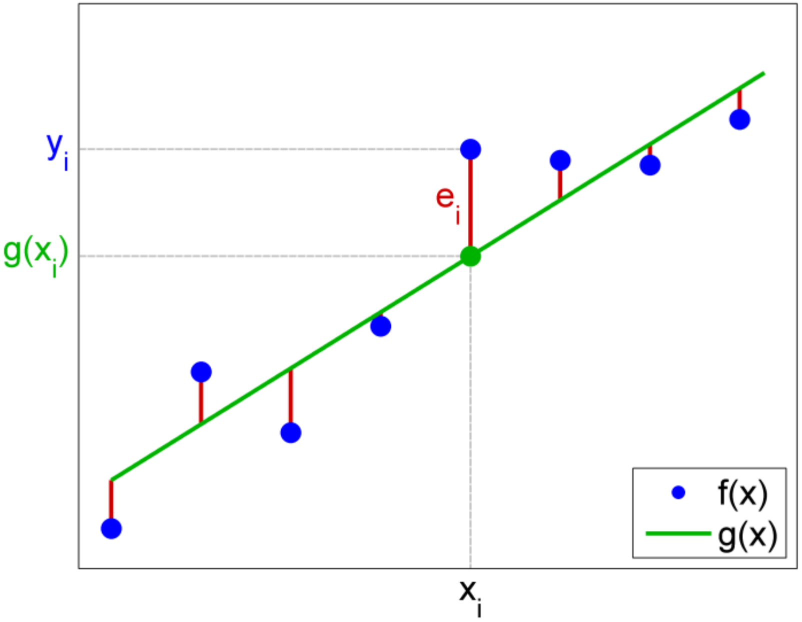
$$\hat{Y}_i = b_0 + b_1 X_i$$

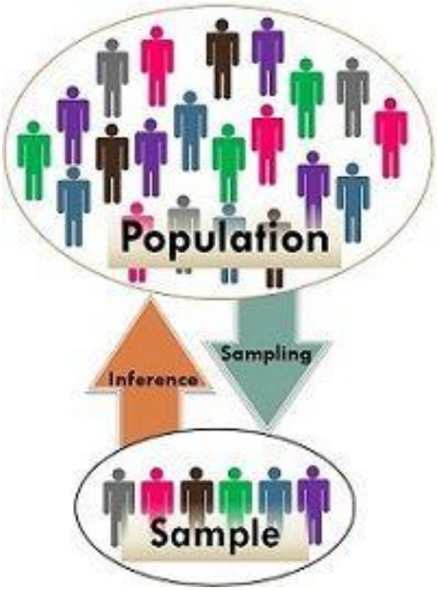


Metod Najmanjih Kvadrata

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

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$





Interpretacija nagiba i odsečka

- b_0 je procenjena (iz podataka) srednja vrednost Y kada je vrednost X nula.

- b_1 je procenjena promena srednje vrednosti Y za povećanje X za jednu jedinicu.

$$\boxed{\hat{Y}_i = b_0 + b_1 X_i}$$

Jednostruka linearna regresija - primer

- Agent za nekretnine želi da ispita uticaj veličine placa na kome se nalazi kuća i cene te kuće.

- Podaci su uzorak od 280 kuća
  - Zavisna promenljiva (Y) = cena kuće u dolarima
  - Nezavisna promenljiva (X) = površina placa u kvadratnim metrima

Jednostruka linearna regresija - primer

- Skup podataka o 280 kuća iz kanadskog grada Windsor (1987)
- Pored cene (*price*) i površine placa (*lotsize*) postoji još 11 atributa:

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

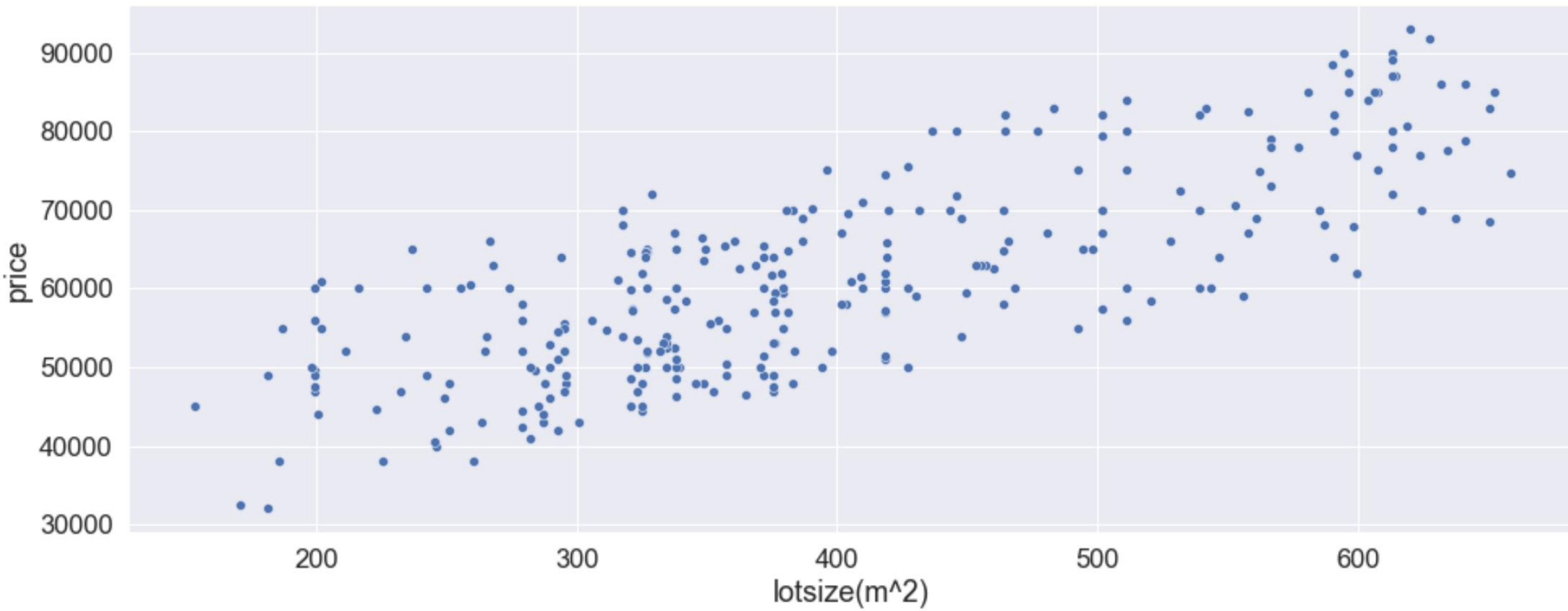
- U ovom primeru koristimo samo cenu i površinu placa, dok ćemo ostale upotrebiti kasnije tokom kursa.

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

Primer – dijagram rasipanja

• Sa dijagrama možemo da primetimo da skuplje kuće imaju veću površinu, i obrnuto.

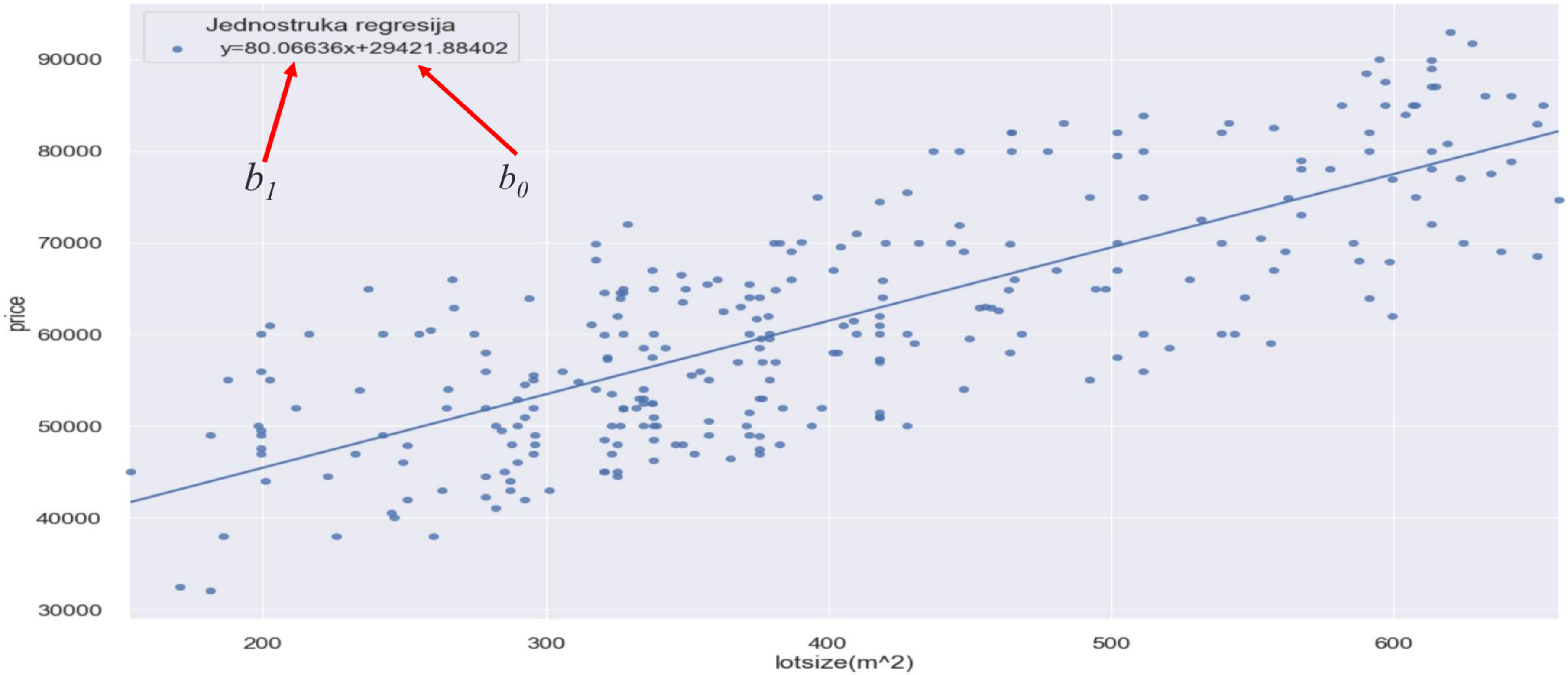


- Takođe vidimo da se većina tačaka ne nalazi na pravoj liniji što znači:
  - da postoji značajna linearna veza (trend), a ne egzaktan linearni odnos između cene kuća i površine placa.

Primer – jednačina jednostruke linearne regresije dobijena iz podataka

- Pomoću metoda *LinearRegression* iz biblioteke *scikit* dobijamo:

Cena = 80.07 * *površina_placa(m^2)* + 29421.88



- Recimo da želimo da predvidimo cenu kuće sa površinom placa od 100m²

Cena = 80.07 * *100* + 29421.88 = 37428.88

- Koliko je dobra naša predikcija? Time se bavimo u nastavku.

Primer – interpretacija osdečka b_o

Cena = 80.07 * *površina_placa(m^2)* + 29421.88

- b_0 procenjena (iz podataka) srednja vrednost Y kada je vrednost X nula.
- Imajući u vidu da kuća sa površinom placa od nula m^2 nema smisla možemo zaključiti da:
  - interpretacija b_0 nema uvek praktičnu vrednost u regresionoj analizi.

Primer – interpretacija nagiba b_1

Cena = 80.07 * *površina_placa(m^2)* + 29421.88

- b_1 je procenjena promena srednje vrednosti Y za povećanje X za jednu jedinicu.
- Za ovaj primer vrednost 80.07 kaže nam da se prosečna cena kuće poveća za 80.07 dolara kada se površina placa poveća za 1 m^2.

Evaluacija modela linearne regresije

- Postoji mnogo načina za evaluciju prediktivnih modela.

- Jedan od načina je primena modela na nepoznate podatke i onda merenje njegovih prefomansi npr. pomoću srednje vrednosti kvadrata grešaka.
  - Ovaj način je univerzalan za sve prediktivne modele.
  - Ovaj način obradićemo detaljno pomoću demonstracija na vežbama.

- Na ovom predavanju prikazaćemo proces evaluacije koji je specifičan za linearnu regresiju.

Evaluacija modela linearne regresije

- Na ovom predavanju pokazaćemo kako možemo da proverimo:
  1. Da li populacioni model dobro modeluje populaciju?
     Odnosno, da li je pretpostavka o tome da postoji linearna veza između X i Y tačna ili pogrešna?
  2. U kom intervalu realnih brojeva se nalaze parametri populacionog modela?
  3. U kom intervalu realnih brojeva se nalaze predikcije populacionog modela?

- Provere vršimo koristeći uzorak podataka koji smo upotrebili za formiranje našeg modela i statističke alate.

- Proces vršenja provera naziva se zaključivanje o modelu linearne regresije.

- Da bi zaključivanje bilo validno moraju da važe Pretpostavke Linearne Regesije – koje prikazujemo u nastavku.

Pretpostavke linearne regresije

- Model predstavlja celu populaciju, dok parametre modela procenjujemo iz jednog uzorka populacije pomoću Metoda Najmanjih Kvadrata (MNK).

- Metoda Najmanjih Kvadrata ćemo u ovom kontekstu zvati estimator parametara.

- Pretpostavke date u nastavku odnose se na kombinaciju estimatora i samog modela linearne regresije, zato se često nazivaju:

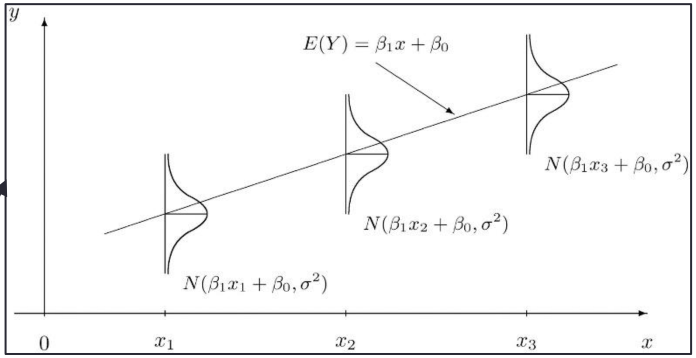
  Pretpostavke MNK linearne regresije

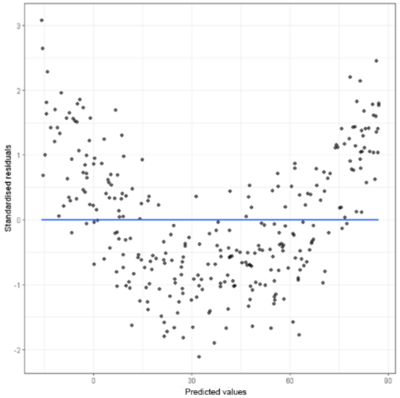
- Dok predstavljamo pretpostavke naglasićemo koje se odnose samo na estimator, a koje i na esitmator i na model.

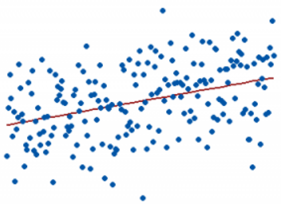
Pretpostavke MNK Linearne regresije L.I.N.E

- *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 drugo 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

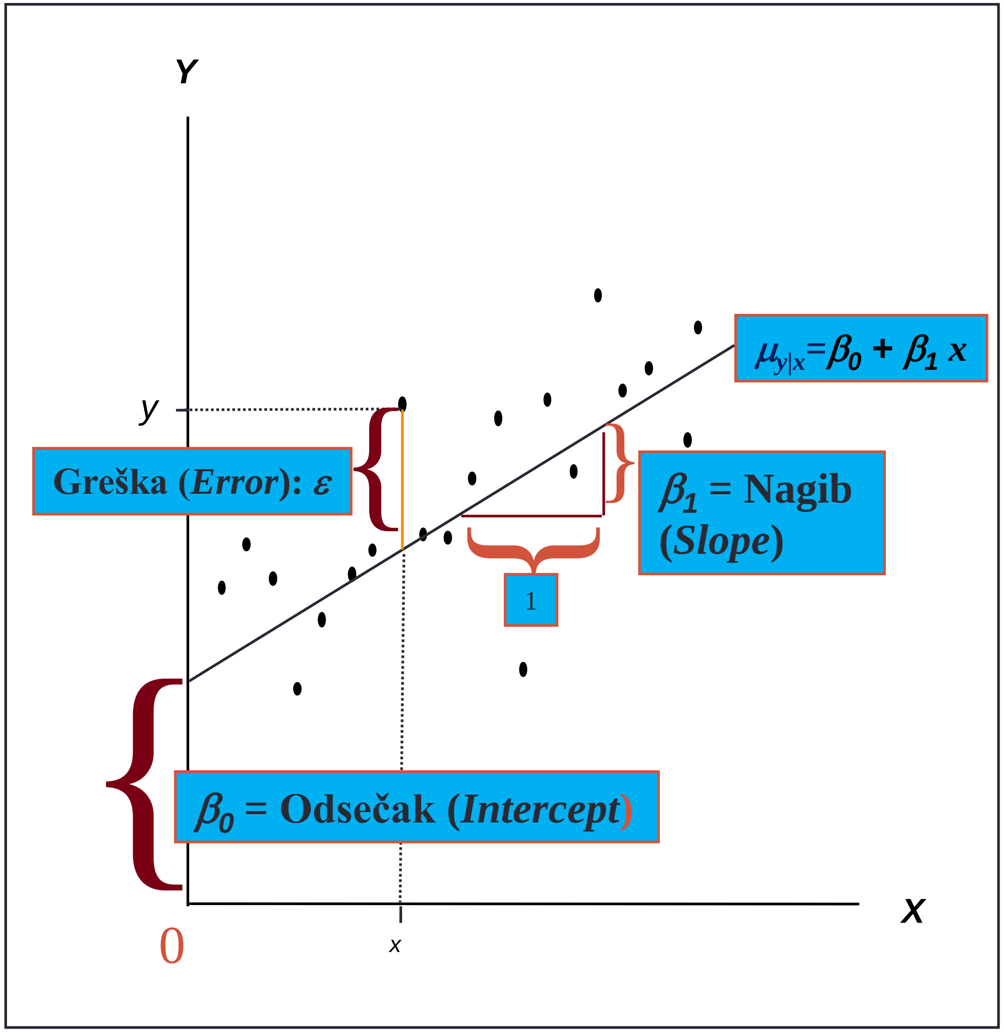
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$







Pretpostavke MNK Linearne regresije
Vizualizacija modela – podsećanje



Model linearne regresije modeluje vezu očekivane vrednosti Y za dato X:

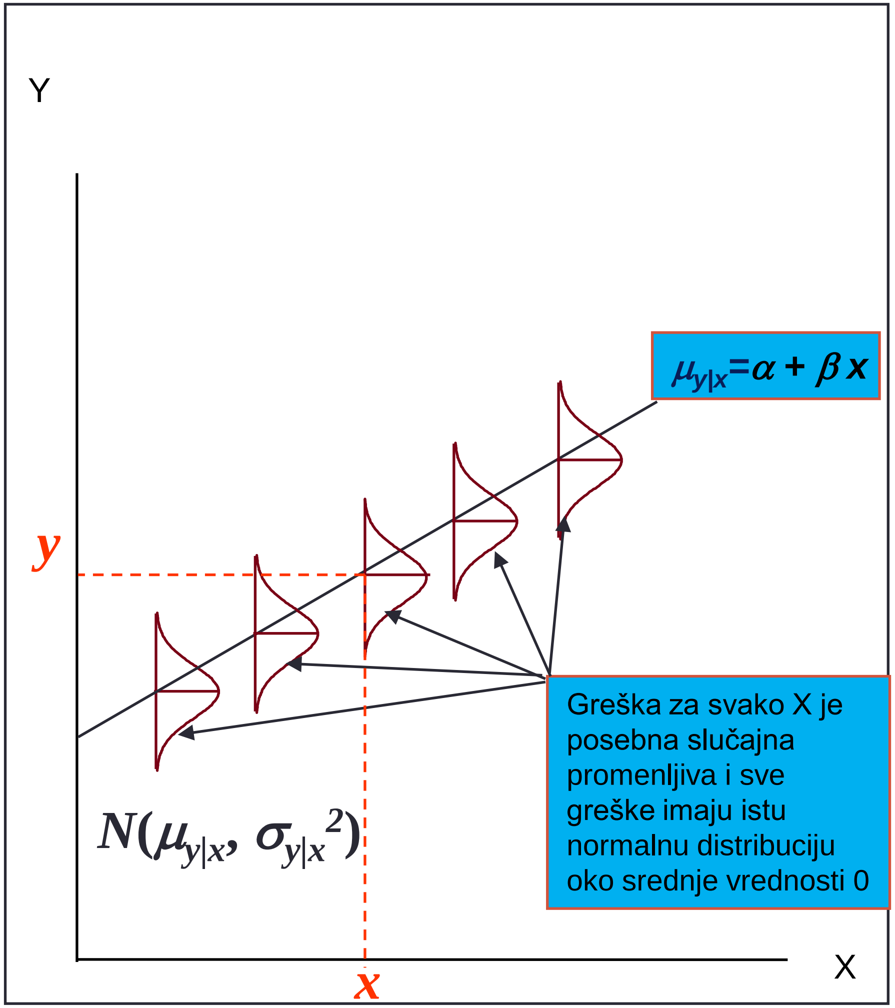
$$\mu_{y|x} = \beta_0 + \beta_1 x$$

Stvarne (tačne vrednosti) Y (y) razlikuju se od očekivane vrednosti ($\mu_{y|x}$) za grešku koju ne može da objasni regresioni model i koja je slučajna promenljiva (ε):

$$y = \mu_{y|x} + \varepsilon$$
$$= \beta_0 + \beta_1 x + \varepsilon$$

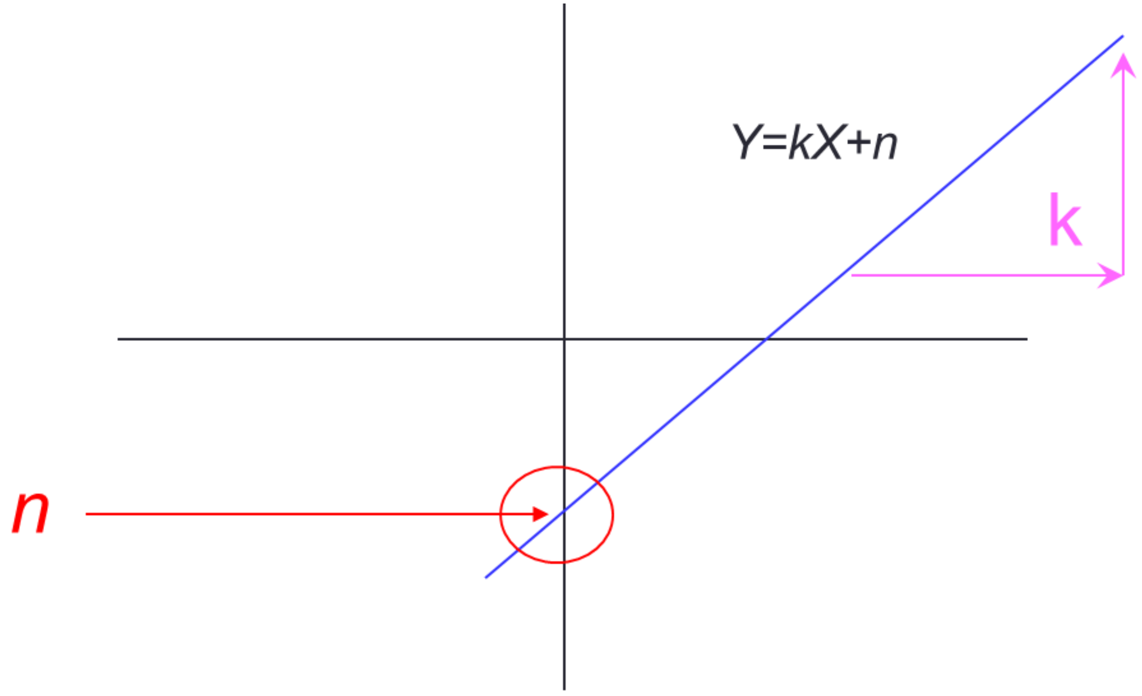
Pretpostavke MNK Linearne regresije
Vizualizacija pretpostavki

- Odnos između X i Y je linearan
- Greške ε su statistički nezavisne
- Greške su normalno distribuirane oko srednje vrednosti 0 za svako dato X
- Distribucija grešaka oko srednje vrednosti 0 ima ima jednaku varijansu σ^2 za svako dato X.
- Odnosno: $\varepsilon \sim N(0,\sigma^2)$



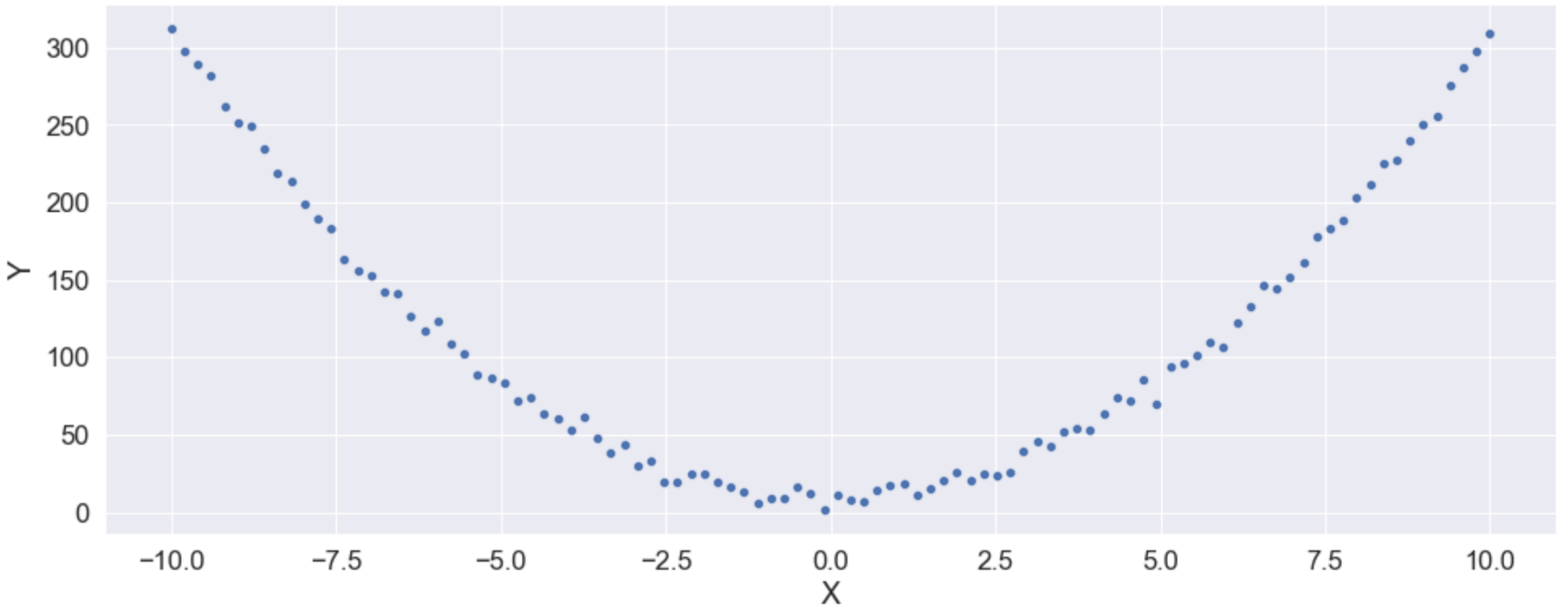
Pretpostavke MNK Linearne regresije Linearnost – tumačenje

- Ova pretpostavka naglašava da ako odnos između X i Y nije linearan da onda koristimo pogrešan model.



Linearnost u smislu parametara

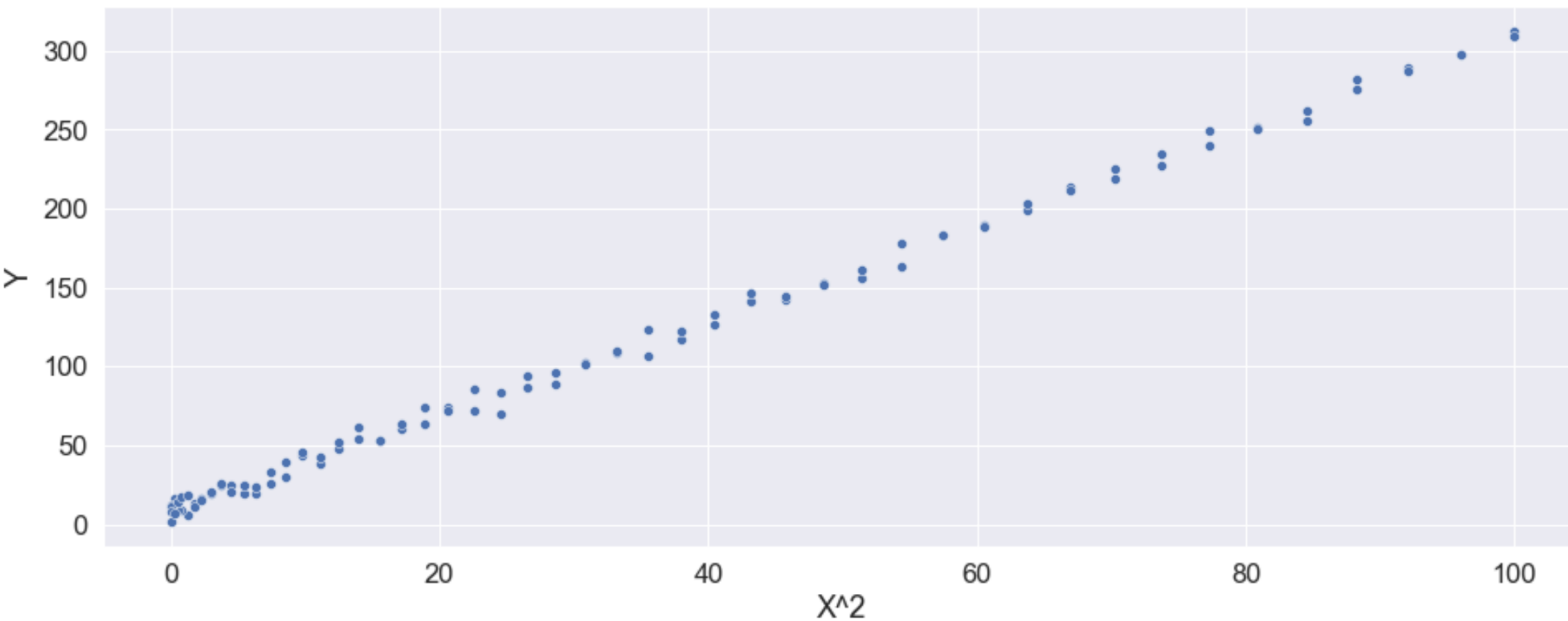
- Linearna regresija je alat koji nam omogućava da modelujemo odnos između X i Y koji nije linearan.
- Da bi to postigli moramo da pronađemo transformaciju X koja će imati linearan odnos sa Y.
- Recimo da su X i Y imaju povezanost kao na sledećem grafiku:



Linearnost u smislu parametara

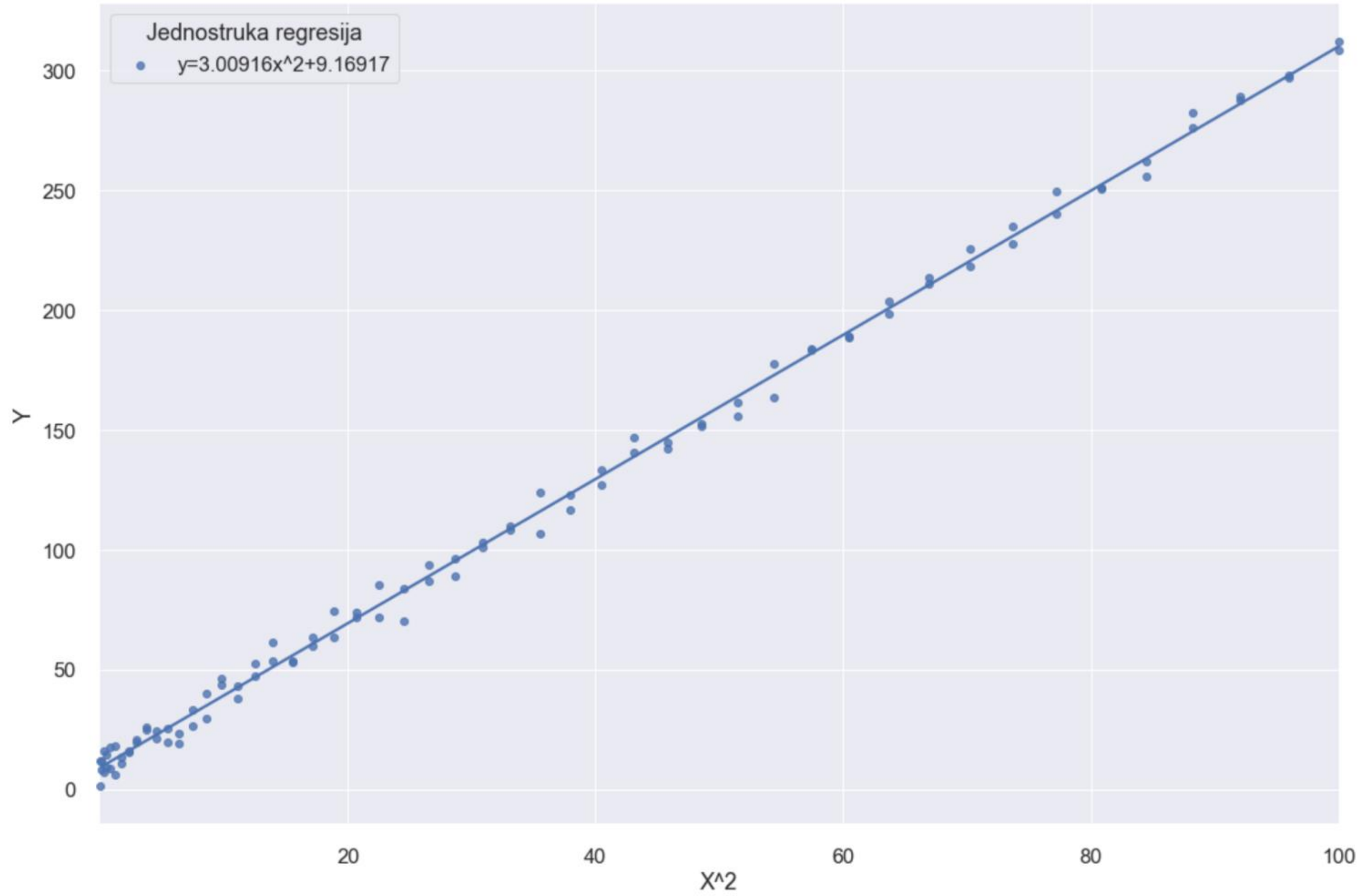
- Sa prethodnog grafika se vidi da je veza između X i Y parabola i da ne možemo da je modelujemo pomoću prave linije.

- Međutim ako umesto X kao nezavisnu promenjljivu koristmo X^2 tada dobijamo sledeći grafik:



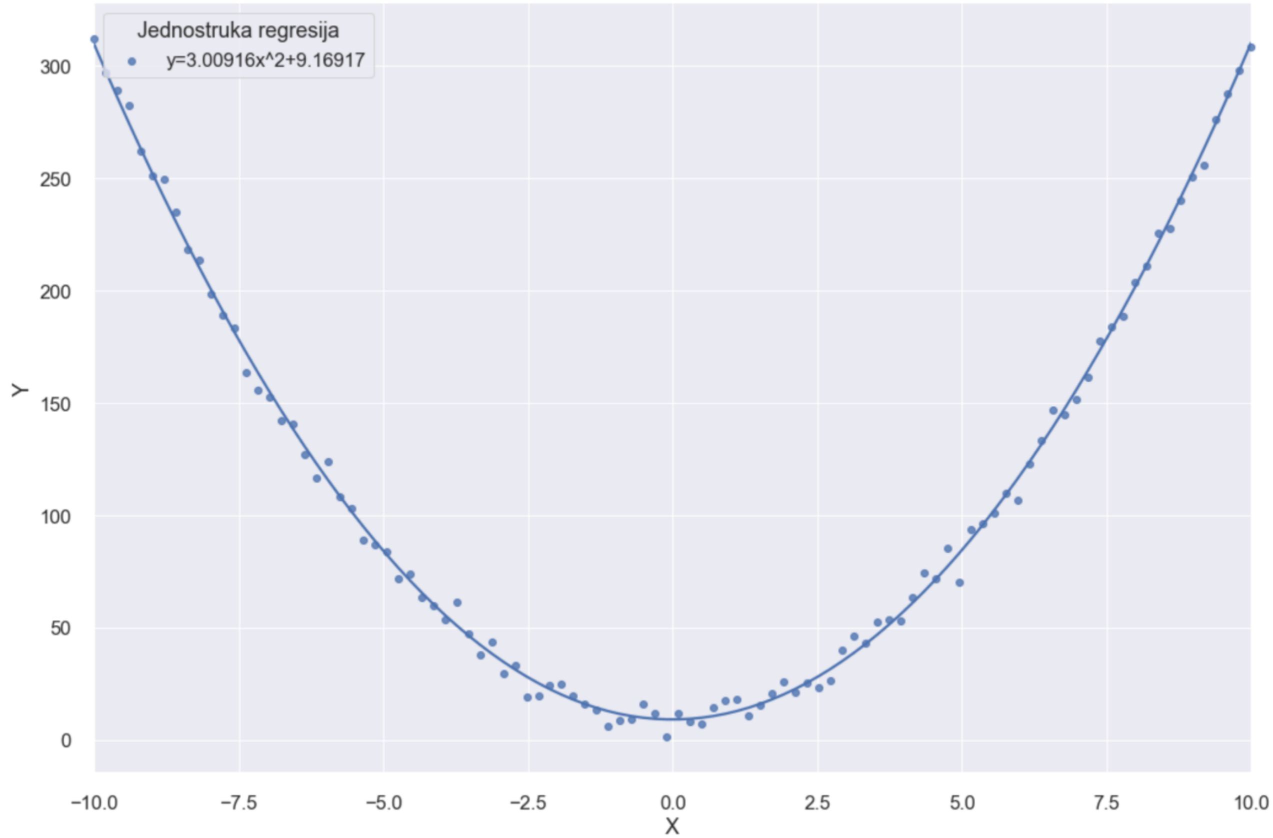
Linearnost u smislu parametara

- Sa grafika se vidi da postoji linearna veza između nezavisne promenjljive X^2 i Y.
- To znači da možemo da koristimo linearnu regresiju, pa imamo:



Linearnost u smislu parametara

- Prikazujemo prethodno dobijeni model $3*X^2 + 9.1$, ali tako da je na x-osi nezavisna promenjljiva X, a ne X^2.



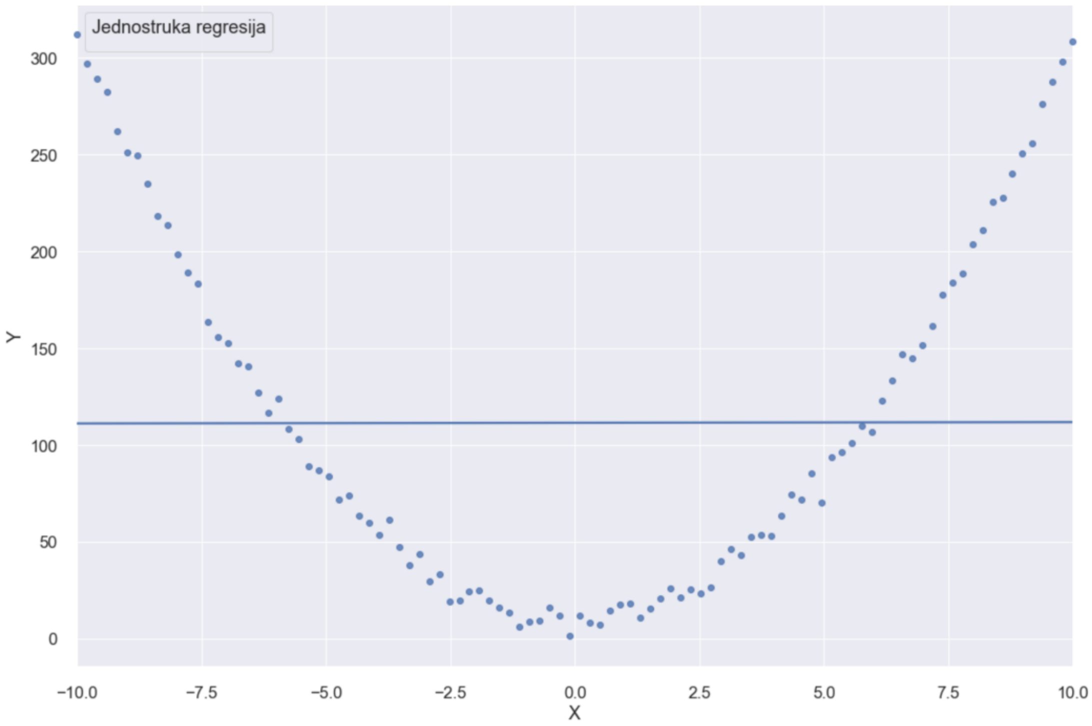
Linearnost u smislu parametara

- Sa grafika se vidi da smo uspeli da upotrebimo linearnu regesiju da modelujemo vezu između X i Y.
- Model koji smo koristili je $3*X^2$+ 9.1 odnosno linearna kombinacija nagiba vrednosti 3, odsečka vrednosti 9.1 i nezavisne promenljive X^2.
- Dakle linearnost se odnosi na vezu između parametara i nezavisne promenjljive, a nezavisna promenjivlja može biti i transformacija promenjljive X iz podataka.
- Takve transformacije se najčeće nazivaju osobine ili na engleskom *features*.
- Dok je broj nezavisnih promenjljivih manji od 3 sa grafika je relativno lako uvideti da li postoje odgovarajuće transformacije.
- U suprotnom, potrebno je analizirati podatke i eksperimentisati sa različitim transformacijama.

Linearnost – testiranje

- U slučaju jednostruke linearne regresije dovoljno je da pogledamo grafik rasipanja.

- Ukoliko na grafiku postoje očigledna velika odstupanja podataka od regresione prave, očigledno je da je linearan model pogrešan izbor.

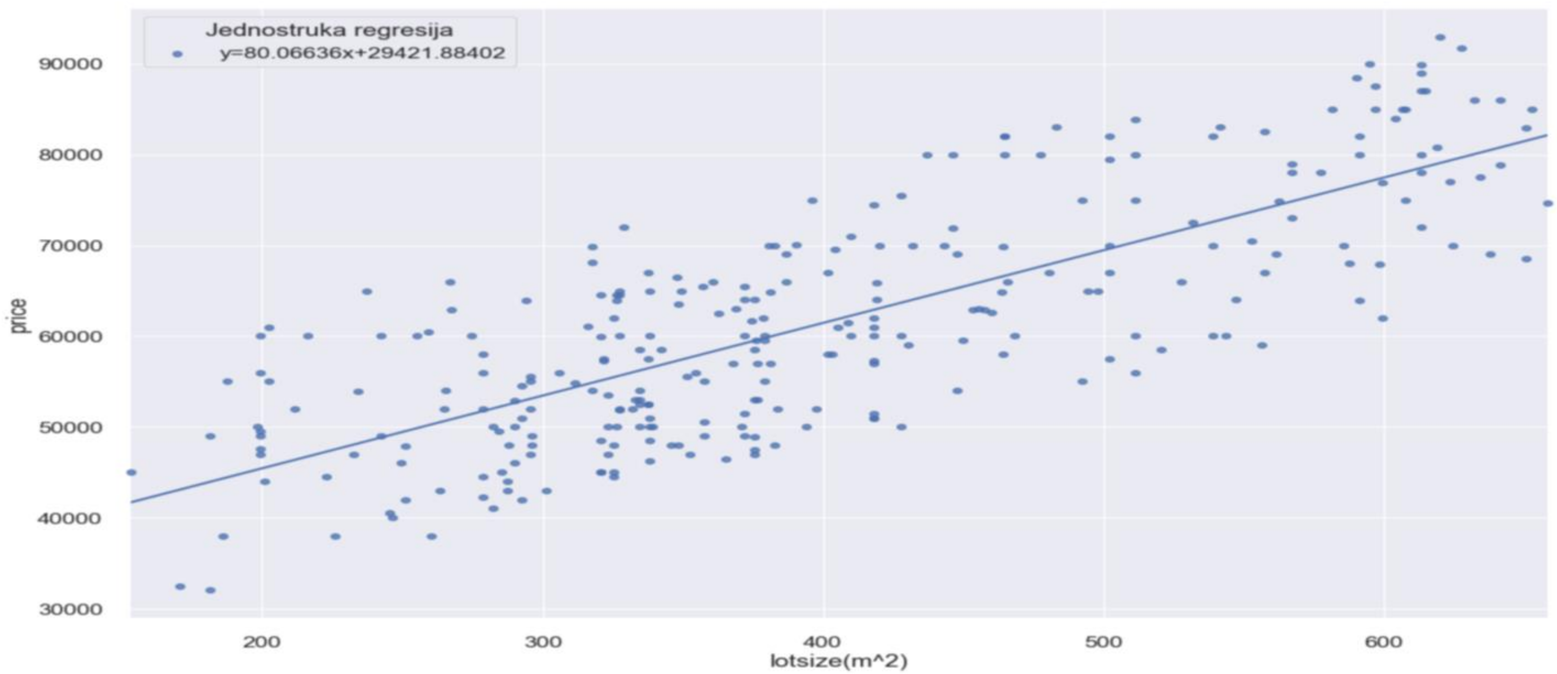


Linearnost – testiranje

- U slučaju jednostruke linearne regresije dovoljno je da pogledamo grafik rasipanja sa regresionom pravom.

- Ukoliko na grafiku postoje očigledna velika odstupanja podataka od regresione prave, očigledno je da je linearan model pogrešan izbor.

Primer: cene kuća



Linearnost – narušenost pretpostavke i potencijalna rešenja

- Linearnost je osnovna pretpostavka linearne regresije.

- Ako je ova pretpostavka narušena model koji smo formirali nije validan ni za predikciju niti za regresionu analizu.

- U nekim slučajevima možemo da uvidimo da postoji odgovarajuća transformacija X koja nam može omogućiti upotrebu linearne regresije.
  - Kao u primeru sa parabolom na prethodnim slajdovima.

- U drugim slučajevima prikladnije je koristiti neki od nelinearnih modela npr. trenutno vrlo aktuelne neuronske mreže.

Nezavisnost grešaka ε_i - tumačenje

- Prisustvo grešaka modela je normalna pojava. Podaci su iz realnog sveta i retko imaju savršenu linearnu (ili neku) drugu vezu.

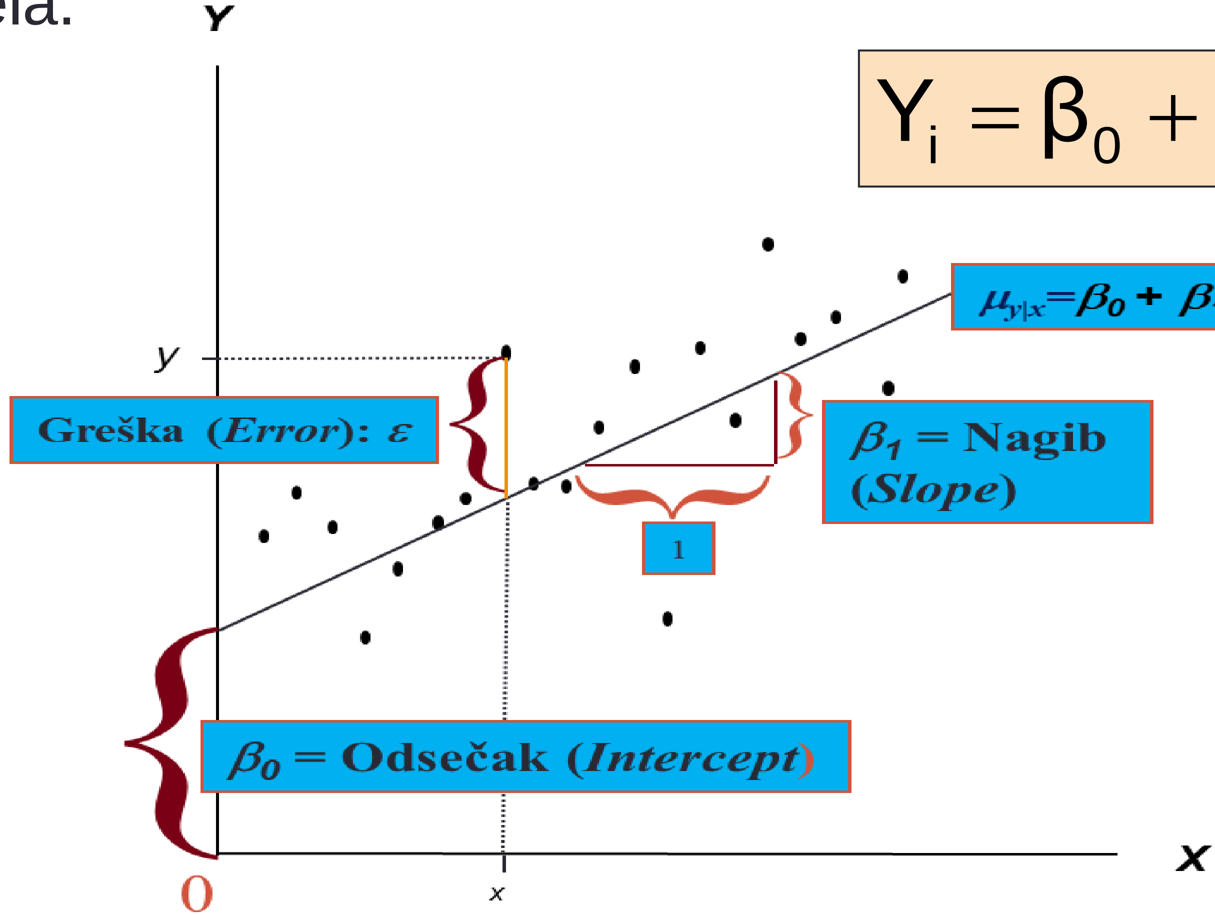
- Pretpostavka kaže da ako znamo nešto o nekoj grešci ε_i to nam ne govori ništa o bilo kojoj drugoj grešci ε_j.

- Ova pretpostavka je vezana za MNK estimator ali takođe može da nam otkrije da li su vrednosti Y međusobno nezavisne što je veoma značajano za ispravan model.
  - Međusobna povezanost Y vrednosti najčešće se javlja kod vrednosti koje se mere kroz vreme (vremenskih serija).

Nezavisnost grešaka ε_i - tumačenje

- Greške ε_i su odstupanja Y od regresione prave koja su deo modela.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Nezavisnost grešaka ε_i - tumačenje

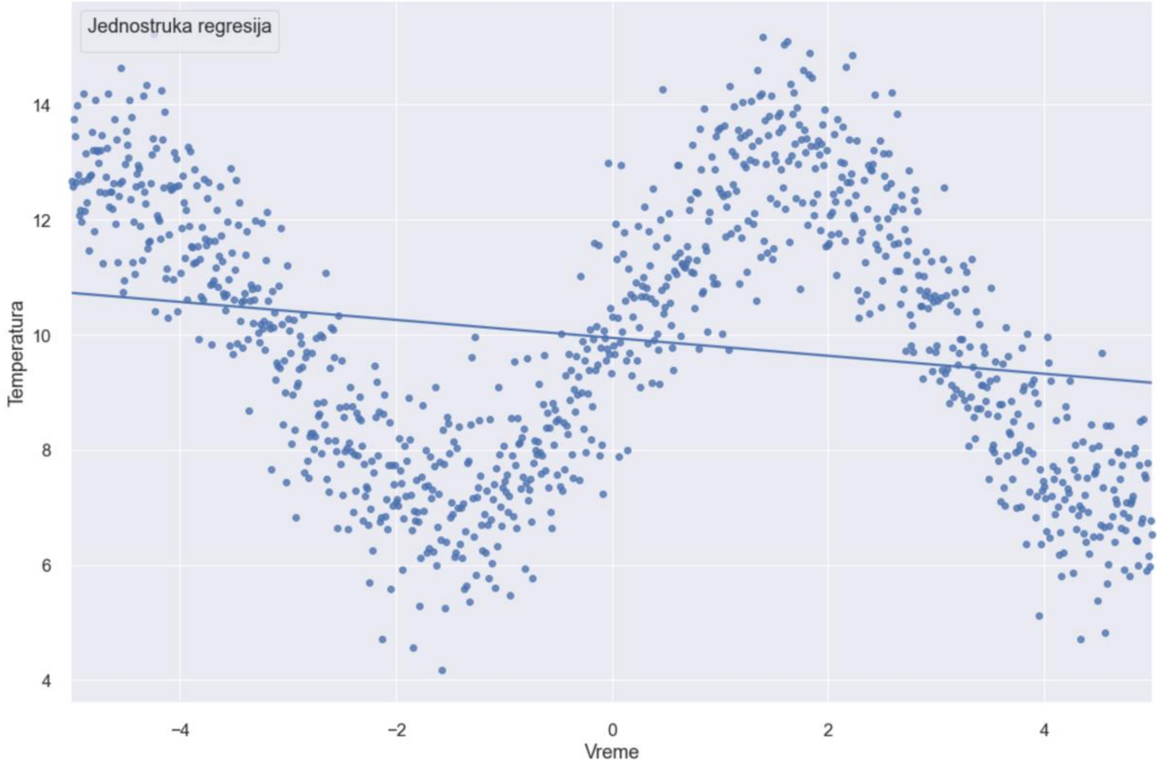
- Sa grafika sa prethodnog slajda se lako može uvideti da ako greške ε_i i ε_j nisu nezavisne onda nisu nezavisni ni podaci Y_i i Y_j.

- Intuitivno, ako pomoću jedne greške (ili Y) možemo da predvidimo drugu (ili drugo Y), onda bi tu informaciju trebalo da koristmo u modelu.
  - Odnosno nezavisna promenjljiva X očigledno nije dovoljna.

- Posmatrajmo grafik na sledećem slajdu.

Nezavisnost grešaka ε_i - tumačenje

- Grafik predstavlja ilustrativni primer promene temperature vazduha po vremenu.
  - Na grafiku je i prava dobijena linearnom regresijom sa nezavisnom promenjljivom Vreme.
  - Sa grafika se vidi da prava nije dobar model trenda u podacima.
- Očigledno da trenutna temperatura zavisi od prethodne.
- U tom smislu, informacije o prethodnim temperaturama bi trebalo da uključimo u model.



Nezavisnost grešaka ε_i - testiranje

- Ovu pretpostavku testiramo pomoću *Durbin-Watson* statističkog testa i analiziranjem grafika reziduala.

- Reziduali e_i su procene grešaka ε_i, koje dobijamo iz uzorka podataka.

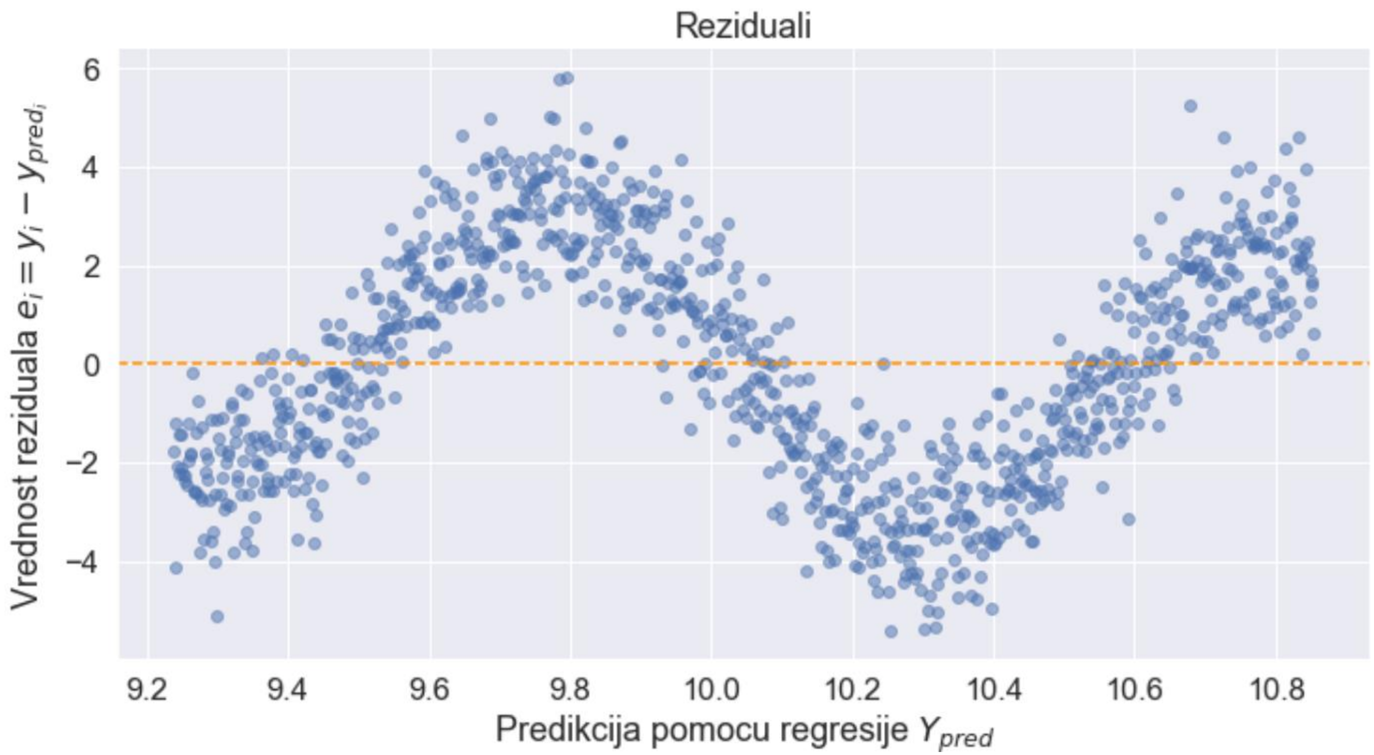
$$e_i = y_i - \hat{y}_i$$

- gde je $\hat{y}_i$ predikcija regresionog modela za x_i.

- Veza reziduala e_i i grešaka ε_i je ista kao i veza između β_0 i β_1 sa b_0 i b_1.
  - β_0, β_1 i ε su teorijski koncepti, tj. parametri i slučajna greška ε populacionog modela, dok su b_0, b_1 i e procenjene vrednosti parametara i slučajne greške dobijene iz uzorka podataka.

Nezavisnost grešaka ε_i – testiranje – grafik reziduala

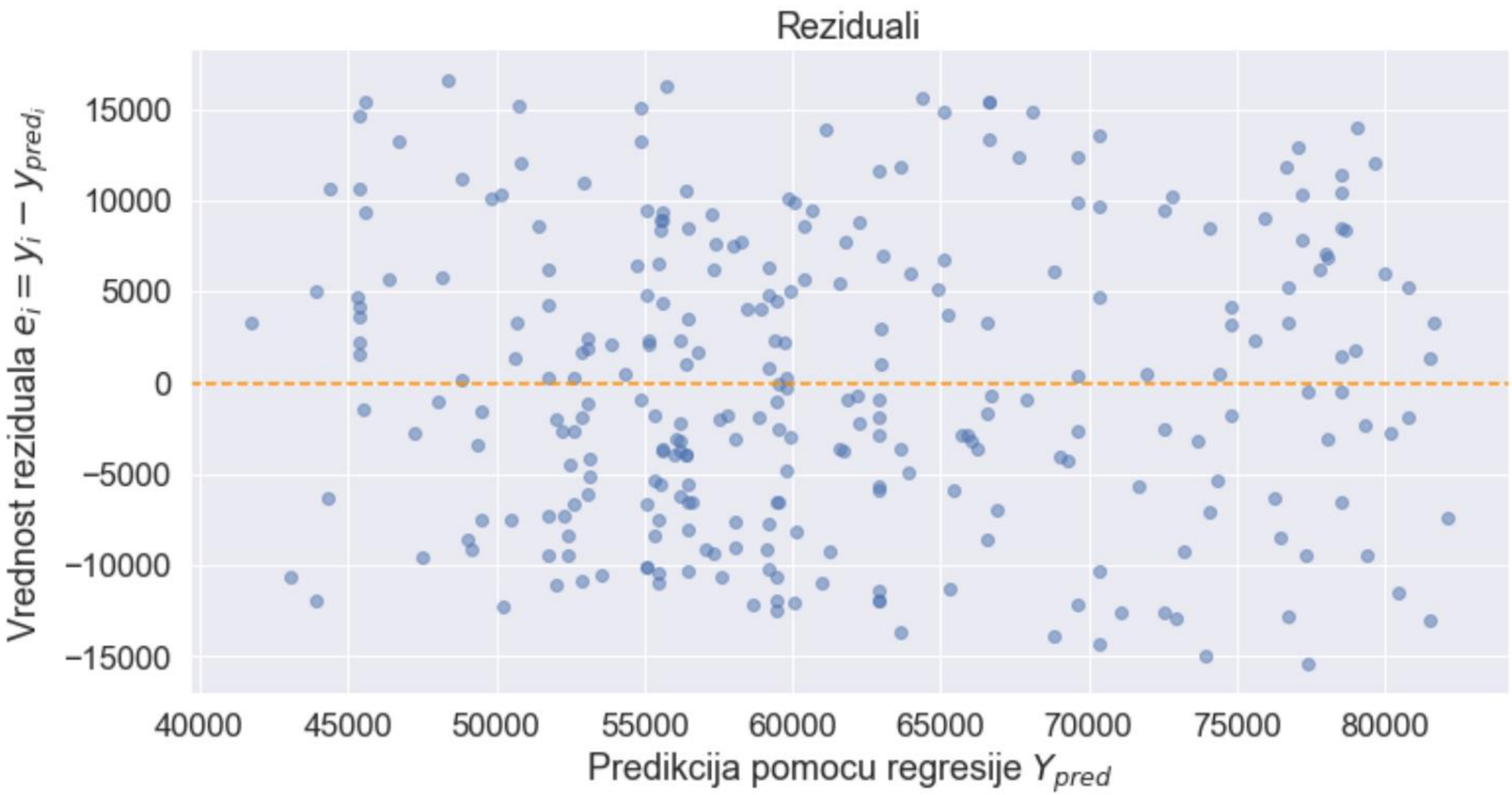
- Ako pretpostavka o nezavisnosti grešaka ne važi za dati uzorak podataka, onda grafik reziduala ima šablon (patern) iz koga se obično može uočiti zavisnost trenutnih vrednosti od prethodnih kao na grafiku ispod.



- Sa grafika se može videti da reziduali imaju šablon, odnosno da se (u ovom slučaju) sledeći rezidual može predvideti pomoću prethodnog.

Nezavisnost grešaka ε_i – testiranje – grafik reziduala

- Prikazujemo sada grafik reziduala za naš primer predikcije cene kuća.



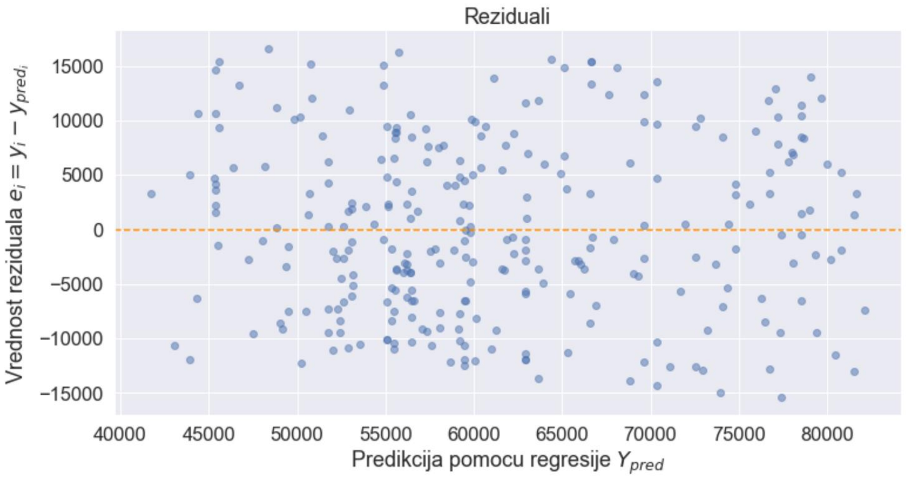
- Sa grafika se može videti da reziduali ne prate nikakav šablon, odnosno da su međusobno nezavisni.

Nezavisnost grešaka ε_i – testiranje – *Durbin-Watson*

- Statistički test koji meri autokorelaciju: korelaciju između vremenske serije (signala) i te iste vremenske serije pomerene za zadati broj vremenskih jedinica unazad.
- *Durbin-Watson* meri autokorelaciju za jednu vremensku jedinicu unazad.
- Za potrebe ovog kursa nećemo detaljno objašnjavati *Durbin-Watson* statistički test već ćemo samo dati interpretaciju rezultata testa.
- U slučaju linearne regresije primenjuje se na rezidualima.
- Raspon vrednosti testa je [0, 4]:
  - Vrednosti u rasponu [1.5, 2] – nema autokorelacije, odnosno greške su nezavisne.
  - Vrednosti u rasponu [0, 1.5) – pozitivna autokorelacija
  - Vrednosti u rasponu (2, 4] – negativna autokorelacija
- Dakle pretpostavka je zadovoljena ako je rezultat u rasponu [1.5, 2].

Nezavisnost grešaka ε_i – testiranje – *Durbin-Watson*

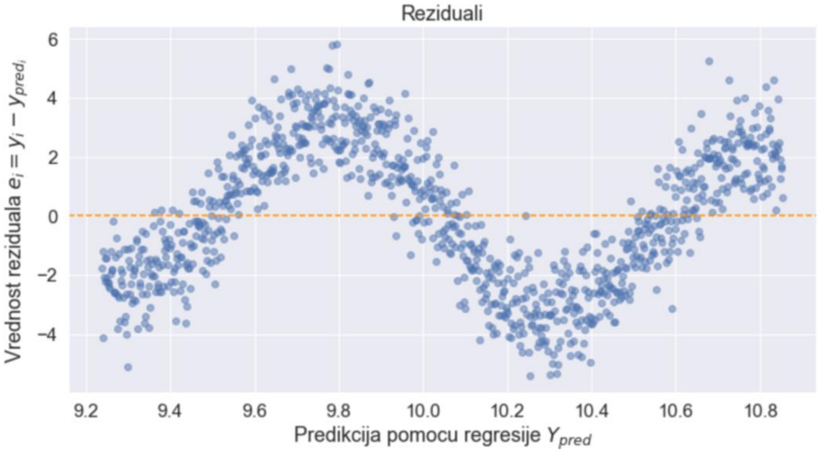
Primer: cene kuća



Durbin-Watson: 1.78
Nema autokorelacije
Pretpostavka važi.

(rezultat u rasponu [1.5, 2])

Primer: temperatura vazduha po vremenu



Durbin-Watson: 0.39
Pozitivna autokorelacija
Pretpostavka ne važi.

Nezavisnost grešaka ε_i – narušenost pretpostavke i MNK

- U slučaju da greške nisu nezavisne MNK nije prikladan estimator za linearnu regresiju.

- Na ovom kursu nećemo se fokusirati na dokaz ove činjenice niti na alternativne estimatore.

- Zainteresovani od vas mogu pročitati više o teoremi Gaus-Markova i npr. o Uopštenom Metodu Najmanjih Kvadrata (Weighted Least Squares).
  - Pored toga postoje tehnike kao što su Robusna linearna regesija ili Regresija pomoću kvantila.

Nezavisnost grešaka ε_i – narušenost pretpostavke i potencijalna rešenja

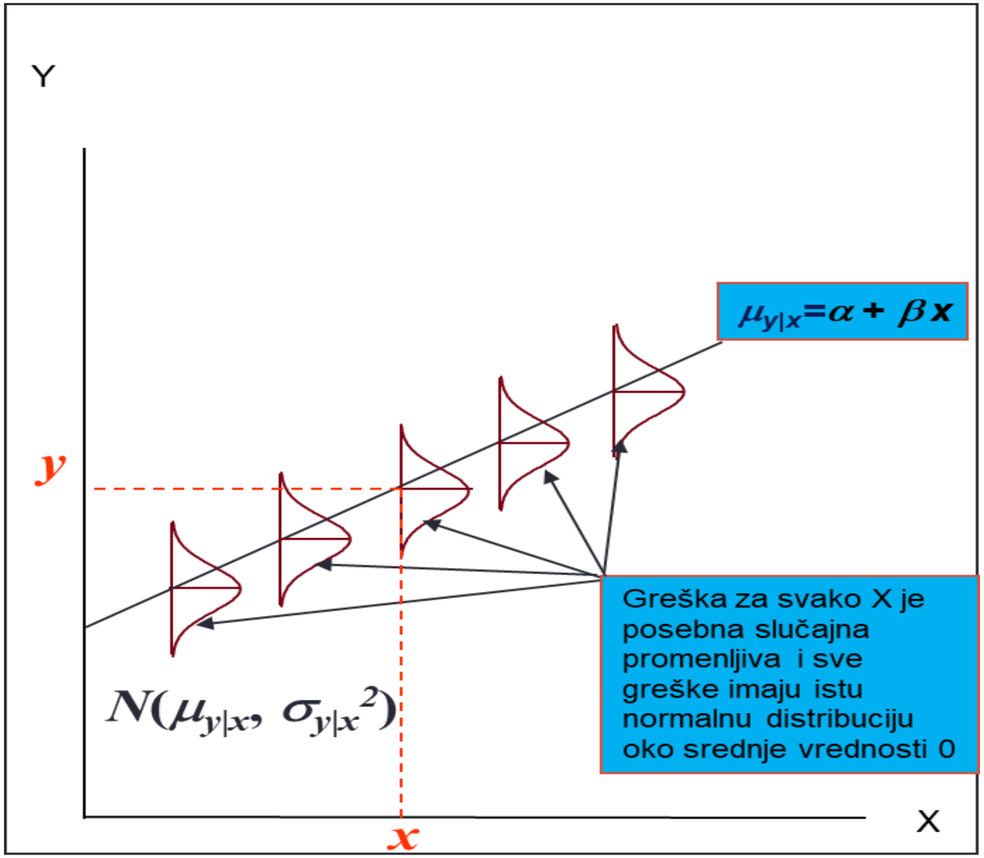
- Narušenost ove pretpostavke čini statističke testove vezane za parametre, kao i intervale poverenja, nevalidnim (odnosno nepouzdanim).
  - Ne možemo zaključiti ništa o tome kako bi izgledao model i kakve bi predikcije bile kada bi promenili uzorak podataka.
  - Ne možemo da pouzdano znamo da li je došlo do preprilagođavanja (overfitting).

- Ako radimo sa podacima koji nisu vremenske serije najčešće rešenje problema je pronalaženje dodatnih nezavisnih promenljivih.
  - Npr. kod primera sa kućama možemo koristi broj soba, broj kupatila itd.

- Kod vremenskih serija postoje tehnike koje ne zahtevaju nove atribute, ali se njima ne bavimo na ovom predavanju.

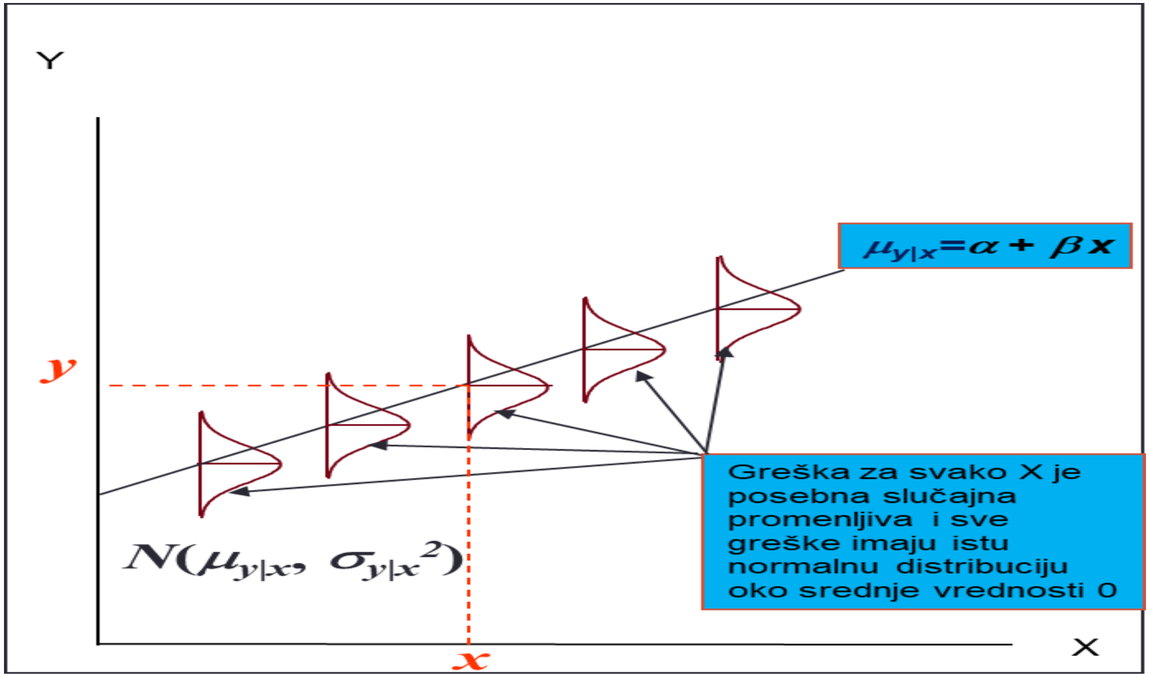
Normalnost grešaka ε_i – tumačenje

- Dve pretpostavke u jednoj: „ (1) Greške ε_i su normalno distribuirane oko (2) srednje vrednosti 0 za svako dato X_i“



Normalnost grešaka ε_i – tumačenje

- Intuitivno, želimo model koji dobro modeluje uzorak podataka koji imamo:
  - Normalna distribucija grešaka znači da imamo puno malih i malo velikih grešaka, koje su podjednako pozitivne i negativne. (1)
  - Srednja vrednost 0 za greške znači da za svako dato X_i regresioni model kao rezultat ima srednju vrednost Y_i u celoj populaciji. (2)



Normalnost grešaka εi – testiranje.

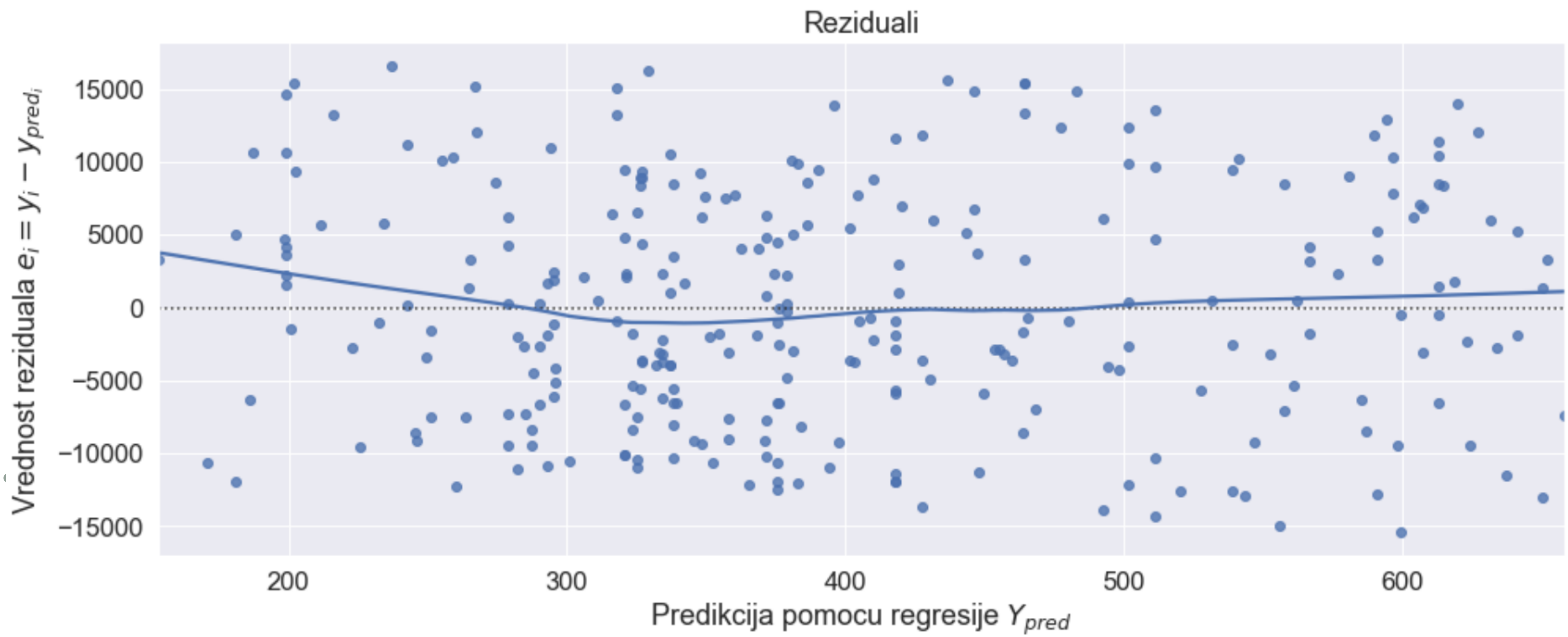
- Objasnićemo prvo testiranje pretpostavke (2).
- Logičan način testiranja bio bi određivanje srednje vrednosti reziduala.
- To nije dobar pristup jer je po definiciji MNK zbir reziduala je uvek 0, pa je samim tim i srednja vrednost reziduala uvek 0.
- Prethodna tvrdnja se za jednostruku regresiju lako može izvesti iz jednog od koraka određivanja odsečka za MNK:

$$\frac{\delta SSE(k,n)}{\delta n} = 2\sum_{i=1}^{n}\underbrace{(y_i - kx_i - n)}_{\text{rezidual za } x_i}(-1) = 0$$

Normalnost grešaka εi – testiranje.

- Pretpostavku (2) testiramo vizualno pomoću grafika rasipanja reziduala na kome imamo liniju* koja nam omogućava da vidimo koliko reziduali odstupaju od nule – želimo „što ravniju“ liniju.

Primer: cene kuća



*Za detalje pogledati o LOWESS (*locally weighted scatterplot smoothing*) liniji.

Normalnost grešaka ε_i – testiranje.

- Pretpostavku (1) – normalna distribucija grešaka – testiramo pomoću: histograma i nekog od statističkih testova (npr. *Shapiro-Wilk* test).

- Histogram je grafik koji prikazuje broj primera (tačaka ili observacija) u podacima koji imaju vrednost koja je u odgovarajućem rasponu.

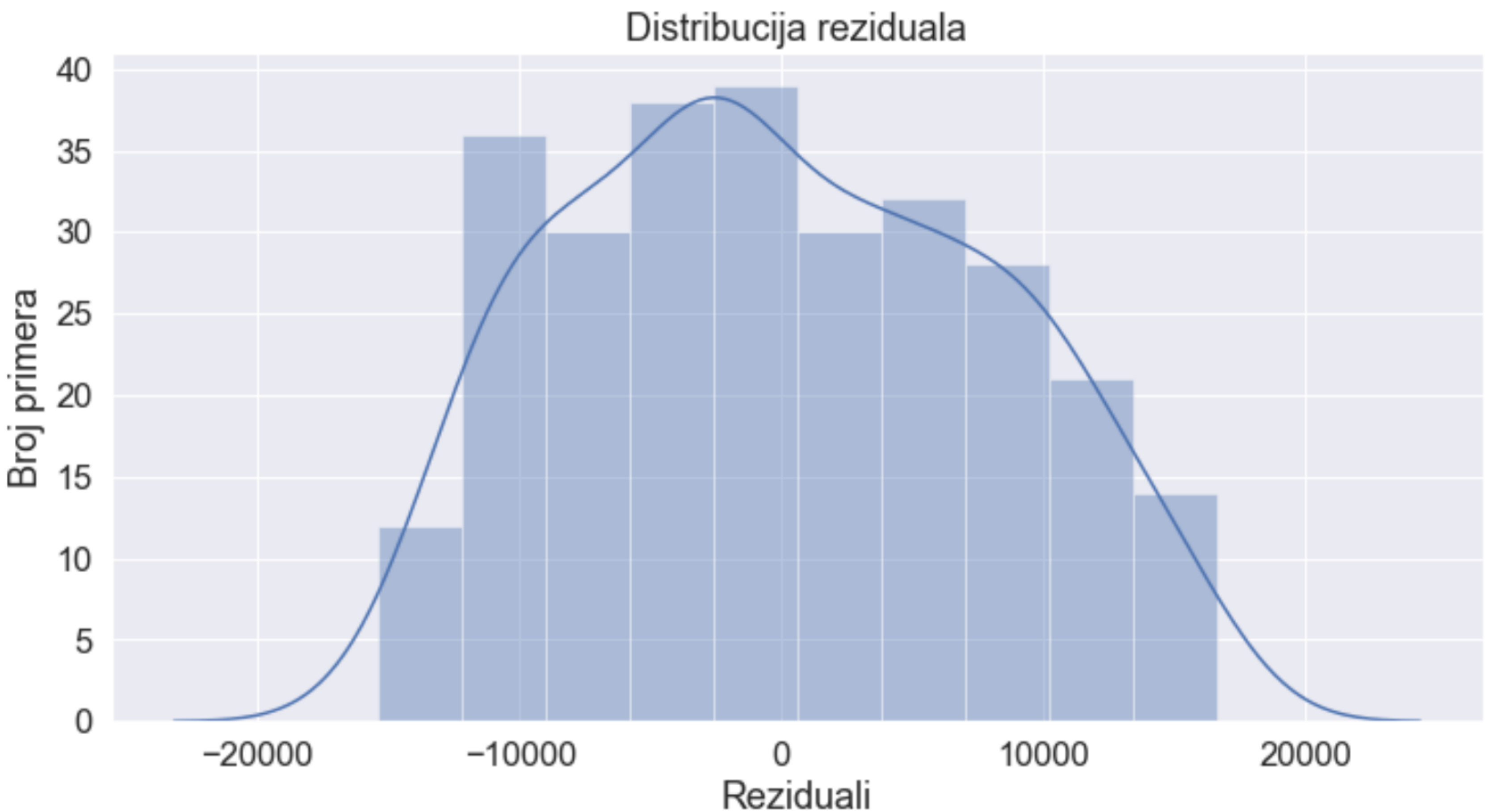
- Rasponi se dobijaju podelom raspona svih primera na različite načine: jednaki delovi, jednaki brojevi primera,...

Normalnost grešaka ε_i – testiranje – *Shapiro-Wilk*

- Statistički test pomoću koga za proizvoljne ulazne podatke možemo da proverimo da li su normalno distribuirani.
- U slučaju linearne regresije primenjuje se na rezidualima.
- Za potrebe ovog kursa nećemo detaljno objašnjavati sam test već ćemo samo dati interpretaciju rezultata testa.
- Raspon vrednosti testa je [0, 1], vrednost veća ili jednaka od 0.05 je indikator da podaci prate normalnu distribuciju.
- Dakle pretpostavka je zadovoljena ako je rezultat u rasponu [0.05, 1].

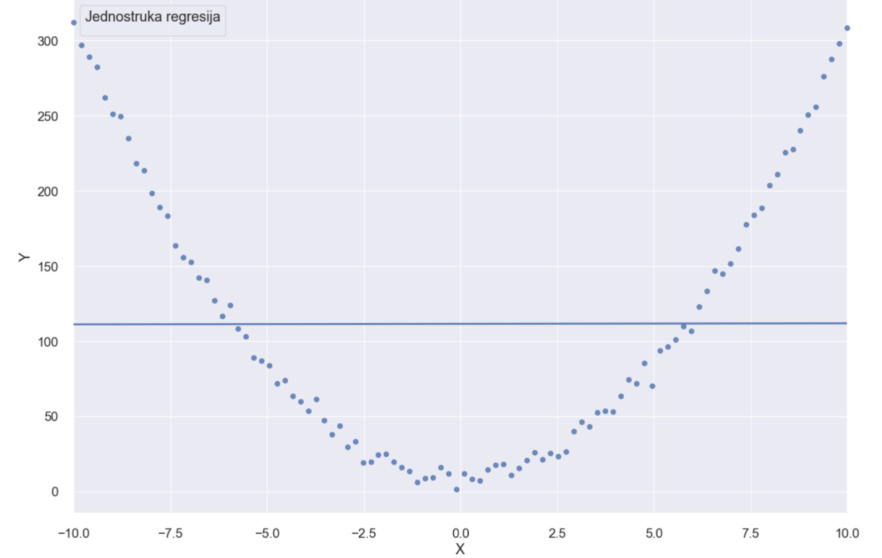
Normalnost grešaka ε_i – testiranje – primeri

Primer: cene kuća



Shapiro-Wilk: $3.82 * 10^{-5}$
Pretpostavka (2) ne važi.

Normalnost grešaka ε_i – narušenost pretpostavke i potencijalna rešenja



- Krenućemo od pretpostavke (2) jer je ona značajnija.

- Ako je pretpostavka (2) narušena, odnosno srednja vrednost grešaka ε_i za dato X_i nije 0, onda linearan model nije prikladan model za dati uzorak podataka.

- U tom slučaju potencijalna rešenja su ista kao i kod narušenosti pretpostavke o linearnosti:
  - Promena modela ili
  - Uvođenje novih nezavisnih promenjljvih ili transofrmacija postojećih

Normalnost grešaka ε_i – narušenost pretpostavke i potencijalna rešenja

- Posledice narušenosti pretpostavke (1) o normalnoj distribuciji grešaka ε_i zavise od veličine uzorka podataka.

- Ako je uzorak podataka mali onda su posledice takve da će zaključivanja o parametrima i predikcijama u slučaju promene uzroka biti nevalidna.
  - Što znači da naš model potencijalno može biti preprilagođen, a mi nemamo način da to utvrdimo.

- Potencijalna rešenja su transformacije zavisne i/ili nezavisne promenljive pomoću logartima ili korena.

Normalnost grešaka ε_i – narušenost pretpostavke i potencijalna rešenja

- Ako je uzorak podataka veliki onda se obično u praksi ova pretpostavka ignoriše.

- Smatra se da su zaključivanja o parametrima i predikcijama u slučaju promene uzroka validna.

- Po relativno savremenoj literaturi (Green 1991)* grubo pravilo za veliki uzorak je:

$$104 + \textit{broj\_zavisnih\_promenljivih}$$

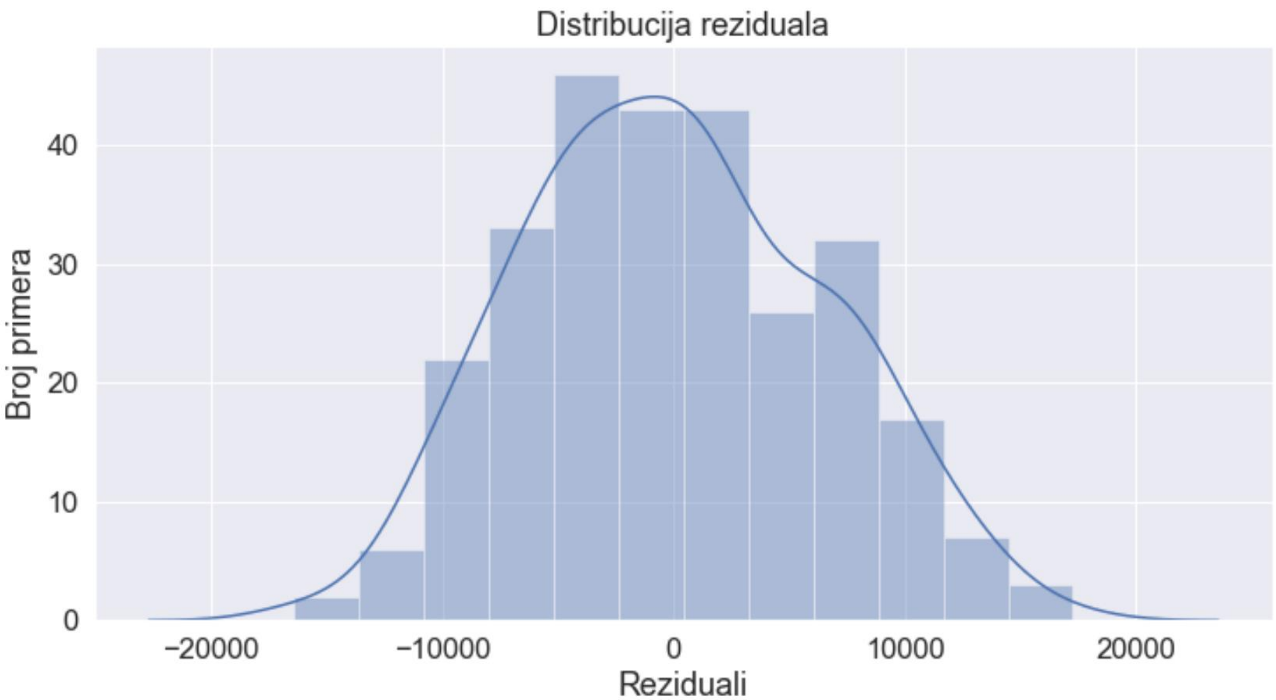
- Kod jednostruke regresije uzorak sa bar 105 primera može se smatrati velikim.

*Green SB. How Many Subjects Does It Take To Do A Regression Analysis. Multivariate Behav Res. 1991 Jul 1;26(3):499-510.

Normalnost grešaka ε_i – uvođenje novih nezavisnih promenljivih – primer cena kuća

- Ako u model uvedemo sve nezavisne promenljive iz skupa podataka dobijamo model višestruke linearne regresije prikazan u tabeli.
- Reziduali takvog modela imaju normalnu distribuciju što se može videti sa grafika, a i iz rezultata Shapiro-Wilk testa.

	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

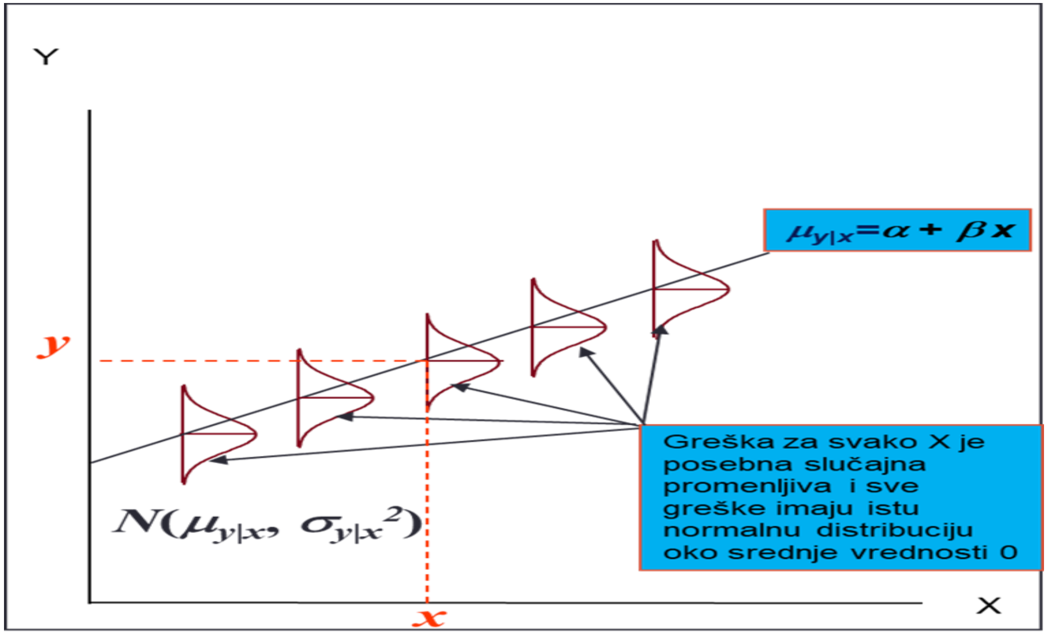


Shapiro-Wilk: 0.124
Pretpostavka (2) važi.

Jednaka varijansa grešaka ε_i – tumačenje

- Pretpostavka: Gausijane grešaka oko regresione prave imaju jednaku varijansu.
  - Vrednost varijanse ne mora da nam bude poznata.

- Intuitivno, želimo da kvalitet našeg modela bude jednak za svako X iz raspona koji imamo u uzroku.



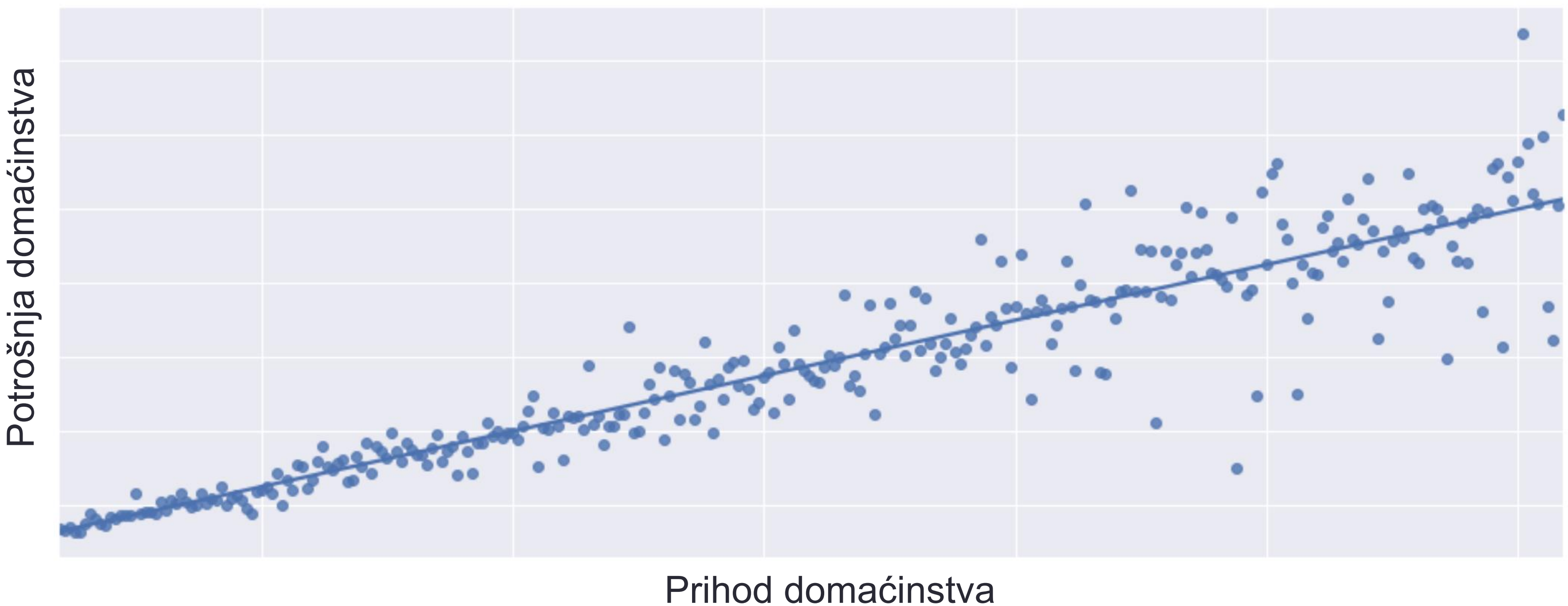
Jednaka varijansa grešaka ε_i – tumačenje

- Recimo da koristimo linearnu regresiju da modelujemo potrošnju nekog domaćinstva u zavisnosti od prihoda:
$$potrošnja = \beta_1 \cdot prihod + \beta_0$$
- Što su prihodi domaćinstva veći to je veća varijabilnost u potrošnji.
  - Veći prihod daje mogućnost izbora u smislu da li će da se troši manje ili više.
- Domaćinstva sa manjim prihodom nemaju mogućnost da mnogo variraju svoju potrošnju.

Jednaka varijansa grešaka ε_i – tumačenje

- Na grafiku ispod dat je ilustrativni primer odnosa prihoda i potrošnje domaćinstva.

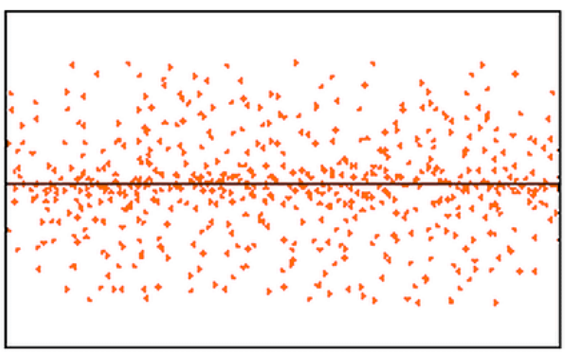
- Sa grafika se vidi da vrednosti potrošnje imaju mnogo veću varijansu za veće prihode.
  - Predikcije regresione prave sa grafika značajno su pouzdanije za manje prihode nego za veće.
  - To je problematično jer nam je cilj da model ima isti kvalitet predikcije u celom rasponu prihoda.
  - Zato je važana pretpostavka o jednakoj varijansi grešaka.



Jednaka varijansa grešaka ε_i – testiranje

- Testiranje se vrši vizualno pomoću grafika raspiranja reziduala.
  - Neravnomerna „rasutost“ reziduala oko 0 može se lako uvideti.
  - Šabloni „levak“ ili „mašna“ su neki od tipičnih primera.

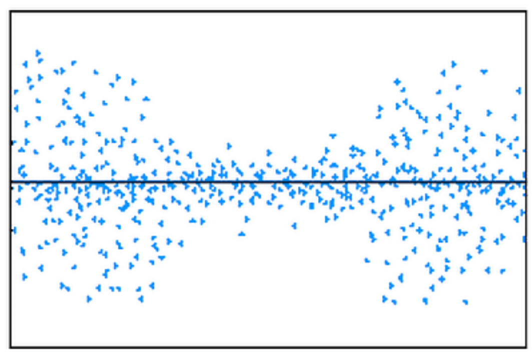
Bez šablona



Random Cloud (No Discernible Pattern)

Pretpostavka
važi

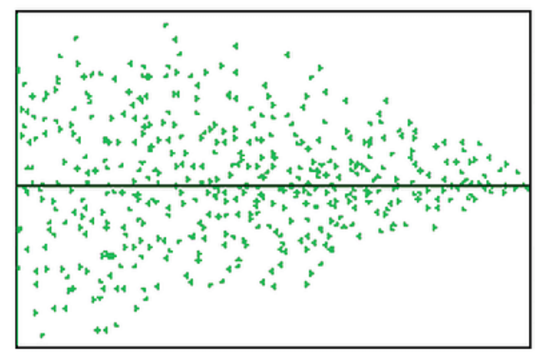
„Mašna“



Bow Tie Shape (Pattern)

Pretpostavka
ne važi

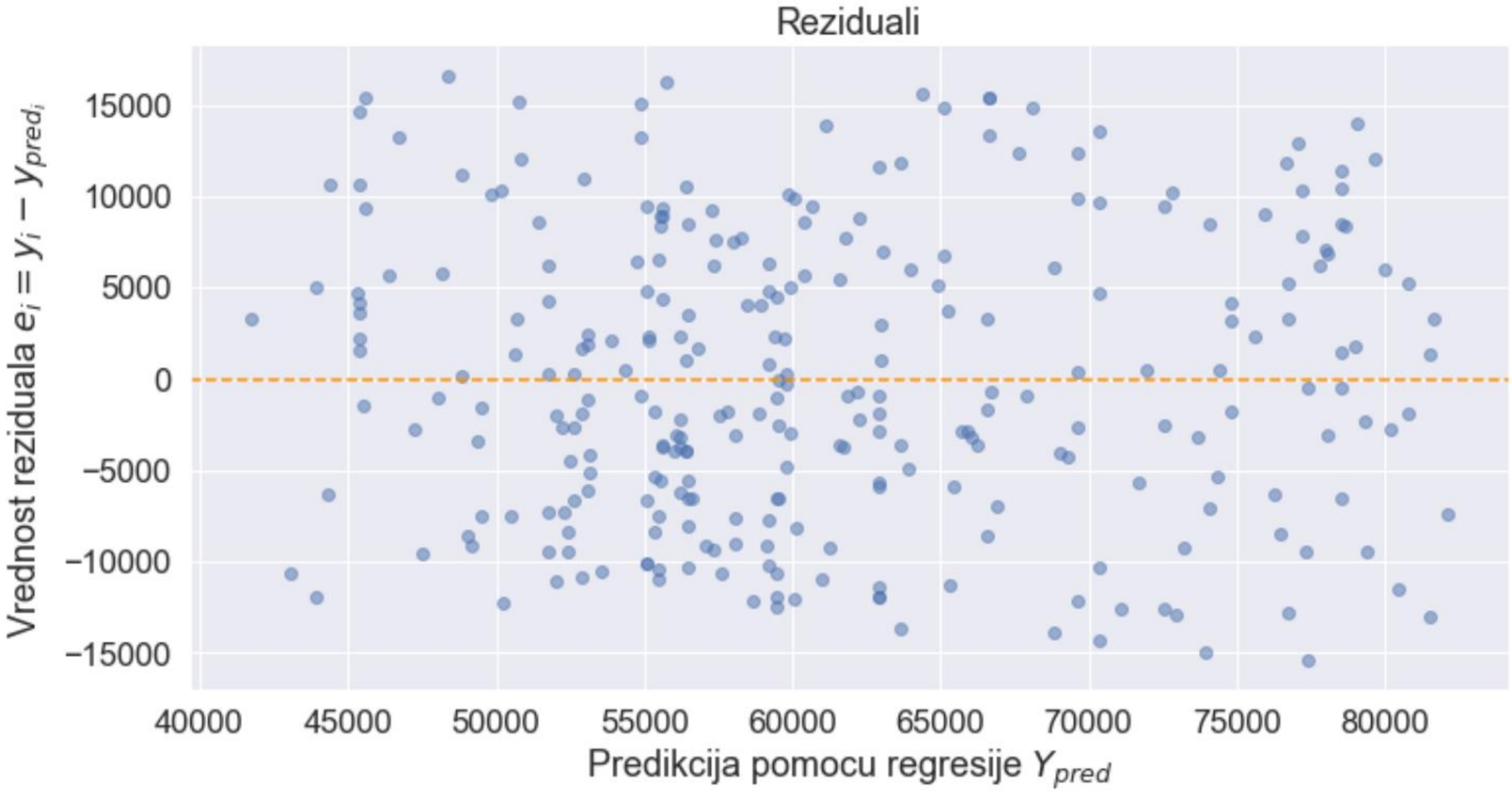
„Levak“



Fan Shape (Pattern)

Pretpostavka
ne važi

Jednaka varijansa grešaka ε_i – testiranje – Primer cene kuća



Na grafiku rasipanja nema šablona – prepostavka o jednakoj varijansi ε_i važi.

Jednaka varijansa grešaka ε_i– narušenost pretpostavke i potencijalna rešenja

- U slučaju da je pretpostavka narušena MNK nije prikladan estimator za linearnu regresiju.
  - Jedna od alternativa je Uopšteni Metod Najmanjih Kvadrata (Weighted Least Squares).

- Pre upotrebe alternativa za MNK vredi pokušati a transformacijom zavisne i/ili nezavisne promenljive pomoću logartima ili korena.

Zaključivanje o modelu linearne regresije

- Nakon testiranja pretpostavki regresije prelazimo na sledeću fazu tumačenja rezultata regresije.
- U ovoj fazi radimo statističko zaključivanje (*inference*) o regresionom modelu.
- Zaključivanje radimo na osnovu uzorka podataka pomoću koga smo odredili parametre modela.
- Proces zaključivanja pruža nam uvid u različite aspekte kvaliteta modela – detaljno objašnjeno u nastavku.

Zaključivanje o modelu linearne regresije – linearna veza između X i Y?

- Pretpostavku o linearnosti proverili smo pomoću grafika rasipanja. Sada je proveravamo pomoću testiranja statističkih hipoteza.

- Kada testiramo statističke hipoteze mi želimo da proverimo da li je neka razlika vrednosti koju smo primetili u uzorku realna?
  - Da li važi za celu populaciju ili samo za naš uzorak?

- Na primer, da li je razlika u broju kardiovaskularnih bolesti ljudi sa i bez visokog krvnog pritiska realna?

Zaključivanje o modelu linearne regresije – t-test

- Kada testiramo statističku hipotezu koristimo statistički test.
- Test meri odnos *signala* i *buke* (*signal to noise ratio*).
- Razliku između vrednosti zvaćemo signal.
- Buka je mera nesigurnosti u vrednosti (iz uzorka) koje testiramo.
- Buka je količnik dve vrednosti:
  1. Mere varijabilnosti vrednosti kada se promeni uzorak.
     - Koliko smo pouzdani u naše procene broja kardiovaskularnih bolesti iz primera sa prethodnog slajda.
  2. Veličine uzorka
     - Veći uzorak „ublažava“ varijabilnost vrednosti koju merimo.
     - Nije isto ako broj kardiovaskularnih bolesti procenjujemo iz populacije od 100 ili 100.000 ljudi.
- Što je buka veća manje smo sigurni i signal mora biti jači.
  - Kao kada pričate sa nekim u svadbi ili u čitaonici.

Zaključivanje o modelu linearne regresije – t-test

- Hipoteza koju testiramo je da li je β_1=0, odnosno da li postoji linearna veza između X i Y?
  - Ako je β_1=0 onda ne postoji linearna veza u modelu:

$$Y = \beta_1 X + \beta_0 + \varepsilon = 0X + \beta_0 + \varepsilon = \beta_0 + \varepsilon$$

- Za test koristimo b_1 koji je procena β_1 iz uzorka.

- Po definiciji testa signal nam je razlika između b_1 i nule.

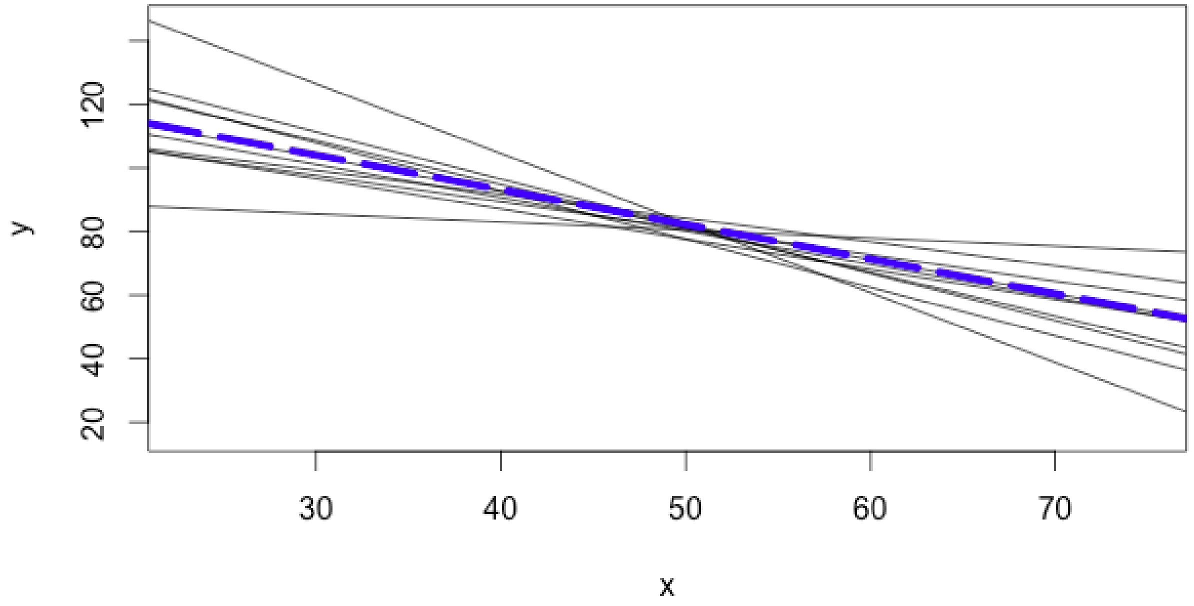
- Buka nam je standardna greška nagiba koja meri koliko bi nagib varirao za promene uzoraka podataka – sledeći slajd.

Zaključivanje o modelu linearne regresije – standardna greška nagiba

- Nagib regresione prave b_1 dobijen pomoću MNK je procena populacionog nagiba β_1 (isprekidana linija na slici)

- Procena b_1 varira u zavisnosti od uzorka podataka.

- Standardna greška nagiba je procena te varijabilnosti.



Zaključivanje o modelu linearne regresije – standardna greška nagiba

- Stadardna greška nagiba (b_1) računa se na sledeći način:

$$\boxed{S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum (X_i - \overline{X})^2}}}$$

$$S_{YX} = \sqrt{\frac{SSE}{n-2}}$$ = Standardna greška predikcija modela

- U ovom slučaju veličina uzorka deo je formule za standardnu grešku, pa na taj način i samog odnosa signal-buka.
  - Što je *n* veće to će S_{b1} biti manje – veći uzorak „ublažava“ varijabilnost tj. smanjuje buku .

Zaključivanje o modelu linearne regresije – T-vrednost

- Nakon definsanja signala i buke definišemo i njihov odnos:

$$T - vrednost = \frac{signal}{buka} = \frac{b_1 - 0}{s_{b_1}} = \frac{b_1}{s_{b_1}}$$

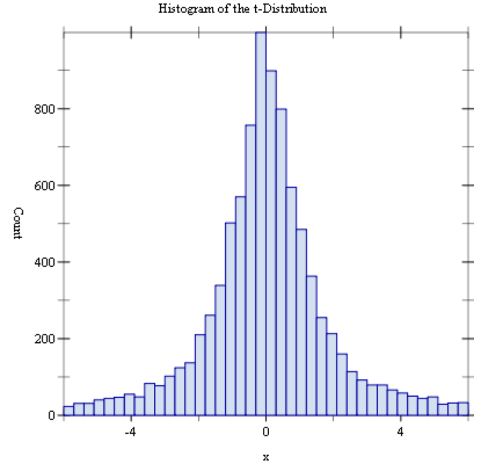
- T-vrednost meri odnos signala i buke.

- Ono što nas dalje zanima je šta nam govori konkretna T-vrednost.

Zaključivanje o modelu linearne regresije – T-vrednost

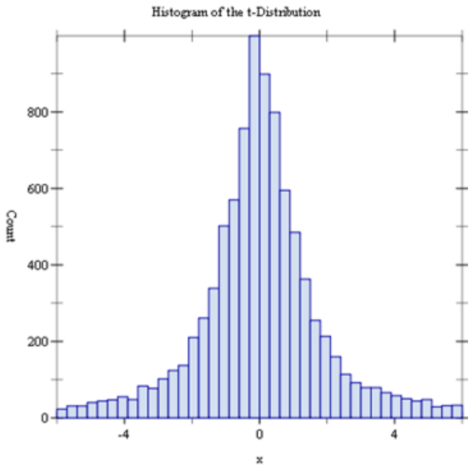
- Bilo bi idealno kada bi mogli da znamo kakve su T-vrednosti kada naša hipoteza važi.
  - Onda bi mogli da uporedimo dobijenu T-vrednost sa tim vrednostima i damo zaključak o važenju hipoteze.

- Jedna mogućnost je da sami pronađemo populaciju primera u kojoj važi pretpostavka (β_1=0), da prikupimo puno uzoraka, odredimo T-vrednosti i kreiramo histogram.



Zaključivanje o modelu linearne regresije – t-distribucija

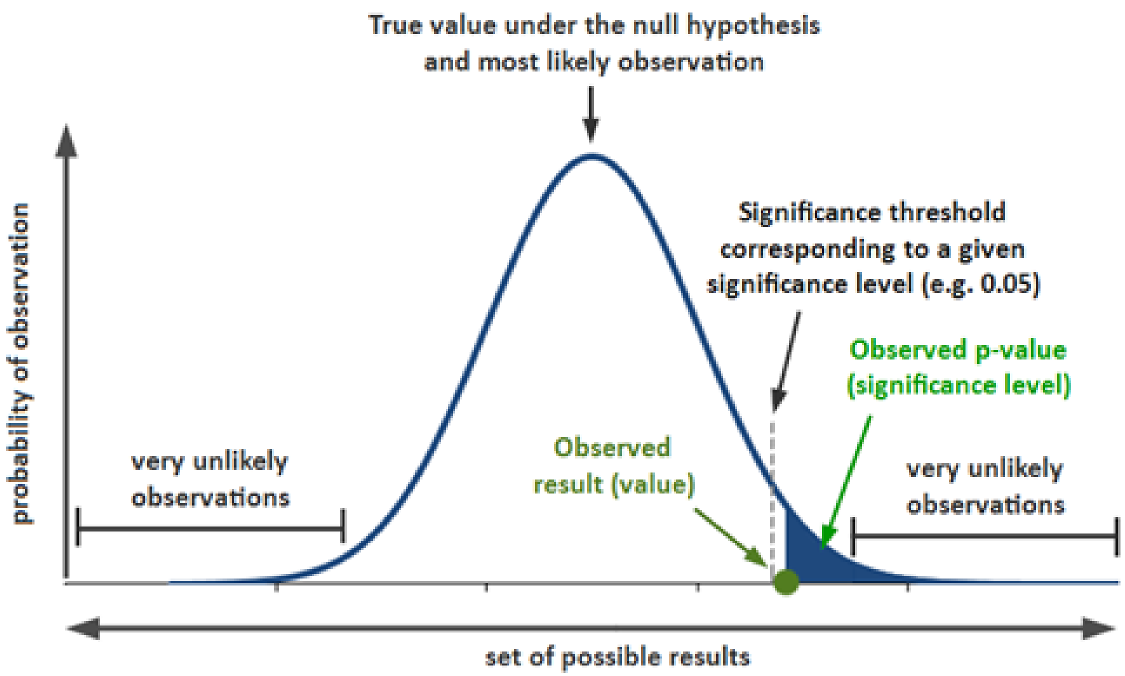
- Pomoću histograma možemo da zaključimo koliko je T-vrednost koju smo dobili česta ako hipoteza važi.
  - Na primer, vrednosti veće od 4 ili manje od -4 su retke, dok su vrednosti oko nule česte.



- Dakle, ako naša T-vrednost nije česta onda bi mogli da zaključimo da hipoteza ne važi.
- Postupak koji smo opisali se i koristi ali bez potrebe da sami formiramo histogram.
- Istraživanjima je utvrđeno kako izgledaju histogrami T-vrednosti u slučaju kada važi hipoteza koja se testira (signal je mali).
- Oblik histograma određen je t-distribucijom – distribucija koju prate T-vrednosti.

Zaključivanje o modelu linearne regresije – P-vrednost

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



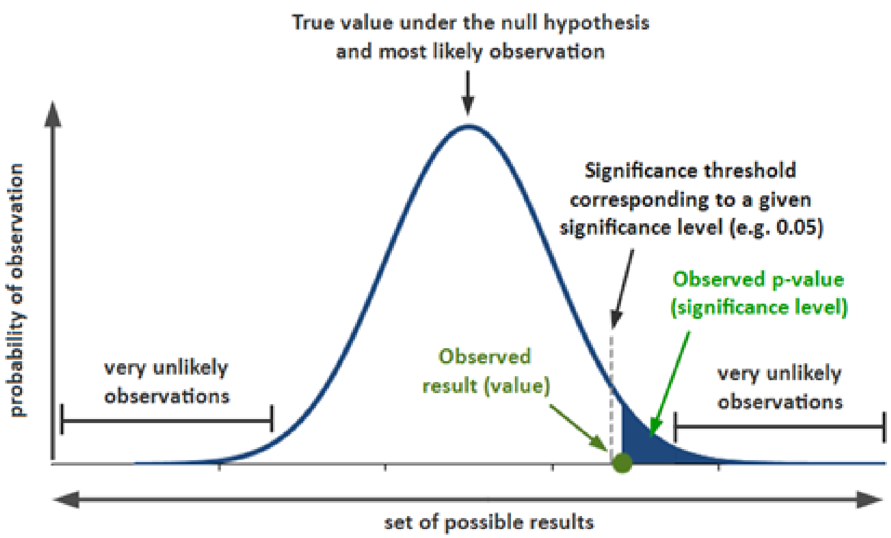
Zaključivanje o modelu linearne regresije – P-vrednost

- Želimo što manju p-vrednost. Prag koji se koristi u praksi je 0.05.
  - Ako je p-vrednost ≤ 0.05 onda zaključujemo da signal postoji, odnosno da je $\beta_1 \neq 0$.

- Veći deo distribucije zauzimaju T-vrednosti koje bi dobili da važi $\beta_1=0$ jer je tako t-distrbucija formirana.

- Ako je p-vrednost ≤ 0.05 to znači da postoji ≤ 5% verovatnoće da ćemo dobiti našu T-vrednost ako važi $\beta_1=0$.

- Odnosno možemo sa 95% sigurnosti da zaključimo da je $\beta_1 \neq 0$.

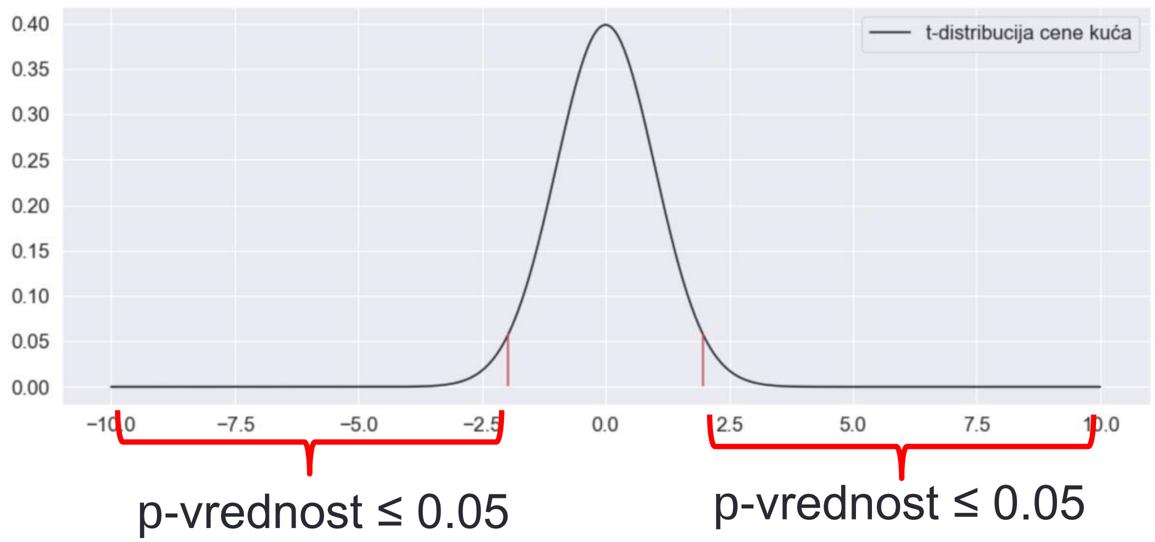


Zaključivanje o modelu linearne regresije – P-vrednost - Primer

- T- i p- vrednosti ne izračunavamo ručno već koristimo softver.
- Za naš primer sa cenama kuća pomoću Python biblioteke *statsmodels* dobijamo:

	coef	std err	t	P>|t|
const	2.942e+04	1608.768	18.288	0.000
lotsize(m^2)	80.0664	3.856	20.764	0.000

- T-vrednost za β_1 je 20.764 dok je p-vrednost nula, tj. sigurno ispod 0.05.



Zaključivanje o modelu linearne regresije – Interval poverenja

- Interval poverenja je raspon vrednosti u kome će se potencijalno naći prava vrednost parametra koji procenjujemo na osnovu uzorka.

- Prava vrednost odnosi se na vrednost koju bismo dobili kada bi nam uzorak bila cela populacija.

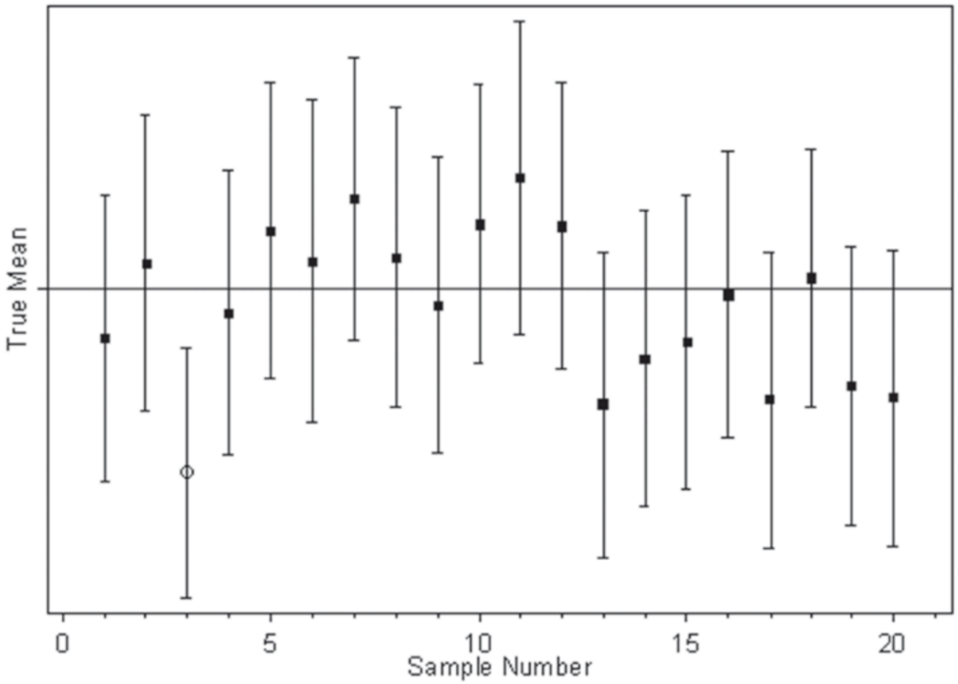
- Prava vrednost nam nije dostupna, zato i radimo procene.

- Intervali poverenja kreiraju se za određeni procenat pouzdanosti.

Zaključivanje o modelu linearne regresije – Interval poverenja - interpretacija

- Interpretacija intervala poverenja za 95% pouzdanosti: ako imamo 100 različitih uzoraka podataka i za svaki kreiramo intreval poverenja sa pouzdanošću od 95%, onda će 95 tih intervala da sadrži pravu vrednost parametra, a 5 neće (slika ispod).

- Znači biranjem veće pouzdanosti mi povećavamo verovatnoću da će baš interval koji smo mi dobili da sadrži pravu vrednost, ali nemamo garanciju da je tako.



Zaključivanje o modelu linearne regresije – Interval poverenja za β_1

- Za linearnu regresiju kreiraju se intervali poverenja za parametre i predikcije.

- Tipično se bira nivo pouzanosti od 95%.

- U nastavku prvo ćemo pokazati na koji način se kreira intreval poverenja za parametar β_1.

Zaključivanje o modelu linearne regresije – Interval poverenja za β_1 - Primer

- Interval poverenja za naš primer sa cenama kuća pomoću:

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.942e+04	1608.768	18.288	0.000	2.63e+04	3.26e+04
lotsize(m^2)	80.0664	3.856	20.764	0.000	72.476	87.657

- Sa pouzdanošću od 95% zaključujemo da je β_1 između 72.4 i 87.6.

- Na osnovu našeg uzorka možemo da sa pouzdanošću od 95% da tvrdimo da se cena kuće u populaciji poveća za iznos između 72.4 i 87.6 dolara kada se površina placa poveća za 1m^2.

Zaključivanje o modelu linearne regresije – Interval predikcije

- Interval predikcije je raspon vrednosti u kome će se potencijalno naći prava vrednost zavisne promenjlijve Y za neko zadato X_d.

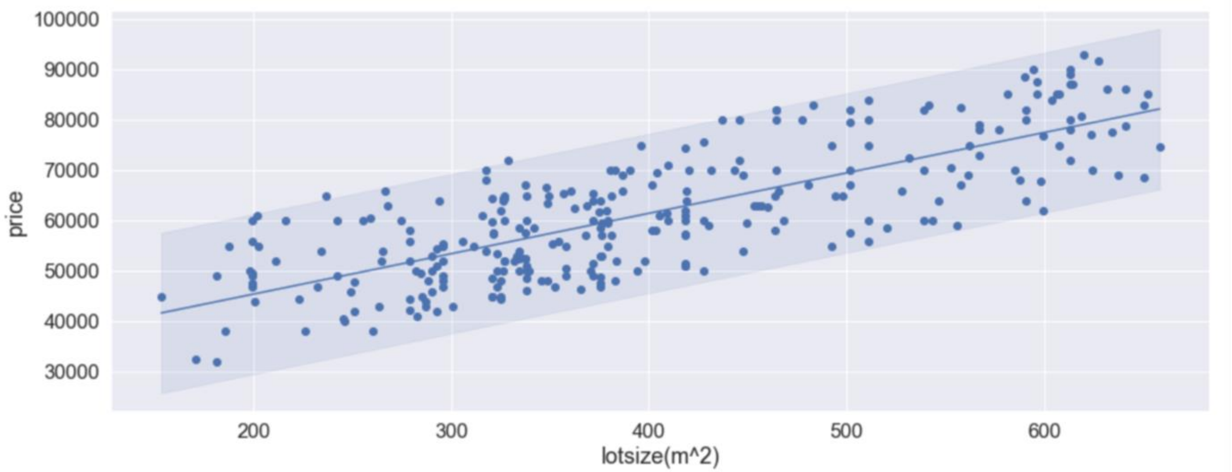
- Pod terminom prava vrednost misli se na vrednost koju bih populacioni model vratio ako bi u njega uneli X_d:

$$Y_d = \beta_1 \cdot X_d + \beta_0 + \epsilon_d$$

- Gde su β_0 i β_1 parametri populacionog modela, X_d neka data tačka, a ε_d je greška populacionog modela za X_d.

Zaključivanje o modelu linearne regresije – Interval predikcije

- Naravno, mi ne znamo parametre populacionog modela, ni slučajnu grešku ε_d.
- Iz tog razloga kreiramo intreval predikcije kako bi sa određenom pouzdanošću procenili gde će biti vrednost Y_d.
- Recimo da nam je dat populacioni model cena kuća u Srbiji (modelovan pomoću svih kuća u Srbiji) i da znamo vrednosti svih slučajnih grešaka.
- Onda bi recimo Y_d bila cena kuće koju bi vratio populacioni model za neko X_d na primer od $132m^2$.



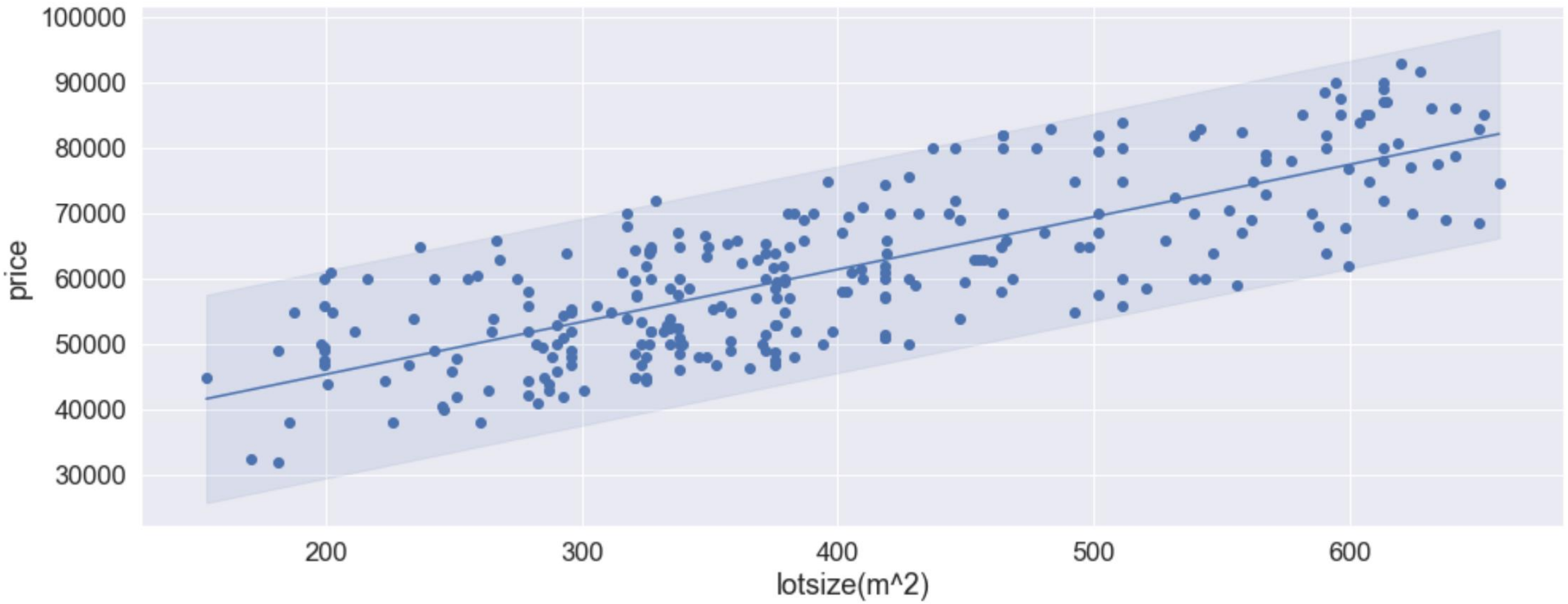
Zaključivanje o modelu linearne regresije – Interval predikcije - interpretacija

- Kao i kod intervala poverenja, moramo da odaberemo pouzdanost.
- Tipično se koristi pouzdanost od 95%.
- Interpretacija za pouzdnost od 95%: ako imamo 100 različitih uzoraka i za svaki kreiramo interval predikcije (za 95% pouzdanost) za neko dato X_d onda će 95 tih intervala sadržati pravu vrednost Y_d.

Zaključivanje o modelu linearne regresije – Interval predikcije – Primer cene kuća

- Pomoću Python biblioteke statsmodels izračunali smo intervale predikcije za svaki primer iz skupa podataka cena kuća.

- Sa grafika se vidi da su intervali predikcije ravnomerni za ceo raspon X.



Zaključivanje o modelu linearne regresije – Interval predikcije – Primer cene kuća

- Pomoću *Python* biblioteke *statsmodels* izračunali smo interval predikcije sa pouzdanošću od 95% za kuću sa površinom 658.47m^2:

[66257.84, 98030.10]

- To znači da postoji verovatnoća od 95% da baš ovaj interval sadrži predikciju populacionog modela za x_d = 658.47m^2.

- Interval je dosta širok imajući u vidu cenu u dolarima.

- Širina je posledica relativno velike prosečne greške modela od 7993.22 dolara.

- Jedan od načina da se smanji prosečna greška je upotrebom dodatnih atributa pored površine placa – to je tema sledećeg predavanj.

- Napomena: Povećanje skupa podataka je takođe dobar način da se smanji prosečna greška, ali proces pribavljanja novih podataka tipično nije lak.