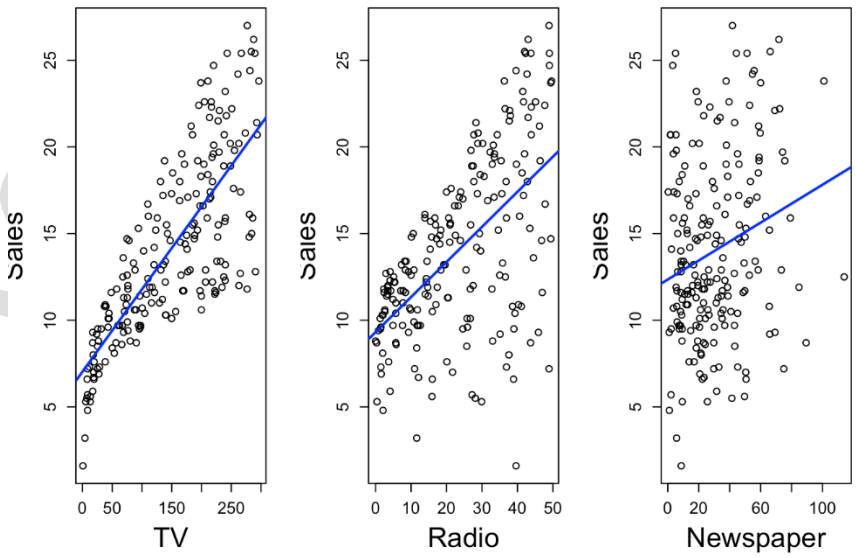
# Statističko učenje

Zamislite da ste zaposleni u ulozi Data Scientist-a u kompaniji koja se bavi prodajom određenog proizvoda na 200 različitih lokacija. U isto vreme kompanija ima određeni budžet za marketing koji troši na različite medije: novine, TV i radio. Mi ne možemo direktno da utičemo na prodaju, ali možemo da raspolažemo, tačnije raspodelimo, budžet na različite medije. Ako utvrdimo da postoji veza između reklamiranja i prodaje, tada možemo da podesimo raspodelu novca i na taj način indirektno utičemo na povećanje prodaje.

Drugim rečima: Naš zadatak je da razvijemo precizan model koji može da nam predvidi prodaju na osnovu raspodele budžeta na tri medija.

R code:

```
lm.radio=lm(Sales ~ Radio)
lm.tv = lm(Sales ~ TV)
lm.newspaper = lm(Sales ~ Newspaper)
par(mfrow = c(1,3))
plot(TV, Sales, cex.lab=2, cex.axis=1.2)
abline(lm.tv, col="blue", lty=1, lwd=2)
plot(Radio, Sales, cex.lab =2, cex.axis=1.2)
abline(lm.radio, col="blue", lty=1, lwd=2)
plot(Newspaper, Sales, cex.lab=2, cex.axis=1.2)
abline(lm.newspaper, col="blue", lty=1, lwd=2)
```




U ovom slučaju *budžeti za medije* su ulazne promenljive (eng. input variables), dok je *prodaja* izlazna promenljiva (eng. *output variable*).

$X_1$ - TV budžet
$X_2$ - Radio budžet
$X_3$ - Novine budžet

U literaturi možete naći i sledeće nazive za ulazne promenljive: *input variables, predictors, independent variables, features*, i najčešće se obeležavaju sa $X_{index}$.

Za izlazne promenljive možete čuti nazive: *output variables*, *response, dependent variable*, i najčešće se obeležavaju sa $Y_{index}$.

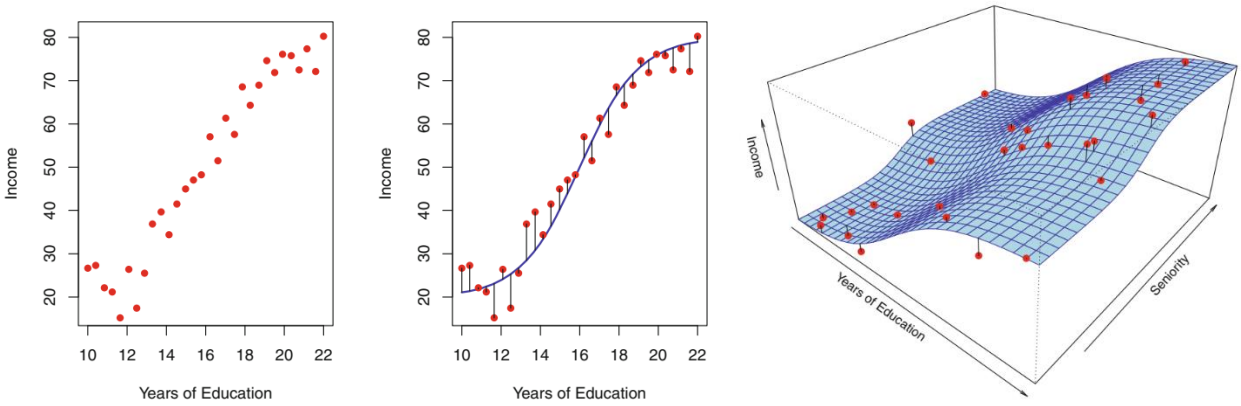
**PRIMERI**: Kvadratura kuće → Cena kuće (regresija)
Godište pacijenta → Stanje pluća (dobro/loše, klasifikacija)

Pretpostavimo da je izlazna promenljiva $Y$ kvantitativnog tipa (može da bude i *kvalitativa*) i da imamo $p$ različitih prediktora, $X_1, X_2, \dots, X_p$. Pretpostavimo da postoji neka veza između izlaza $Y$ i ulaza $X = (X_1, X_2, \dots, X_p)$, koja može biti napisana u opštem obliku:

$$Y = f(X) + e, \tag{1}$$

gde je $f$ nepoznata funkcija, a $e$ je slučajna greška (random error term) koja ne zavisi od $X$, i ima prosek koji ima vrednost 0. Za $f$ se kaže da predstavlja sistematične (systematic) informacije koje $X$ daje za $Y$.

Kao drugi primer prikazan je grafik prihoda u odnosu na godine obrazovanja za 30 pojedinaca. Drugi grafik uključuje još jednu promenljivu – godine iskustva (stepen senioriteta).



Grafik nam govori da bi neko mogao veoma uspešno da predvidi prihod koristeći samo godine koje je neko potrošio na obrazovanje. Međutim, funkcija $f$ koja povezuje ulaznu promenljivu sa izlaznom promenljivom je generalno nepoznata, nju treba otkriti. U ovom slučaju treba proceniti $f$ na osnovu posmatranih tačaka (observacija).


**Stastičko učenje** je skup različitih pristupa za određivanje, tj. aproksimaciju (eng. *estimation*) funkcije $f$.

## Zašto određujemo funkciju $f$? Dva su razloga za to.

Zbog **predikcije** (eng. *prediction*) i **zaključivanja** (eng. *inference*).

# Predikcija

U mnogim situacijama skup ulaznih promenljivih $X$ je odmah dostupan, ali vrednost izlaza nije lako dobiti. U ovom slučaju, pošto je greška $e$ u proseku jednaka nuli, mi možemo da predvidimo $Y$ na sledeći način:

$$\hat{Y} = \hat{f}(X),$$

gde je $\hat{f}$ naša aproksimacije funckije $f$, a $\hat{Y}$ predstavlja predikciju za $Y$. U ovom slučaju se $\hat{f}$ tretira kao crna kutija, i od nje se očekuje da precizno predviđa vrednosti za $Y$.

**PRIMER**: pacijent i negov rizik za ozbiljnu neželjenu reakciju na određeni lek.

Tačnost kako $\hat{Y}$ predviđanja $Y$ zavisi od dve veličine: greške koja se može smanjiti (eng. *reducible error*) i greške koja se ne može smanjiti (eng. *irreducible error*). Generalno, $\hat{f}$ neće biti pefektna aproksimacija za $f$, tako da i ova netačnost dovodi do određene greške. Ova greška se može smanjiti, jer potencijalno možemo poboljšati tačnost $\hat{f}$ koristeći najprikladniju statističku tehniku učenja za procenu funkcije $f$. *Međutim, čak i kada bi bilo moguće formirati savršenu aproksimaciju za $f$ tako da naš procenjeni odgovor dobje oblik $\hat{Y} = f(X)$, naše predviđanje bi i dalje imalo neku grešku u njemu!* ZAŠTO? To je zato što je $Y$ takođe funkcija od $e$ (vidi jednačinu (1)), koju po definiciji nije moguće predvideti pomoću $X$. Zbog toga promenljivost koja potiče od $e$ takođe utiče na tačnost naših predviđanja. Ovo je poznato kao *irreducible* greška, jer koliko god dobro procenili $f$, ne možemo smanjiti grešku koju sa sobom nosi $e$.

### Zašto je *irreducible* greška veća od nule?

1) Veličina $e$ može sadržati neizmerene promenljive koje su korisne u predviđanju $Y$: pošto ih ne merimo, $f$ ih ne može koristiti za predviđanje.
2) Veličina $e$ takođe može sadržati nemerljive varijacije. Na primer, rizik od neželjenih reakcija može da varira za datog pacijenta određenog dana, u zavisnosti od proizvodnih varijacija samog leka ili opšteg osećaja pacijenta tog dana. Dodatni primer je fudbal.

Neka je data aproksimacija $\hat{f}$ i skup prediktora $X$ koji određuju $\hat{Y} = \hat{f}(X)$. Ako su $\hat{f}$ i $X$ fiksirani, tada važi

$$E\left(Y - \hat{Y}\right)^2 = E\left[f(X) + e - \hat{f}(X)\right]^2 = [f(X) - \hat{f}(X)]^2 + Var(e),$$

gde $E\left(Y - \hat{Y}\right)$ predstavlja prosek, ili očekivanu vrednost, kvadarata razlike između predviđene i prave vredosti $Y$, dok $Var(e)$ predstavlja varijansu greške $e$.

**Jedan od ciljeva ovog predmeta je vezan za tehnike za procenu $f$ sa ciljem minimiziranja reducibilne greške.**

## Zaključivanje (Inference)

Često smo zainteresovani za razumevanje načina na koji $X_1, X_2, \dots, X_p$ utiču na $Y$. U ovoj situaciji želimo da procenimo $f$, ali naš cilj nije nužno da predvidimo $Y$. Umesto toga, želimo da razumemo odnos između $X$ i $Y$, ili tačnije, kako se $Y$ menja u zavisnosti od $X_1, X_2, \dots, X_p$. Sada se $\hat{f}$ ne tretira kao crna kutija, jer želimo da znamo njen tačan oblik. Pod ovim uslovima neko može biti zainteresovan za odgovore na sledeća pitanja:

- **Koji su prediktori povezani sa odgovorom?** Čest je slučaj da je samo mali deo dostupnih prediktora u suštini povezan sa $Y$. Identifikovanje nekoliko važnih prediktora među velikim brojem mogućih promenljivih može biti izuzetno korisno, u zavisnosti od primene.
- **Kakav je odnos između odgovora i svakog prediktora?** Neki prediktori mogu imati pozitivan odnos sa $Y$, u tom smislu da je povećanje prediktora povezano sa povećanjem vrednosti od $Y$. Ostali prediktori mogu imati suprotan odnos. U zavisnosti od složenosti za $f$, odnos između odgovora i datog prediktora takođe može zavisiti od vrednosti ostalih prediktora.
- **Može li se odnos između $Y$ i svakog prediktora adekvatno prikazati pomoću linearne jednačine ili je veza složenija?** Istorijski gledano, većina metoda za procenu $f$ imala je linearni oblik. U nekim situacijama je takva pretpostavka razumna ili čak poželjna. Ali često

je pravi odnos složeniji, i u tom slučaju linearni model možda neće pružiti tačan prikaz odnosa između ulaznih i izlaznih promenljivih.

U narednom periodu videćemo niz metoda koji spadaju u grupu onih metoda kojima je bitnija predikcija, nekima je bitnije zaključivanje, a negde ćemo videti i kombinaciju obe vrste.

PRIMERI:

Posmatrajmo kompaniju koja je zainteresovana za sprovođenje kampanje direktnog marketinga. *Na osnovu zapažanja demografskih promenljivih izmerenih kod svakog pojedinca, cilj je identifikovanje pojedinaca koji će odgovoriti pozitivno na slanje pošte*. U ovom slučaju demografske promenljive služe kao prediktori, a odgovor na marketinšku kampanju (bilo pozitivan ili negativan) služi kao ishod. Kompaniju ne zanima dubinsko razumevanje odnosa između svakog pojedinačnog prediktora i odgovora; umesto toga, kompanija jednostavno želi tačan model za predviđanje odgovora pomoću prediktora. Ovo je primer modeliranja za predviđanje (**prediction**).

Suprotno tome, uzmite u obzir primer o oglašavanju sa prve strane skripte. Menadžeri mogu biti zainteresovani za odgovore na pitanja kao što su:

- Koji mediji doprinose prodaji?
- Koji mediji generišu najveći podsticaj prodaje?
- Koliko je povećanje prodaje povezano sa datim povećanjem TV-a oglašavanje?

Ova situacija spada u paradigmu zaključivanja (**inference**).

Drugi primer uključuje modeliranje određene marke proizvoda koju kupac može kupiti na osnovu promenljivih kao što su cena, lokacija prodavnice, nivoi popusta, cena konkurencije itd. U ovoj situaciji nekoga zaista može najviše zanimati kako svaka od pojedinačnih promenljivih utiče na verovatnoću kupovine. Na primer, kakav će efekat imati promena cene proizvoda na prodaju? Ovo je primer modeliranja za zaključivanje (**inference**).

Konačno, moglo bi se sprovesti određeno modeliranje kako za predviđanje tako i za zaključivanje. Na primer, u okruženju nekretnina, može se tražiti da se vrednosti kuća povežu sa inputima kao što su stopa kriminala, zona, udaljenost od reke, kvalitet vazduha, škole, nivo prihoda zajednice, veličina kuća i tako dalje. U ovom slučaju nekoga može zanimati kako pojedine ulazne promenljive utiču na cene – tj. koliko će dodatna vrednost vredeti kuća ako ima pogled na reku? Ovo je problem koji je vezan za **zaključivanje**. Na drugoj strani, neko može jednostavno biti zainteresovan za predviđanje vrednosti kuće s obzirom na njene karakteristike: da li je ova kuća potcenjena ili precenjena? Ovo je problem **predviđanja**.

U zavisnosti od toga da li je naš krajnji cilj predviđanje, zaključivanje ili kombinacija ova dva cilja, različite metode za procenu $f$ mogu biti prikladne. Na primer, linearni modeli omogućavaju relativno jednostavno i interpretabilno zaključivanje, ali možda neće dati tako tačna predviđanja kao neki drugi pristupi. Suprotno tome, neki od krajnje nelinearnih pristupa o kojima ćemo biti reči kasnije mogu potencijalno pružiti prilično tačna predviđanja za $Y$, ali to se postiže na račun manje intepretabilnog modela za koji je zaključivanje veliki izazov.

## Kako određujemo $f$?

Podrazumevaćemo da je sa $n$ označen broj različitih podataka (observacija) sa kojima raspolažemo. Ove podatke nazivamo *trening skupom*, jer ih koristimo za treniranje (učenje) metoda kako da aproksimiraju $f$. Neka je sa $x_{ij}$ označena vrednost $j$-tog prediktora za observaciju $i$, pri čemu je $i = 1, 2, \ldots, n$ i $j = 1, 2, \ldots, p$. Neka je sa $y_i$ označen izlaz za observaciju $i$. Tada trening podatke možemo prikazati $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, gde je $x_i = (x_{i1}, x_{i2,} \ldots, x_{ip})^T$.

Cilj nam je da primenimo neku statističku metodu učenja nad trening podacima kako bismo procenili nepoznatu funkciju $f$. Drugim rečima, želimo da pronađemo funkciju $\hat{f}$ takvu da je $Y \approx \hat{Y} = \hat{f}(X)$ za bilo koju observaciju $(X, Y)$.

Većina statističkih metoda učenja za ovaj zadatak se može svrstati u grupu *parametarskih* ili *neparametarskih* metoda.

### Parametarske metode

Parametarske metode uključuju pristup koji se zasniva na modelu iz dva koraka.

1. Prvo se pravi pretpostavka o obliku funkcije $f$. Na primer, jedna vrlo jednostavna pretpostavka je da je $f$ linearno zavisna od $X$:

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

   Ovo je primer *linearnog modela*. Sada, umesto da se mora proceniti potpuno proizvoljna $p$-dimenzionalna funkcija $f(X)$, potrebno je samo proceniti $p + 1$ koeficijenata $\beta_0, \beta_1, \ldots, \beta_p$.

2. Nakon odabira modela potreban nam je postupak koji koristi podatke za trening za treniranje modela.

$$Y \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Jedan od najčešće korišćenih pristupa je metoda najmanjih kvadrata (eng. *least squares*), ali naravno, postoje i drugi pristupi za određivanje ovih koeficijenata.

Upravo opisani pristup zasnovan na modelu naziva se **parametarskim**; zamenjuje problem procene funkcije $f$ problemom ocenjivanja skupa parametara. Potencijalni nedostatak parametarskog pristupa je taj što se model koji odaberemo obično neće podudarati sa pravim nepoznatim oblikom $f$. Ako je izabrani model predaleko od pravog oblika funkcije $f$, naša procena će biti loša. Ovaj problem se može rešiti izborom fleksibilnih modela koji mogu da se uklope u mnogo različitih mogućih oblika za $f$. Ali generalno, izbor fleksibilnijeg modela zahteva procenu većeg broja parametara.

Ovi složeniji modeli mogu dovesti do pojave poznate kao prekomerno učenje iz podataka, preučavanje (eng. *overfitting*), što u suštini znači da aproksimacija previše dobro prati greške.

MODELI:

- Logistic Regression
- Linear Discriminant Analysis
- Perceptron
- Naive Bayes
- Simple Neural Networks

## Neparametarske metode

Neparametarske metode ne daju eksplicitne pretpostavke o funkcionalnom obliku od $f$. Umesto toga, ove metode traže procenu za $f$ koja se približava dostupnim observacijama što je više moguće, a da pritom modeli ne budu previše prosti ili previše složeni.

Glavni nedostatak neparametarskih pristupa je taj što problem procene za $f$ ne svode na mali broj parametara, već je potreban veoma veliki broj observacija (daleko više nego što je obično potrebno za parametarski pristup) da bi se dobila precizna procena za $f$. Dakle, očekuje se da imate dosta veliki skup podataka za treniranje.

MODELI:

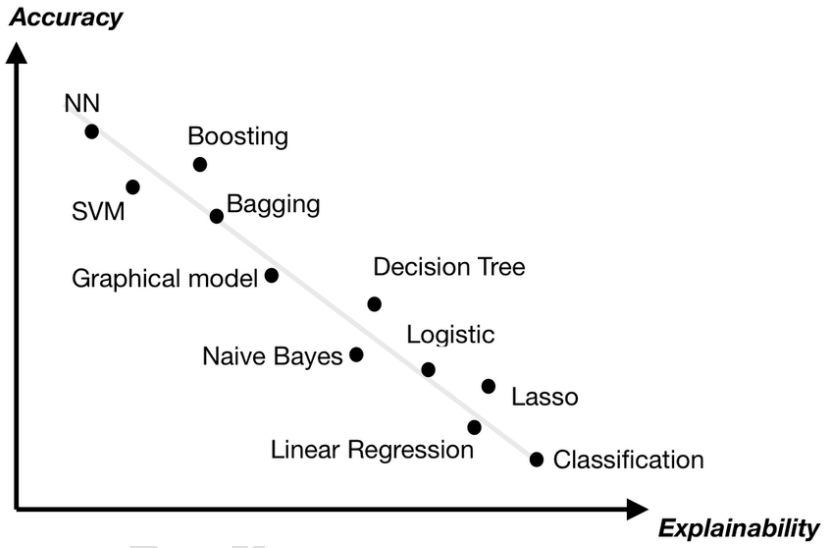
- k-Nearest Neighbors
- Decision Trees like CART and C4.5
- Support Vector Machines

## Kompromis između tačnosti predviđanja modela i interpretabilnosti

Od mnogih metoda koje ćemo da pomenemo/naučimo, neke su manje fleksibilne, tj. restriktivnije, u smislu da mogu da generišu samo relativno mali raspon oblika za procenu $f$. Na primer, linearna regresija je relativno nefleksibilan pristup.

Druge metode, poput *thin plate splines* znatno su fleksibilnije, jer mogu generisati mnogo širi spektar mogućih oblika za procenu $f$.

Postoji nekoliko razloga zbog kojih bismo više voleli restriktivnije modele. Ako nas uglavnom zanima zaključivanje, onda su restriktivni modeli mnogo interpretabilniji. Dok, na primer, *boosting* metode dovode do tako komplikovanih procena za $f$ da je teško razumeti kako je bilo koji pojedinačni prediktor povezan sa odgovorom.



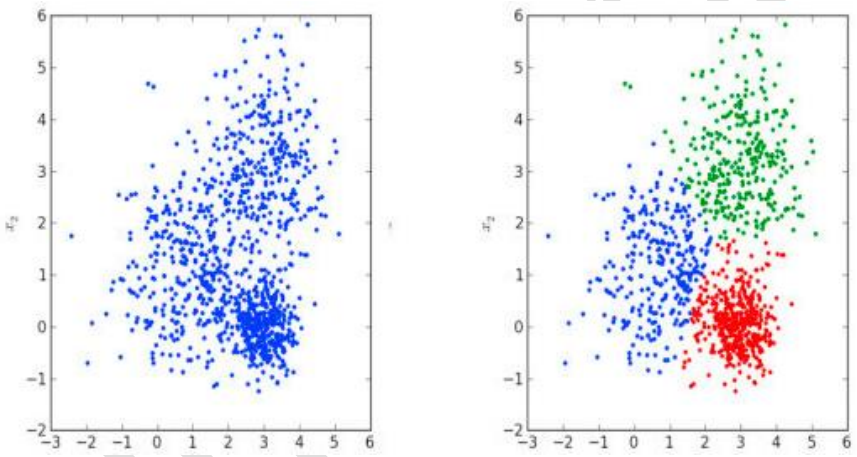

## Supervised Versus Unsupervised Learning

**Prethodno šta ste naučili:** predviđaju se vrednosti za jedan ili više izlaza, ili preciznijim rečnikom, predviđaju se vrednosti zavisne promenljive $Y = (Y_1, Y_2, \dots, Y_m)$ za dati skup ulaza ili prediktora $X^T = (X_1, X_2, \dots, X_p)$. Neka je $x_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$ ulaz za $i$ -tu observaciju u trening skupu, i neka je $y_i$ izlazna vrednost koju predviđamo.

Nove predikcije se zasnivaju na trening primerima $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ prethodno rešenih slučajeva gde su pridružene vrednosti svih parova unapred poznate. Ovaj proces kreiranja modela se naziva ***nadgledano učenje*** ili „učenje sa učiteljom“.

Metafora „student“ predstavlja odgovor $\hat{y}_i$ za svaki ulaz $x_i$ u trening skupu, dok „učitelj“ daje tačan odgovor i grešku koja je pridružena svakom studentovom odgovoru. Ovo se obično karakteriše nekom funkcijom gubitka $L(y, \hat{y})$, kao na primer $L(y, \hat{y}) = (y - \hat{y})^2$.

Šta je **nenadgledano učenje** ili **„učenje bez učitelja“**? U ovom slučaju imamo skup observacija $(x_1, x_2, \dots, x_N)$ slučajnog $p$-vektora $X$ koji imaju zajedničku gustinu $Pr(X)$. Cilj nenadgledanog učenja je da se direktno zaključi o svojstvima funkcije gustine bez pomoći supervizora ili učitelja koji sve vreme pruža tačne odgovore. *Veoma često dimenzije ovih vektora su veće nego što su to vektori u slučaju nadgledanog učenja, a svojstva koja treba otkriti su dosta veoma komplikovana.* Ovi faktori su nešto ublaženi činjenicom da $X$ predstavlja sve promenljive koje se posmatraju; nije potrebno donositi zaključke kako se svojstva $Pr(X)$ menjaju, pod uslovom da se menjaju vrednosti drugog skupa promenljivih.



## Regresioni i klasifikacioni problemi

Promenljive se mogu okarakterisati kao *kvantitativne* ili *kvalitativne* (kategorijske).

# Procena tačnosti modela

*Zašto je potrebno uvoditi toliko različitih statističkih pristupa učenju, a ne samo jedan najbolji metod?* U statistici važi pravilo da nema „besplatnog ručka“ (eng. *no free lunch*): nijedna metoda ne dominira nad svim ostalim metodama, nad svim mogućim skupovima podataka. Na određenom skupu podataka, jedna određena metoda može najbolje raditi, ali neka druga metoda može bolje raditi na sličnom, ali različitom skupu podataka. Stoga je važan zadatak odlučiti koja metoda daje najbolje rezultate za dati skup podataka. Izbor najboljeg pristupa može biti jedan od najizazovnijih delova izvođenja statističkog učenja u praksi.

Da bismo procenili učinak statističke metode učenja na datom skupu podataka, potreban nam je način da izmerimo koliko se njegova predviđanja zapravo podudaraju sa posmatranim podacima.

Za regresione probleme najčešće korišćena matrika je *mean squared error (MSE)* koja je data formulom:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

gde je $\hat{f}(x_i)$ predikcija koju $\hat{f}$ daje za $i$-tu observaciju.