Statistika

Vanja Wagner

5. Statistički testovi

Testiranje pripadnosti diskretnoj distribuciji

Promatramo diskretno obilježje s konačno mnogo vrijednosti: $x_1, x_2, \ldots, x_k$. Modeliramo to obilježje u danoj populaciji sa slučajnom varijablom X.

Cilj: ustvrditi je li razdioba obilježja u populaciji jednako zadanoj razdiobi.

Testiranje pripadnosti diskretnoj distribuciji

Promatramo diskretno obilježje s konačno mnogo vrijednosti: $x_1, x_2, \ldots, x_k$. Modeliramo to obilježje u danoj populaciji sa slučajnom varijablom X.

Cilj: ustvrditi je li razdioba obilježja u populaciji jednako zadanoj razdiobi.

Hipoteza: H_0 : distribucija obilježja X u populaciji jednaka je zadanoj distribuciji:

$$\begin{pmatrix} x_1 & x_2 & \ldots & x_k \\ p_1 & p_2 & \ldots & p_k \end{pmatrix},$$

odnosno, testiramo hipotezu da je $\mathbf{P}(X = x_i) = p_i$, $i = 1, \ldots, k$. $(H_1 : \text{ne } H_0)$

Testiranje pripadnosti diskretnoj distribuciji

Promatramo diskretno obilježje s konačno mnogo vrijednosti: $x_1, x_2, \ldots, x_k$. Modeliramo to obilježje u danoj populaciji sa slučajnom varijablom X.

Cilj: ustvrditi je li razdioba obilježja u populaciji jednako zadanoj razdiobi.

Hipoteza: H_0 : distribucija obilježja X u populaciji jednaka je zadanoj distribuciji:

$$\begin{pmatrix} x_1 & x_2 & \ldots & x_k \\ p_1 & p_2 & \ldots & p_k \end{pmatrix},$$

odnosno, testiramo hipotezu da je $\mathbf{P}(X = x_i) = p_i$, $i = 1, \ldots, k$. (H_1 : ne H_0)

Testiranje provodimo na temelju uzorka (mjerenja promatranog obilježja u populaciji), koji je reprezentiran frekvencijskom tablicom:

Kategorija	Frekvencija
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_k	f_k
Ukupno	n

Pri tome je:
f_i - frekvencija vrijednosti x_i u uzorku (opažena frekvencija, O_i)
n - duljina uzorka

Pearsonov χ^2 test pripadnosti distribuciji

Ovu hipotezu ćemo testirati preko Pearsonovog χ^2 testa.

Pearsonov χ^2 test pripadnosti distribuciji

Ovu hipotezu ćemo testirati preko Pearsonovog χ^2 testa.

Statistika: Za zadanu testnu razdiobu

$$\begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}$$

i uzorak veličine n definiramo **teorijske frekvencije**:

$$E_i = n \cdot p_i.$$

Rezultat možemo prikazati tablično

	Frekvencije	
Kategorija	Opažena	Teorijska
x_1	f_1	E_1
x_2	f_2	E_2
$\vdots$	$\vdots$	$\vdots$
x_k	f_k	E_k
Ukupno	n	n

Uočimo da E_i predstavlja očekivani broj pojava vrijednosti x_i u uzorku duljine n za obilježje X ako vrijedi

Pearsonov χ^2 test pripadnosti distribuciji

Definiramo statistiku

$$Hi^2 = \sum_i \frac{(f_i - E_i)^2}{E_i}.$$

Pearsonov χ^2 test pripadnosti distribuciji

Definiramo statistiku

$$Hi^2 = \sum_i \frac{(f_i - E_i)^2}{E_i}.$$

Ako vrijedi H_0, Hi^2 je distribuirana prema χ^2 razdiobi s $k - p - 1$ stupnjeva slobode:

$$Hi^2 \sim \chi^2(k - p - 1).$$

Ovdje je p broj parametara koje smo morali procijeniti da bismo posve definirali testnu razdiobu.

Pearsonov χ^2 test pripadnosti distribuciji

Definiramo statistiku

$$Hi^2 = \sum_i \frac{(f_i - E_i)^2}{E_i}.$$

Ako vrijedi H_0, Hi^2 je distribuirana prema χ^2 razdiobi s $k - p - 1$ stupnjeva slobode:

$$Hi^2 \sim \chi^2(k - p - 1).$$

Ovdje je p broj parametara koje smo morali procijeniti da bismo posve definirali testnu razdiobu.

Pretpostavka χ^2 testa: $E_i \geqslant 5$ (postoje razne modifikacije)

Pearsonov χ^2 test pripadnosti distribuciji

Definiramo statistiku

$$Hi^2 = \sum_i \frac{(f_i - E_i)^2}{E_i}.$$

Ako vrijedi H_0, Hi^2 je distribuirana prema χ^2 razdiobi s $k - p - 1$ stupnjeva slobode:

$$Hi^2 \sim \chi^2(k - p - 1).$$

Ovdje je p broj parametara koje smo morali procijeniti da bismo posve definirali testnu razdiobu.

Pretpostavka χ^2 testa: $E_i \geqslant 5$ (postoje razne modifikacije)

Kritično područje: uz razinu značajnosti $\alpha \in (0, 1)$ je

$$[\chi^2_{1-\alpha}(k - p - 1), \infty\rangle,$$

pri čemu je $\chi^2_{1-\alpha}(k - p - 1)$ $1 - \alpha$ kvantil χ^2 razdiobe s $k - p - 1$ stupnjeva slobode.

Primjer.

Istraživač za slučajni izbor u uzorak koristi igraću kocku. Prije početka izbora uzorka htio je provjeriti je li njegova kocka simetrična (svi brojevi imaju iste vjerojatnosti). Kocku je bacio 120 puta i zabilježio frekvencije brojeva:

Vrijednost	Opažena frekvencija
1	23
2	15
3	16
4	28
5	17
6	21
Ukupno	120

Je li kocka simetrična?

Primjer.

Istraživač za slučajni izbor u uzorak koristi igraću kocku. Prije početka izbora uzorka htio je provjeriti je li njegova kocka simetrična (svi brojevi imaju iste vjerojatnosti). Kocku je bacio 120 puta i zabilježio frekvencije brojeva:

Vrijednost	Opažena frekvencija
1	23
2	15
3	16
4	28
5	17
6	21
Ukupno	120

Je li kocka simetrična?Označimo s X rezultat bacanja kocke. Nulta hipoteza je:

$$H_0: \; X \sim \begin{pmatrix} 1 & 2 & \dots & 6 \\ 1/6 & 1/6 & \dots & 1/6 \end{pmatrix}.$$

Primjer.

Istraživač za slučajni izbor u uzorak koristi igraću kocku. Prije početka izbora uzorka htio je provjeriti je li njegova kocka simetrična (svi brojevi imaju iste vjerojatnosti). Kocku je bacio 120 puta i zabilježio frekvencije brojeva:

Vrijednost	Opažena frekvencija
1	23
2	15
3	16
4	28
5	17
6	21
Ukupno	120

Je li kocka simetrična?Označimo s X rezultat bacanja kocke. Nulta hipoteza je:

$$H_0 : \; X \sim \begin{pmatrix} 1 & 2 & \dots & 6 \\ 1/6 & 1/6 & \dots & 1/6 \end{pmatrix}.$$

Izračunamo teorijske frekvencije ($E_i = 120 \cdot \frac{1}{6} = 20$) i u Excelu pozovemo funkciju

```
=CHISQ.TEST(f1:f6,E1:E6).
```

Funkcija vraća p-vrijednost testa. U ovom primjeru je $p = 0.287 \geqslant 0.05$ pa ne odbacujemo pretpostavku da je kocka simetrična.

Primjer.

Želimo ispitati je li broj rekombinacija X tijekom mejoze na jednom kromosomu neke biljne vrste Poissonova slučajna varijabla,

$$\mathbf{P}(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 0, 1, 2, \ldots$$

za neki parametar λ (očekivani broj rekombinacija). Pretpostavljamo da imamo podatke nakon $n = 60$ mejoza danih tablicom

Broj rekombinacija	0	1	2	3	Ukupno
Frekvencija	32	15	9	4	60

Primjer.

Želimo ispitati je li broj rekombinacija X tijekom mejoze na jednom kromosomu neke biljne vrste Poissonova slučajna varijabla,

$$\mathbf{P}(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 0, 1, 2, \dots$$

za neki parametar λ (očekivani broj rekombinacija). Pretpostavljamo da imamo podatke nakon $n = 60$ mejoza danih tablicom

Broj rekombinacija	0	1	2	3	Ukupno
Frekvencija	32	15	9	4	60

Da bismo procijenili parametar λ, prisjetimo se da je očekivanje Poissonove razdiobe upravo jednako λ, stoga je prirodan procjenitelj upravo aritmeticka sredina uzorka $\overline{x}$. Kako je

$$\hat{\lambda} = x = \frac{0 \cdot 32 + 1 \cdot 15 + 2 \cdot 9 + 3 \cdot 4}{60} = 0.75$$

možemo odrediti teorijske frekvencije

$$E_i = 60 \cdot \frac{0.75^j}{j!} e^{-0.75}.$$

Primjer.

Dobivamo tablicu

Broj rekombinacija	0	1	2	3	Ukupno
Opažena frekvencija	32	15	9	4	60
Teorijska frekvencija	28.34	21.26	7.97	1.99	59.56

Primjer.

Dobivamo tablicu

Broj rekombinacija	0	1	2	3	Ukupno
Opažena frekvencija	32	15	9	4	60
Teorijska frekvencija	28.34	21.26	7.97	1.99	59.56

Uočimo:

- Teorijska frekvencija za razred 3 je manja od 5 - grupiramo ga s razredom 2.
- Ukupna teorijska frekvjencija je $59.56 < 60$, što je posljedica toga da Poissonova slučajna varijabla može poprimiti i vrijednosti $\geqslant 4$.

Primjer.

Dobivamo tablicu

Broj rekombinacija	0	1	2	3	Ukupno
Opažena frekvencija	32	15	9	4	60
Teorijska frekvencija	28.34	21.26	7.97	1.99	59.56

Uočimo:

- Teorijska frekvencija za razred 3 je manja od 5 - grupiramo ga s razredom 2.
- Ukupna teorijska frekvjencija je $59.56 < 60$, što je posljedica toga da Poissonova slučajna varijabla može poprimiti i vrijednosti $\geqslant 4$.

$\longrightarrow$ grupiramo zadnja dva razreda i preimenujemo ga u razred $\geqslant 2$ te njegovu teorijsku frekvenciju odredimo kao

$$E_2 = 60 \cdot \mathbf{P}(Y \geqslant 2) = 60 - E_0 - E_1.$$

Dobivamo tablicu frekvencija

Broj rekombinacija	0	1	$\geqslant 2$	Ukupno
Opažena frekvencija	32	15	13	60
Teorijska frekvencija	28.34	21.26	10.4	60

Primjer.

Ne možemo kao u prošlom primjeru koristiti istu Excel naredbu za dobivanje p-vrijednosti, obzirom da imamo **jedan procijenjeni parametar**.

Stoga ćemo sami odrediti vrijednost Hi^2 testne statistike, $Hi^2 = 2.94$, te pomoću nje p-vrijednost pozivanjem funkcije

$$p = \texttt{CHISQ.DIST.RT(2.94, 3-1-1)|} = 0.086410733.$$

Kako je $p \geqslant 0.05$ ne odbacujemo pretpostavku da broj rekombinacija slijedi Poissonovu distribuciju.

Ostali χ^2-testovi

Dva često korištena χ^2 testa su test **nezavisnosti** i test **homogenosti** dvaju obilježja X i Y. Provođenje ova dva testa je veoma slično, ali interpretacija je drukčija:

- test nezavisnosti - nulta hipoteza je da su obilježja X i Y **nezavisna**;
- test homogenosti - nulta hipoteza je da su obilježja X i Y (ili isto obilježje promatrano u dvije populacije) **jednako distribuirana**;

6. Korelacija i linearna regresija

Korelacija slučajnih varijabli

Za slučajne varijable X i Y **kovarijanca** se definira kao

$$Cov(X, Y) = \mathbf{E}\left[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))\right] = \mathbf{E}[X \cdot Y] - \mathbf{E}[X] \cdot \mathbf{E}[Y].$$

Korelacija slučajnih varijabli

Za slučajne varijable X i Y **kovarijanca** se definira kao

$$Cov(X, Y) = \mathbf{E}\left[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))\right] = \mathbf{E}[X \cdot Y] - \mathbf{E}[X] \cdot \mathbf{E}[Y].$$

Kovarijanca mjeri zajedničku varijaciju varijabli X i Y (izražena je u mjernim jedinicama slučajnih varijabli X i Y).

Korelacija slučajnih varijabli

Za slučajne varijable X i Y **kovarijanca** se definira kao

$$Cov(X,Y) = \mathbf{E}\left[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))\right] = \mathbf{E}[X \cdot Y] - \mathbf{E}[X] \cdot \mathbf{E}[Y].$$

Kovarijanca mjeri zajedničku varijaciju varijabli X i Y (izražena je u mjernim jedinicama slučajnih varijabli X i Y).

S druge strane, **korelacija**:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y},$$

nema mjernu jedinicu i poprima vrijednosti iz skupa $[-1, 1]$.

Korelacija slučajnih varijabli

Za slučajne varijable X i Y **kovarijanca** se definira kao

$$Cov(X,Y) = \mathbf{E}\left[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))\right] = \mathbf{E}[X \cdot Y] - \mathbf{E}[X] \cdot \mathbf{E}[Y].$$

Kovarijanca mjeri zajedničku varijaciju varijabli X i Y (izražena je u mjernim jedinicama slučajnih varijabli X i Y).

S druge strane, **korelacija**:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y},$$

nema mjernu jedinicu i poprima vrijednosti iz skupa $[-1, 1]$.

Može se pokazati da ako su X i Y **nezavisne** slučajne varijable, tada je

$$E(X \cdot Y) = E(X) \cdot E(Y),$$

pa je

$$Cov(X,Y) \quad = \quad 0 \text{ i } Corr(X,Y) = 0.$$

Koreliracija i linearna povezanost

Za **linearno povezane** slučajne varijable X i Y:

$$Y = a \cdot X + b$$

vrijedi

$$Var(Y) = a^2 Var(X), \quad \text{tj.} \quad \sigma_Y = |a|\sigma_X \implies Corr(X, Y) = \begin{cases} 1, & a > 0 \\ -1, & a < 0 \end{cases}.$$

Koreliracija i linearna povezanost

Za **linearno povezane** slučajne varijable X i Y:

$$Y = a \cdot X + b$$

vrijedi

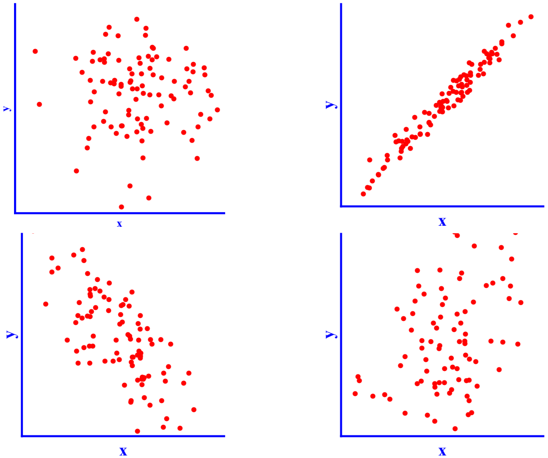
$$Var(Y) = a^2 Var(X), \quad \text{tj.} \quad \sigma_Y = |a|\sigma_X \implies Corr(X,Y) = \begin{cases} 1, & a > 0 \\ -1, & a < 0 \end{cases}.$$

Bitna svojstva:

- Slučajne varijable X i Y **nezavisne** $\rightarrow$ $Corr(X,Y) = 0$.
- Slučajne varijable X i Y **linearno povezane** $\rightarrow$ $Corr(X,Y) = \pm 1$.
- Uočimo, ako su X i Y linearno povezane onda su i zavisne.
- Može se pokazati da $Corr(X,Y) = \pm 1 \rightarrow X$ i Y su linearno povezane.
- Uočimo da $Corr(X,Y) = 0$ ne znači da su X i Y nezavisne slučajne varijable.

Dijagram raspršenosti

Za varijable X i Y promatramo (X, Y)-graf: *scatterplot*, graf raspršenosti.



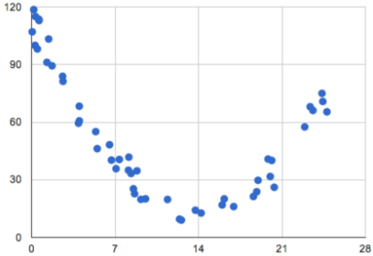
Razmislite koji graf odgovara uzorcima za obilježja (X, Y) koja imaju korelaciju:
a) $\rho = 0$; b) $\rho = 0.46$; c) $\rho = 0.98$; d) $\rho = -0.82$.

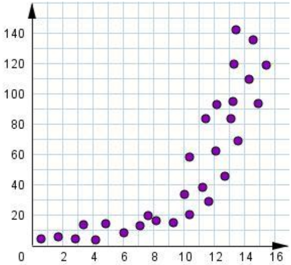
Linearna povezanost i transformacije

Koleracija mjeri **linearnu** povezanost dvije varijable. Što ako su varijable povezane, ali ne linearno?

Linearna povezanost i transformacije

Koleracija mjeri **linearnu** povezanost dvije varijable. Što ako su varijable povezane, ali ne linearno?
Promotrimo dva primjera nelinearne povezanosti:





Možete li naslutiti tip povezanosti? Kako povezanost nije linearna, u oba slučaja korelacija može biti malena.

Procjena Pearsonova koeficijenta korelacije

Korelaciju slučajnih varijabli (obilježja) X i Y (često nazivanu još Pearsonovim koeficijentom korelacije za populaciju) možemo procijeniti temeljem uzorka za ta dva statistička obilježja. Uzorak duljine n dan je mjerenjima

$$(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n),$$

gdje je (x_i, y_i) izmjerena vrijednost obiljžje X i Y na i-toj jedinki uzorka.

Procjena Pearsonova koeficijenta korelacije

Korelaciju slučajnih varijabli (obilježja) X i Y (često nazivanu još Pearsonovim koeficijentom korelacije za populaciju) možemo procijeniti temeljem uzorka za ta dva statistička obilježja. Uzorak duljine n dan je mjerenjima

$$(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n),$$

gdje je (x_i, y_i) izmjerena vrijednost obiljžje X i Y na i-toj jedinki uzorka.

Pearsonov koeficijent korelacije uzorka dan je s:

$$r = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_X \cdot s_Y},$$

gdje su $\bar{x}$, $\bar{y}$ aritmetičke sredine uzoraka, a s_X, s_Y standardne devijacije uzoraka.

U Excelu koristimo naredbu:

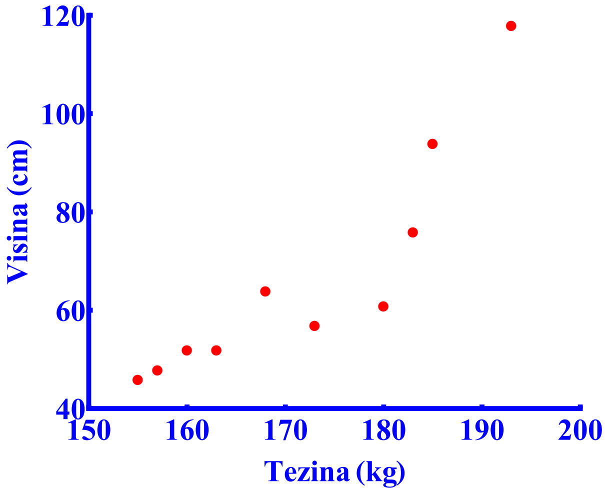
```
=CORREL(X1:Xn;Y1:Yn)
```

Primjer.

Na osnovu uzorka od 10 osoba procijenite koeficijent korelacije za visinu i težinu.

Visina (cm)	**Težina (kg)**
183	76
163	52
180	61
168	64
160	52
157	48
185	94
155	46
193	118
173	57

Dijagram raspršenja:



$$r = 0.89$$

Pearsonov koeficijent korelacije - svojstva i interpretacija

:

- broj iz intervala $[-1, 1]$
- iskazuje smjer i jakost linarne statističke veze između dva obilježja
- r bliži -1 ili 1 $\longrightarrow$ jača korelacija
- $r = 1$ ili $r = -1$ $\longrightarrow$ potpuna povezanost, linearna povezanost
- $r > 0$ $\longrightarrow$ pozitivna korealcija (veći x $\rightarrow$ veći y)
- $r < 0$ $\longrightarrow$ negativna korealcija (veći x $\rightarrow$ manji y)
-
    - $0 - 0.25$ - linearna korelacija slaba
    - $0.25 - 0.64$ - korelacija srednje jačine
    - $0.64 - 1$ - čvrsta korelacija

Interpretacija primjera:

Vratimo se na primjer s visinama i težinama:

- Kod osobe s većom visinom očekujemo i veću težinu (pozitivna koreliranost)
- Kod osobe s većom težinom očekujemo i veću visinu (pozitivna koreliranost)
- Korelacija ne pokazuje uzročno-posljedičnu povezanost!
- Povećanjem težine nećemo povećati visinu.

Testiranje hipoteze o koeficijentu korelacije

Na osnovu uzorka $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ možemo testirati hipotezu

$$H_0 : \quad \rho = 0$$

gdje je ρ Pearsonov koeficijent korelacije (za populaciju).

Testiranje hipoteze o koeficijentu korelacije

Na osnovu uzorka $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ možemo testirati hipotezu

$$H_0 : \quad \rho = 0$$

gdje je ρ Pearsonov koeficijent korelacije (za populaciju).

Koristimo statistiku

$$t = \frac{\sqrt{n-2} \cdot r}{\sqrt{1 - r^2}}$$

koja je uz H_0 distribuirana prema Studentovoj razdiobi: $\quad t \sim t(n-2)$.

Kritično područje na razini značajnosti α je analogno kao i kod t-testa, a ovisi o alternativnoj hipotezi (lijeva/desna jednostrana, dvostrana).

Malo kasnije ćemo vidjeti provođenje testa korelacije u Excelu (uočite da je hipoteza $H_0 : \quad \rho = 0$ ekvivalentna hipotezi $H_0 : \quad a = 0$).

Jednostavna linearna regresija

Jednom kada naslutimo koreliranost dvaju varijabli (obilježja), prirodno se postavlja pitanje kako izgleda linearna povezanost tih varijabli, odnosno procjene parametara a i b iz linearne veze:

$$Y = aX + b,$$

Jednostavna linearna regresija

Jednom kada naslutimo koreliranost dvaju varijabli (obilježja), prirodno se postavlja pitanje kako izgleda linearna povezanost tih varijabli, odnosno procjene parametara a i b iz linearne veze:

$$Y = aX + b,$$

gdje je

- a - koeficijent smjera (*slope*)
- b - slobodni koeficijent (*intercept*)

Jednostavna linearna regresija

Jednom kada naslutimo koreliranost dvaju varijabli (obilježja), prirodno se postavlja pitanje kako izgleda linearna povezanost tih varijabli, odnosno procjene parametara a i b iz linearne veze:

$$Y = aX + b,$$

gdje je

- a - koeficijent smjera (*slope*)
- b - slobodni koeficijent (*intercept*)

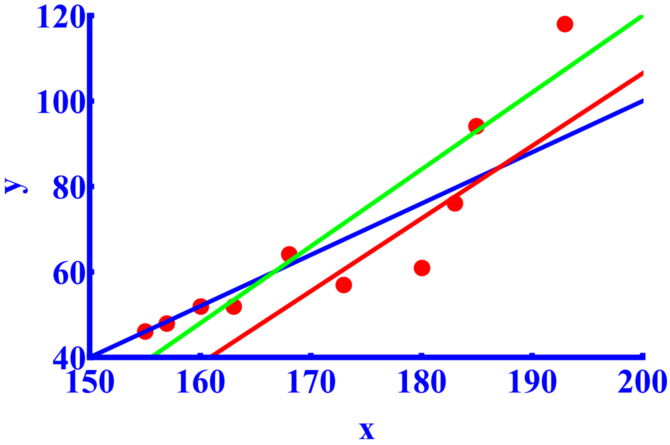
Interpretacija koeficijenata:

- Koeficijent smjera (a) - ukoliko veličinu x povećamo za 1, y će se povećati za a;
- Slobodni koeficijent (b) - za $x = 0$ je $y = b$.

Regresijska analiza - općenito

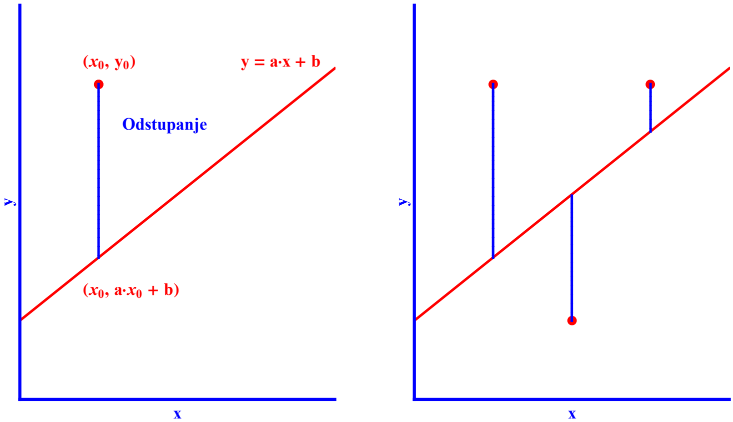
- primjena metoda kojima se analitički (jednadžbom) objašnjava statistička ovisnost jedne varijable o drugoj ili o više drugih
- iz podataka jedne varijable 'prognoziramo' rezultat druge varijable
- zavisna varijabla - varijabla čiju ovisnost objašnjavamo
- nezavisne varijable - objašnjavaju ponašanje zavisne
- zasniva se na modelu
- model je pojednostavljena slika stvarne pojave
- oblik modela ovisi o primjeru kojeg rješavamo
- ako je odnos između dvije pojave oblikom linearan - model jednostavne linearne regresije
- jedna nezavisna varijabla $\longrightarrow$ jednostavna linearna regresija
- više nezavisnih varijabli $\longrightarrow$ multivarijatna regresija

Primjer: podaci za visinu i težinu:



Kako odrediti pravac koji najbolje opisuje podatke?

Metoda najmanjih kvadrata



Za svaki pravac $y = ax + b$ i svaku točku uzorka pogledamo kolika je razlika $y_i - (ax_i + b)$. Tražimo a i b za koje je suma kvadrata tih razlika po svim točkama uzorka najmanja:

$$\min \sum_i (y_i - ax_i - b)^2$$

Regresijski pravac

Pravac koji minimizira srednje kvadratno odstupanje naziva se **regresijski pravac**, a koeficijenti regresijskog pravca nazivaju se **regresijski koeficijenti**.

Regresijski pravac

Pravac koji minimizira srednje kvadratno odstupanje naziva se **regresijski pravac**, a koeficijenti regresijskog pravca nazivaju se **regresijski koeficijenti**.

Eksplicitni izraz za regresijske koeficijente:

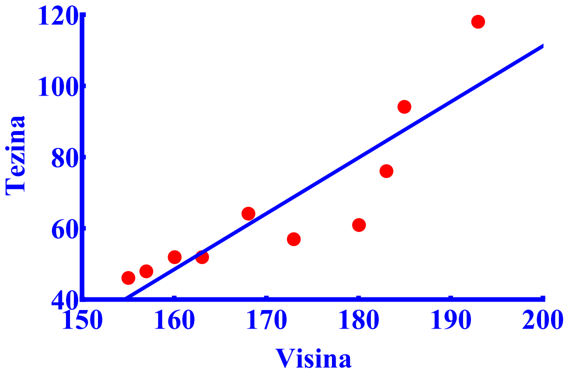
$$\begin{aligned} a &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \\ &= \frac{Cov(X,Y)}{S_X^2} = \\ &= r_{X,Y} \frac{S_Y}{S_x} \\ b &= \bar{Y} - a \cdot \bar{X} \end{aligned}$$

$r_{X,Y}$ - Pearsonov koeficijent korelacije varijabli X i Y

Primjer. Regresijski pravac za podatke o visini i težini

Pokazat ćemo da je regresijski pravac za visine i težine jednak:

$$\text{Težina} = 1.5685 \cdot \text{Visina} - 202.519$$



Linearna regresija u Excelu

U sklopu *Data Analysis* paketa koristimo proceduru *Correlation* koja vraća sljedeći skup analiza:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0,890660908					
R Square	0,793276853					
Adjusted R Square	0,767436459					
Standard Error	11,1458636					
Observations	10					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	3813,757796	3813,758	30,6991	0,000546925	
Residual	8	993,8422037	124,2303			
Total	9	4807,6				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-202,5189988	48,73521949	-4,1555	0,003185	-314,9026164	-90,13538111
Visina	1,568543965	0,283096087	5,540677	0,000547	0,915723218	2,221364711

Regresijeske statistike - analiza

Regression Statistics	
Multiple R	0,890660908
R Square	0,793276853
Adjusted R Square	0,767436459
Standard Error	11,1458636
Observations	10

Koliko dobro regresijski pravac opisuje podatke?

- *Multiple R* - kod linearne regresije to je upravo koeficijent korelacije.
- *R Squared* - **koeficijent determinacije** - kvadrat od *Multiple R* - mjeri koliki udio u ukupnoj varijabilnosti varijable y, možemo objasniti linearnom regresijom.
- *Adjusted R Squared* - *R Squared* prilagođen za broj procijenjenih parametara u modelu (kod jednostavne linearne regresije to su 2 procijenjena parametra)

Testiranje hipoteza o koeficijentima

Uz odrežene pretpostavke na greške modela (razlike između stvarne vrijednosti y_i i procijenjene vrijednosti na pravcu), tzv. Gauss-Markovljeve uvjete, moguće je reći nešto više o samim koeficijentima modela.

Testiranje hipoteza o koeficijentima

Uz odrežene pretpostavke na greške modela (razlike između stvarne vrijednosti y_i i procijenjene vrijednosti na pravcu), tzv. Gauss-Markovljeve uvjete, moguće je reći nešto više o samim koeficijentima modela.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-202,5189988	48,73521949	-4,1555	0,003185	-314,9026164	-90,13538111
Visina	1,568543965	0,283096087	5,540677	0,000547	0,915723218	2,221364711

Testiranje hipoteza o koeficijentima

Uz odrežene pretpostavke na greške modela (razlike između stvarne vrijednosti y_i i procijenjene vrijednosti na pravcu), tzv. Gauss-Markovljeve uvjete, moguće je reći nešto više o samim koeficijentima modela.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-202,5189988	48,73521949	-4,1555	0,003185	-314,9026164	-90,13538111
Visina	1,568543965	0,283096087	5,540677	0,000547	0,915723218	2,221364711

- *Intercept* se odnosi na koeficijent b, *Tezina* se odnosi na koeficijent a (nalazi se uz varijablu X=težina).
- stupac *Coefficients* su procijenjene vrijednosti parametara.
- stupci *t-test* i *p-value* daju vrijednost testne statistike i p-vrijednost u testovima **značajnosti koeficijenata**. To su testovi koji za nultu hipotezu uzimaju $b = 0$, odnosno $a = 0$.
- stupci *t-test* i *p-value* daju vrijednost testne statistike i p-vrijednost u testovima **značajnosti koeficijenata**. To su testovi koji za nultu hipotezu uzimaju $b = 0$, odnosno $a = 0$.
- stupci *Lower 95%* i *Upper 95%* daju donji i gornji rub 95% pouzdanog intervala za vrijednosti koeficijenata.

Značajnost modela

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3813,757796	3813,758	30,6991	0,000546925
Residual	8	993,8422037	124,2303		
Total	9	4807,6			

Značajnost modela

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3813,757796	3813,758	30,6991	0,000546925
Residual	8	993,8422037	124,2303		
Total	9	4807,6			

Promatramo nultu hipotezu da nezavisne varijable (u našem slučaju X = Visina) nisu korelirane sa zavisnom (u našem slučaju Y = Težina), odnosno da promatrani model opisuje Y jednako dobro kao model sa samo slobodnim koeficijentom ($Y = konst.$).

- F je vrijdnost testne statistike testa značajnosti modela;
- *Significance F* je p-vrijednost testa značajnosti modela.