Primijenjena i inženjerska matematika

Statistika - predavanje 7

Statističko zaključivanje - dvije varijable

Korelacija i jednostavna linearna regresija

24.11.2023.

Veza između varijabli

- za parove podataka iz dvije neprekidne slučajne varijable želimo zaključivati o postojanju zavisnosti između njih
- za diskretne slučajne varijable možemo koristiti χ^2 test
- kod neprekidnih slučajnih varijabli zavisnost se može pojaviti na brojne načine
- tipovi veza među varijablama

Deterministička veza

- **deterministička veza** između dvije varijable je veza zadana pravilom oblika

$$y = f(x)$$

  gdje je y zavisna varijabla, x nezavisna varijabla, a $f : \mathbb{R} \to \mathbb{R}$ zadana funkcija
- primjerice, pravilima $y = x + 54$, $y = x^2 - 14x$ i $y = \sin(3x)$ zadane su determinističke veze među varijablama x i y
- za svaku dopuštenu vrijednost nezavisne varijable x možemo izračunati točnu vrijednost zavisne varijable y
- primjer 6.15

Statistički model s aditivnom greškom

- u praktičnim problemima ne možemo očekivati deterministička vezu
- **dijagram raspršenosti** podataka (eng. scatter plot) je prikaz uređenih parova podataka iz dviju varijabli u koordinatnom sustavu
- primjer 6.16

Statistički model s aditivnom greškom

- **regresijska metoda** modeliranja pretpostavlja da možemo uspostaviti funkcijsku vezu ali uz dodanu grešku
- veza između nezavisne varijable x i zavisne slučajne varijable $Y(x)$ će biti oblika

$$Y(x) = f(x) + \varepsilon, \tag{1}$$

  gdje pretpostavljamo da je ε slučajna varijabla koja opisuje grešku u modeliranju
- mnogo nezavisnih slučajnih smetnji u pravilu ima normalnu distribuciju – u primjenama se u klasičnom načinu modeliranja prihvaća da je model adekvatan ako je u njemu postignuta normalna distribuiranost grešaka ε
- primjer 6.17

Statistički model s aditivnom greškom

- sparena mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$ dvaju obilježja koja dolaze od slučajnih varijabli $Y_1, \ldots Y_n$ (čije su realizacije realni brojevi $y_1, \ldots y_n$) i nezavisne varijable x (čije su izmjerene vrijednosti $x_1, \ldots x_n$)
- cilj je utvrditi zavisnost između dvije varijable
- **regresijski model** – matematički model oblika

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

gdje je f realna funkcija jedne realne varijable, a $\varepsilon_1, \ldots, \varepsilon_n$ međusobno nezavisne slučajne varijable takve da je $E\,\varepsilon_i = 0$ i $\mathrm{Var}(\varepsilon_i) = \sigma^2$

Statistički model s aditivnom greškom

- prvi korak u uspostavljanju ovakvih veza među varijablama Y i x prikaz je podataka u dijagramu raspršenosti iz kojeg se lako vidi grupiraju li se sparena mjerenja oko pravca (linearna zavisnost) ili neke krivulje (neka druga funkcijska zavisnost - polinomijalna, logaritamska, . . .).

Regresijski pravac

- pretpostavimo da je graf funkcije f u modelu pravac
- f možemo prikazati formulom $f(x) = \alpha + \beta x$
- slobodni koeficijent α zove se **odsječak na y-osi**, a koeficijent β uz nezavisnu varijablu x zove se **koeficijent smjera** i važan je iz sljedećeg razloga:

  ako je $\beta < 0$ funkcija $f(x) = \alpha + \beta x$ je padajuća
  ako je $\beta > 0$ funkcija $f(x) = \alpha + \beta x$ je rastuća.
- graf funkcije $f(x) = \alpha + \beta x$ nazivamo **regresijskim pravcem**, a koeficijente α i β **regresijskim parametrima**

Statistički model

Linearni regresijski model može se zapisati u obliku

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

Ovdje su:

- $x_1, x_2, \ldots, x_n$ vrijednosti varijable x koje su **izabrane/izmjerene**
- $Y_1, Y_2 \ldots, Y_n$ slučajne varijable (njihove izmjerene vrijednosti su $y_1, \ldots, y_n$)
- α i β su **nepoznati parametri** linearne veze koje treba odrediti u postupku modeliranja – **procijeniti** (to zapravo znači da trebamo **procijeniti regresijski pravac** $y = \alpha + \beta x$)

Statistički model

- $\varepsilon_1, \ldots, \varepsilon_n$ predstavljaju varijable greške koja je dodana na linearnu vezu $(\alpha + \beta x_i)$ – nemjerljive slučajne varijable za koje pretpostavljamo da
  - međusobno su nezavisne
  - sve imaju normalnu distribuciju
  - imaju očekivanje 0
  - sve imaju jednaku varijancu σ^2

Metoda najmanjih kvadrata

- na osnovu podataka želimo procijeniti nepoznate parametre α i β
- tako ćemo dobiti i procjenu nepoznatog regresijskog pravca $y = \alpha + \beta x$
- ako su α i β poznati za svaku izmjerenu vrijednost x_i možemo odrediti broj

$$y_i' = \alpha + \beta x_i$$

- y_i' – **teorijska vrijednost** zavisne varijable u x_i (eng. predicted value) (odgovara vrijednosti očekivanja zavisne varijable u x_i)
- y_i – **izmjerena** ili **eksperimentalna** vrijednost zavisne varijable u x_i (eng. observed value)
- y_i se razlikuje od teorijske vrijednosti pa točke (x_i, y_i), $i = 1, \ldots, n$, uglavnom ne leže na pravcu $y = \alpha + \beta x$

Metoda najmanjih kvadrata

- parametre ćemo odrediti tako da razlike između izmjerenih i teorijskih vrijednosti budu što manje
- metoda koju koristimo naziva se **metoda najmanjih kvadrata**
- procjenu treba odrediti tako da funkciju

$$D(\alpha,\beta) = \sum(\text{eksperimentalne v.} - \text{teorijske v.})^2$$
$$= \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

  učinimo što manjom
- procjene $\widehat{\alpha}$ i $\widehat{\beta}$ regresijskih parametara α i β trebamo odrediti tako da vrijedi:

$$D(\widehat{\alpha},\widehat{\beta}) = \min_{(\alpha,\beta)\in\mathbb{R}^2} D(\alpha,\beta)$$

Metoda najmanjih kvadrata

- procjene $\widehat{\alpha}$ i $\widehat{\beta}$ nazivamo **procjenama u smislu metode najmanjih kvadrata** (eng. least squares estimates) regresijskih parametara α i β
- procjena nepoznatog regresijskog pravca $y = \alpha + \beta x$ je pravac $y = \widehat{\alpha} + \widehat{\beta} x$

Metoda najmanjih kvadrata

- procjene se mogu eksplicitno izraziti:

$$\widehat{\beta} = \frac{s_{xy}}{s_x^2}, \quad \widehat{\alpha} = \overline{y}_n - \widehat{\beta}\,\overline{x}_n,$$

gdje su

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$s_x^2 = \sum_{i=1}^{n}(x_i - \overline{x}_n)^2, \quad s_y^2 = \sum_{i=1}^{n}(y_i - \overline{y}_n)^2,$$

$$s_{xy} = \sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n).$$

Metoda najmanjih kvadrata

- koristeći formulu procijenjenog regresijskog pravca $y = \widehat{\alpha} + \widehat{\beta}x$, za svaku vrijednost x možemo izračunati pripadnu procjenu teorijske vrijednosti – **predikcija**
- $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i$ — predikcija zavisne varijable za vrijednost x_i nezavisne varijable
- odstupanje procijenjene vrijednosti $\widehat{y}_i$ od izmjerene vrijednosti y_i zavisne varijable:

$$e_i = y_i - \widehat{y}_i = y_i - (\widehat{\alpha} + \widehat{\beta}x_i).$$

- $e_1, \ldots, e_n$ zovemo **rezidualima** i možemo ih smatrati procjenama grešaka $\varepsilon_1, \ldots, \varepsilon_n$ iz modela $Y_i = \alpha + \beta x_i + \varepsilon_i$

Metoda najmanjih kvadrata

- suma kvadrata svih reziduala upravo je minimalna postignuta vrijednost za D, tj. $D(\widehat{\alpha}, \widehat{\beta})$, i predstavlja jednu mjeru kvalitete modela koju označavamo SSE (sum of squares of errors):

$$SSE = \sum_{i=1}^{n} e_i^2.$$

- primjer 6.18

Statističko zaljučivanje

- da bismo mogli koristiti dobiveni model potrebno je napraviti analizu prihvatljivosti modela
- istražujemo jesu li ispunjene osnovne pretpostavke klasičnog regresijskog modela: greške modela trebaju biti međusobno nezavisne i jednako distribuirane slučajne varijable s normalnom distribucijom
- dio analize modela koji se provodi u tu svrhu obično se naziva **analiza reziduala**.

Analiza reziduala

Jednakost varijanci grešaka

- imaju li $\varepsilon_1, \ldots \varepsilon_n$ jednake varijance? (homogenost varijanci grešaka)
- zaključujemo na temelju procjena grešaka modela – reziduala $e_1, \ldots, e_n$
- grafički prikažemo reziduale u ovisnosti o predikcijama – dijagram raspršenosti za točke $(\widehat{y}_i, e_i)$, $i = 1, \ldots, n$
- ako u tom dijagramu uočavamo sustavno povećanje ili smanjenje raspršenosti, to je znak da varijance nisu homogene

Analiza reziduala

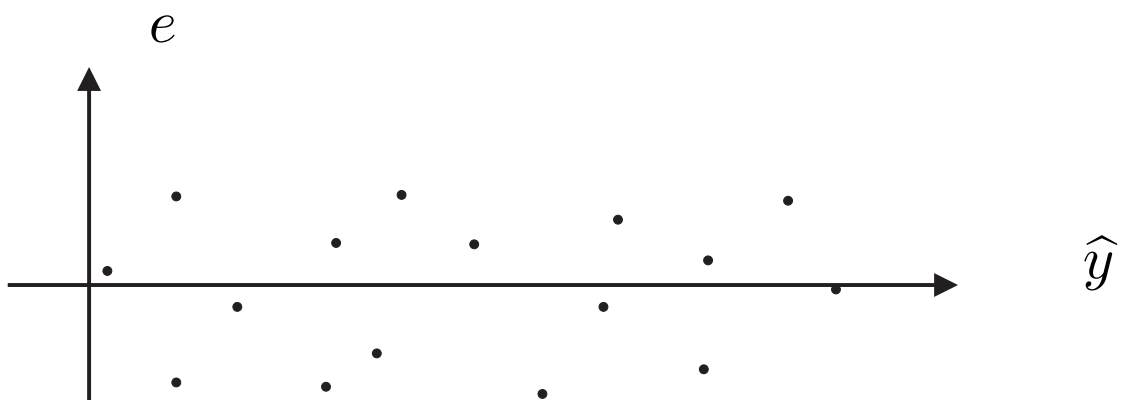

Figure 1: Parovi $(\widehat{y}_i, e_i)$ koji sugeriraju homogenost varijanci reziduala.

Analiza reziduala

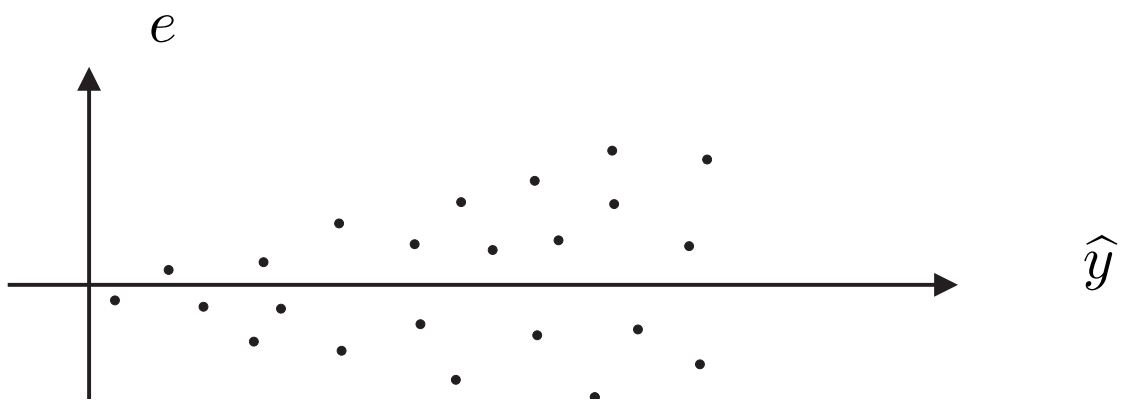

Figure 2: Ovakav raspored parova $(\widehat{y}_i, e_i)$ sugerira stalan rast varijance, dakle varijance nisu homogene.

Analiza reziduala

Nezavisnost grešaka

- jesu li $\varepsilon_1, \ldots \varepsilon_n$ nezavisne?
- zavisnost grešaka može se manifestirati na razne načine
- koristit ćemo dvije grafičke metode:
  - dijagram raspršenosti reziduala u odnosu na vrijednosti nezavisne varijable
  - dijagram raspršenosti parova susjednih reziduala, tj. parova (e_i, e_{i-1}), $i = 2, \ldots, n$
- ako nema nikakve pravilnosti u izgledu dijagrama, nemamo razloga sumnjati u nezavisnost

Analiza reziduala

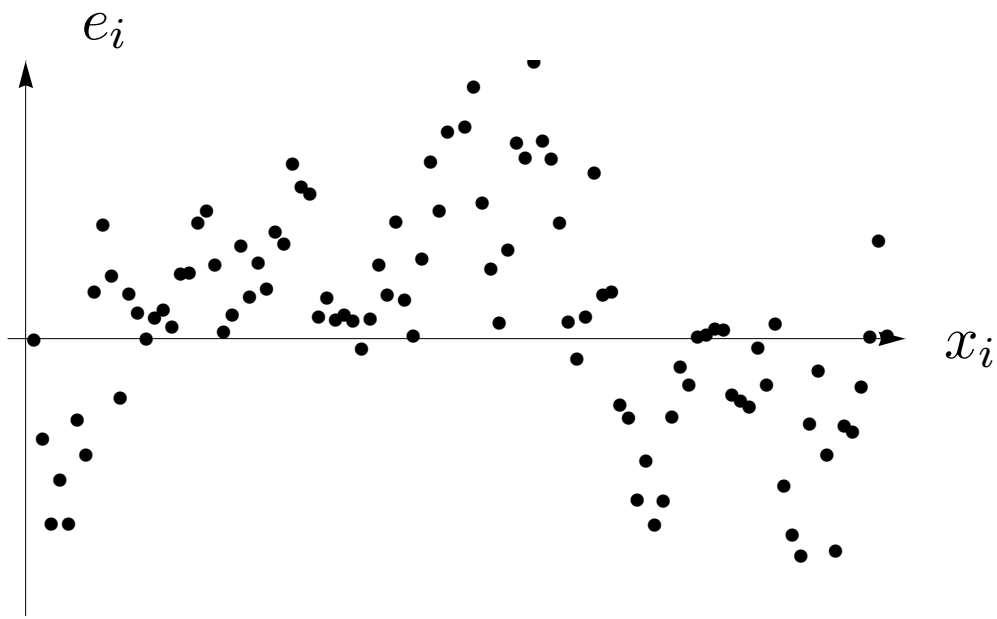

Figure 3: Ovakav raspored parova (x_i, e_i) sugerira međusobnu zavisnost grešaka modela.

Analiza reziduala

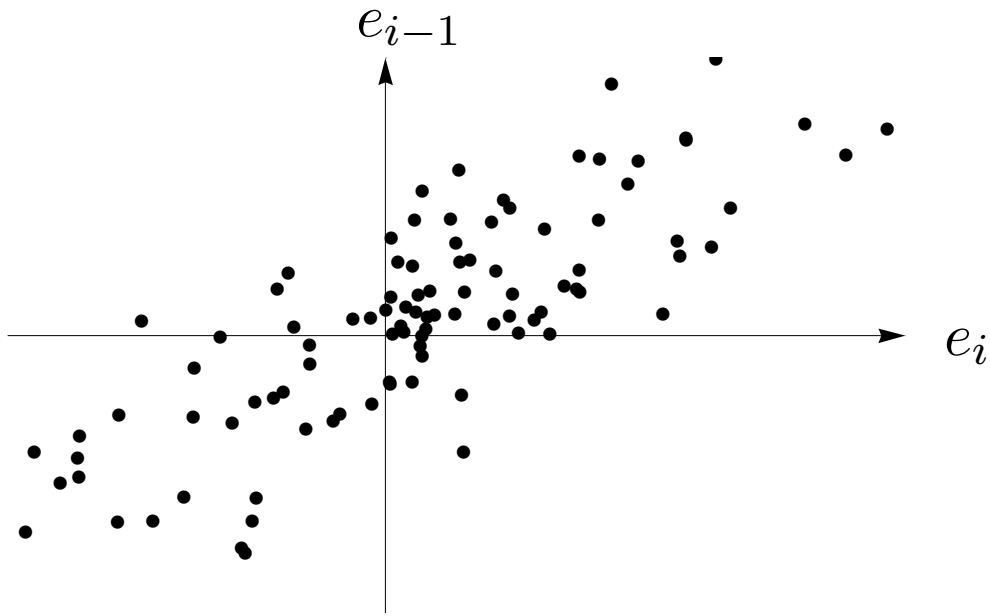

Figure 4: Ovakav raspored parova (e_i, e_{i-1}) sugerira međusobnu zavisnost grešaka modela.

Analiza reziduala

Normalna distribuiranost grešaka

- jesu li $\varepsilon_1, \ldots \varepsilon_n$ normalno distribuirane?
- možemo provjeriti KS testom i Shapiro-Wilk testom na rezidualima $e_1, \ldots, e_n$
- nije nužan uvjet, ali ukoliko ne vrijedi treba biti oprezan u statističkom zaključivanju o modelu

Ako nemamo razloga sumnjati u ispravnost pretpostavki modela, možemo ga koristiti za zaključivanje o vezi između nezavisne i zavisne varijable.

Zaključivanje o koeficijentu smjera regresijskog pravca

- je li model $Y_i = \alpha + \beta x_i + \varepsilon_i$ bolji od nul-modela $Y_i = \alpha + \varepsilon_i$ (modela u kojemu je $\beta = 0$)?
- koji od dva modela bolje opisuje promjene u očekivanju slučajnih varijabli Y_i u ovisnosti o vrijednostima x_i?
- ako je $\beta = 0$, takav regresijski pravac bio bi paralelan s x-osi pa promjena vrijednosti nezavisne varijable ne bi rezultirala promjenom očekivanja zavisne varijable
- to možemo utvrditi statističkim testom čije su hipoteze

$$H_0 : \beta = 0,$$
$$H_1 : \beta \neq 0.$$

Zaključivanje o koeficijentu smjera regresijskog pravca

- test se temelji na test-statistici čiju vrijednost $\hat{t}$ za eksperimentalne vrijednosti x_i i y_i računamo formulom

$$\hat{t} = \frac{s_x \cdot \widehat{\beta}}{s} \sqrt{n-1},$$

gdje je

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2}, \quad s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}},$$

a $\widehat{\beta}$ procjena regresijskog koeficijenta β metodom najmanjih kvadrata

- ako je nul-hipoteza istinita, test-statistika ima Studentovu distribuciju s $(n-2)$ stupnja slobode

Zaključivanje o koeficijentu smjera regresijskog pravca

- na temelju realizacije $\hat{t}$ test statistike računamo pripadnu p-vrijednost na sljedeći način:

$$p = P(|T| \geq |\hat{t}|)$$

  gdje je T slučajna varijabla koja ima Studentovu distribuciju s $(n-2)$ stupnja slobode
- tako izračunatu p-vrijednost uspoređujemo s razinom značajnosti α i donosimo odluku kako slijedi:
  - ako je $p \leq \alpha$, odbacujemo nul-hipotezu i na razini značajnosti α prihvaćamo alternativnu hipotezu, tj. podaci potvrđuju da se promjene u vrijednosti nezavisne varijable odražavaju na promjene u očekivanju zavisne varijable na razini značajnosti α (model ima smisla)
  - ako je $p > \alpha$, nemamo dovoljno argumenata tvrditi da se promjene u vrijednosti nezavisne varijable odražavaju na promjene u očekivanju zavisne varijable na razini značajnosti α

Dio varijabilnosti objašnjen modelom

- koliki je dio promjena u eksperimentalnim vrijednostima zavisne varijable objašnjen dobivenim modelom?
- **koeficijent determinacije** — R^2 definiran je izrazom

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}, \quad R^2 \in [0, 1].$$

- koeficijent determinacije R^2 – u kolikoj mjeri je rasipanje eksperimentalnih vrijednosti zavisne varijable objašnjeno linearnom funkcijom $x \mapsto \alpha + \beta x$, a u kolikoj se mjeri radi o tzv. rezidualnom ili neobjašnjenom rasipanju (tu informaciju očitavamo iz broja $(1 - R^2)$)

Dio varijabilnosti objašnjen modelom

- velika vrijednost koeficijenta determinacije (R^2 blizu 1) ukazuje na to da linearan model objašnjava velik dio raspršenosti u eksperimentalnim vrijednostima zavisne varijable (samo mali dio je ostao neobjašnjen modelom i treba ga pripisati slučajnoj grešci)
- modeli kod kojih je R^2 mali nisu informativni za opis varijable Y korištenjem vrijednosti nezavisne varijable x jer opisuju samo mali dio varijablnosti u podacima iz Y, dok je veliki dio ostao neobjašnjen modelom
- primjeri 6.19, 6.20

Mjere korelacije i asocijacije

- korelacija i asocijacija predstavljaju veze između varijabli koje ih čine zavisnima
- značajnost i jakost tih veza možemo iskazati različitim mjerama

Koeficijent korelacije

- neka je (X, Y) dvodimenzionalan slučajni vektor kojemu svaka komponenta ima varijancu
- X i Y ne moraju nužno biti diskretne slučajne varijable
- **Koeficijent korelacije** je broj definiran izrazom:

$$\rho = \frac{E(X-\mu)(Y-\nu)}{\sigma_X \sigma_Y},$$

gdje su

$$\mu = EX, \quad \nu = EY, \quad \sigma_X = \sqrt{Var\, X}, \quad \sigma_Y = \sqrt{Var\, Y}.$$

Koeficijent korelacije

Za koeficijent korelacije vrijede sljedeće činjenice:

- $\rho \in [-1, 1]$
- ako su X i Y nezavisne slučajne varijable tada je $\rho = 0$
- $Y = aX + b$, gdje je $a > 0$, onda i samo onda ako je $\rho = 1$
- $Y = aX + b$, gdje je $a < 0$, onda i samo onda ako je $\rho = -1$.

Ako je $\rho = 0$, kažemo da su slučajne varijable X i Y **nekorelirane**.

- ako su X i Y nezavisne tada su i nekorelirane – $\rho = 0$
- ako je $\rho \neq 0$, tada su X i Y zavisne

Općenito, obrat ne vrijedi:

- ako su nekorelirane, ne moraju biti i nezavisne

Koeficijent korelacije

- zavisnost između slučajnih varijabli X i Y možemo potvrditi ako pokažemo da je njihov koeficijent korelacije različit od 0
- ako je koeficijent korelacije 1 ili -1, onda znamo i tip veze između X i Y – linearna veza
- za procjenu koeficijenta korelacije postoji nekoliko procjenitelja
- ovdje ćemo koristiti Pearsonov koeficijent korelacije (koristi se za neprekidne slučajne varijable)
- ako su $(x_1, y_1), \ldots, (x_n, y_n)$ parovi nezavisnih realizacija slučajnog vektora (X, Y), onda se iznos Pearsonova korelacijskog koeficijenta računa pomoću izraza

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y}_n)^2}}.$$

Koeficijent korelacije

- uočimo: r^2 je jednak koeficijentu determinacije koji mjeri jakost linearne veze između varijabli u linearnom regresijskom modelu
- da bismo korištenjem procjene koeficijenta korelacije potvrdili zavisnost slučajnih varijabli, potrebno je odbaciti statističku hipotezu

$$H_0 : \rho = 0$$

nasuprot

$$H_1 : \rho \neq 0$$

- pretpostavljamo normalnost distribucije slučajnog vektora (X, Y)
- za testiranje navedene nul-hipoteze računamo vrijednost test statistike po formuli:

$$\hat{t} = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}}.$$

- ako je nul-hipoteza istinita, test statistika ima Studentovu distribuciju s $(n-2)$ stupnjeva slobode

Koeficijent korelacije

- označimo li s T slučajnu varijablu koja ima Studentovu distribuciju s $(n-2)$ stupnjeva slobode, pripadnu p-vrijednost određujemo na uobičajeni način:
  - $p = P(|T| \geq |\hat{t}|)$ za alternativnu hipotezu oblika $H_1 : \rho_{XY} \neq 0$
- tako izračunatu p-vrijednost uspoređujemo s razinom značajnosti α i donosimo odluku:
  - ako je $p \leq \alpha$, odbacujemo nul-hipotezu i na razini značajnosti α prihvaćamo alternativnu hipotezu, tj. kažemo da su slučajne varijable X i Y korelirane (pa onda i zavisne)
  - ako je $p > \alpha$, nemamo razloga odbaciti nul-hipotezu, tj. kažemo da nemamo dovoljno argumenata tvrditi da su X i Y korelirane (ne znači da su nezavisne)
- primjer 6.21, automobili.sta

Spearmanov koeficijent korelacije ranga

- neka je (X, Y) dvodimenzionalan slučajni vektor
- **Spearmanov koeficijent korelacije ranga** ρ_S pokazuje u kojoj mjeri se veza između dvije varijable X i Y može opisati monotonom funkcijom

Vrijedi sljedeće:

- $\rho_S \in [-1, 1]$
- ako $\rho_S = 0$ ne postoji monotona veza između X i Y
- ako $\rho_S > 0$ veza između X i Y je rastuća
- ako $\rho_S < 0$ veza između X i Y je padajuća

Spearmanov koeficijent korelacije ranga

Za uzorka $(x_1, y_1), \ldots, (x_n, y_n)$ nezavisnih realizacija slučajnog vektora (X, Y) Spearmanov koeficijent može se procijeniti na sljedeći način:

- sortiramo niz $x_1, \ldots, x_n$ i niz $y_1, \ldots, y_n$ u rastućem poretku
- odredimo r_{x_i} redni broj (**rang**) podatka x_i u sortiranom nizu
- odredimo r_{y_i} redni broj (rang) podatka y_i u sortiranom nizu
- procjena za Spearmanov koeficijent je tada

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} (r_{x_i} - r_{y_i})^2}{n(n^2 - 1)}$$

Spearmanov koeficijent korelacije ranga

- ako želimo potvrditi postojanje monotone veze, potrebno je odbaciti hipotezu

$$H_0 : \rho_S = 0$$

nasuprot

$$H_1 : \rho_S \neq 0$$

- o tome je li veza rastuća ili padajuća zaključujemo na osnovu predznaka r_S
- za testiranje navedenih hipoteza računamo vrijednost test statistike po formuli:

$$\hat{t} = \frac{\sqrt{n-2}\, r_S}{\sqrt{1 - r_S^2}}.$$

- ako je nul-hipoteza istinita, test statistika asimptotski ima Studentovu distribuciju s $(n-2)$ stupnjeva slobode

Spearmanov koeficijent korelacije ranga

- označimo li s T slučajnu varijablu koja ima Studentovu distribuciju s $(n-2)$ stupnjeva slobode, pripadnu p-vrijednost određujemo na uobičajeni način:
  - $p = P(|T| \geq |\hat{t}|)$ za alternativnu hipotezu oblika $H_1 : \rho_S \neq 0$
- tako izračunatu p-vrijednost uspoređujemo s razinom značajnosti α i donosimo odluku:
  - ako je $p \leq \alpha$, odbacujemo nul-hipotezu i na razini značajnosti α prihvaćamo alternativnu hipotezu, tj. kažemo da su postoji monotona veza između slučajnih varijabli X i Y (posebno, i zavisne su)
  - ako je $p > \alpha$, nemamo razloga odbaciti nul-hipotezu, tj. kažemo da nemamo dovoljno argumenata tvrditi da između X i Y postoji monotona veza
- primjer 6.21, automobili.sta

↭