LEKCIJA. Korelācija un lineārā regresija

Jānis Valeinis

FMOF

14/04/2021

Nodarbības tēma

1. Korelācija un lineārā regresija;
2. Daudzfaktoru regresija un citi modeļi.

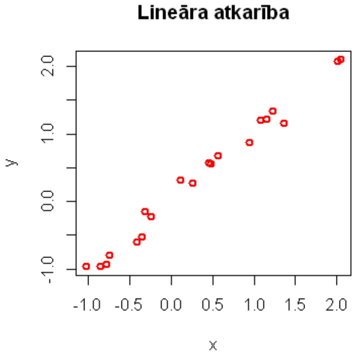
Korelāciju un regresiju analīzes mērķi un apraksts

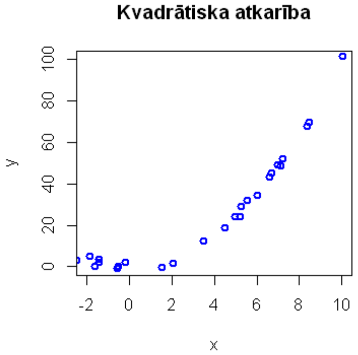
- Dotiem gadījuma lielumiem X un Y, analizēt to saistību jeb atkarību (*asociation*, *dependence*), korelāciju (*correlation*);
- Prognozēt Y, izmantojot X: regresiju analīze;
- Parasti doti divdimensiju novērojumi (x_1, y_1), (x_2, y_2),…, (x_n, y_n), kas iegūti vienlaicīgi novērojot divus gadījuma lielumus X un Y;
- Daudzdimensiju regresiju analīze (*multiple regression*) tiek lietota, ja Y atkarīgs no daudziem faktoriem X_1, X_2, …, X_p, kur p - interesējošo faktoru skaits.
- Vārdu **korelācija** parasti matemātiskajā statistikā lieto tieši saistībā ar *lineāru saistību jeb atkarību*!

Pozitīva saistība

Pozitīva saistība (asociācija, atkarība)

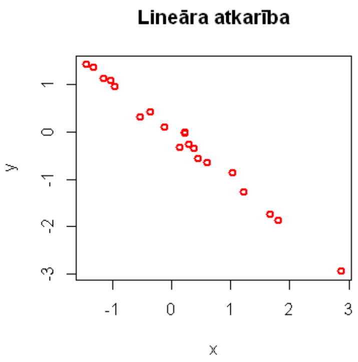
Divi mainīgie ir pozitīvi saistīti, ja viena mainīgā Y lielas vērtības parādās saistībā ar lielām otra mainīgā X vērtībām. Tāpat mazas Y vērtības tiek novērotas vienlaicīgi ar mazām X vērtībām (Garums un svars parasti ir pozitīvi saistīti).





Negatīva saistība (asociācija, atkarība)

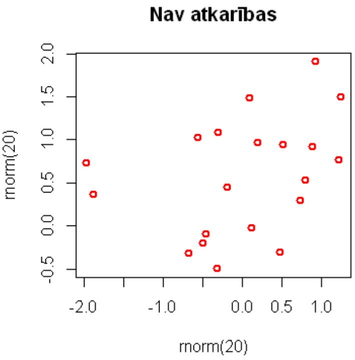
Divi mainīgie ir negatīvi saistīti, ja viena mainīgā Y lielas vērtības parādās saistībā ar mazām otra mainīgā X vērtībām. Tāpat mazas Y vērtības tiek novērotas vienlaicīgi ar lielām X vērtībām (Augsts pieprasījums parasti parādās pie zemām cenām).

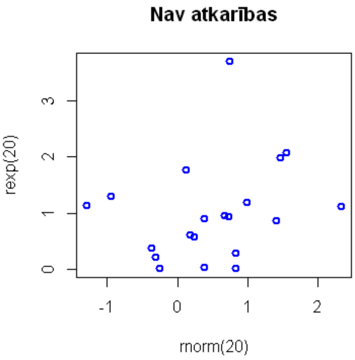


Nav saistības (asociācijas, atkarības)

Nav saistības jeb asociācijas

Ja divi mainīgie nav saistīti, tad parasti izkliedes grafikā nav redzamas nekādas sakarības (*no pattern*). Piemēram, svars un ienākums nav atkarīgi viens no otra.





Pīrsona korelācijas koeficients

Pīrsona korelācijas koeficients

Ja X un Y ir divi gadījuma lielumi, tad Pīrsona korelācijas koeficients

$$\rho_{XY} = \frac{cov(X, Y)}{\sqrt{DX}\sqrt{DY}},$$

kur $cov(X, Y) = E(X - EX)(Y - EY)$ sauc par kovariāciju, DX un DY ir X un Y dispersijas.

Pīrsona korelācijas koeficienta īpašības

- $-1 \leq \rho_{XY} \leq 1$;
- Ja X un Y neatkarīgi, tad $\rho_{XY} = 0$;
- Ja $\rho_{XY} = 1$, tad $P(Y = aX + b) = 1$, kur $a > 0$ (pastāv perfekta pozitīva korelācija jeb lineāra atkarība). Savukārt, ja $\rho_{XY} = -1$, tad $P(Y = aX + b) = 1$, kur $a < 0$.

Pīrsona izlases korelācijas koeficients

Pīrsona izlases korelācijas koeficients

Dotiem novērojumiem (x_1, y_1), (x_2, y_2),..., (x_n, y_n) Pīrsona izlases korelācijas koeficients definēts

$$\hat{\rho}_{XY} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}},$$

kur

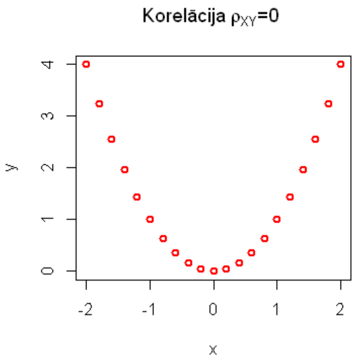
$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

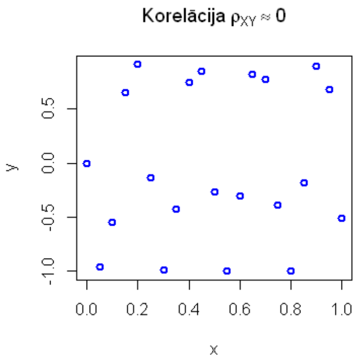
$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \; S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Pīrsona korelācijas koeficients kā lineārs mērs

Uzmanību!

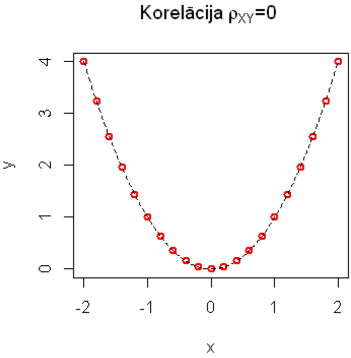
Pīrsona korelācijas koeficients ir lineārs saistības jeb atkarības mērs!

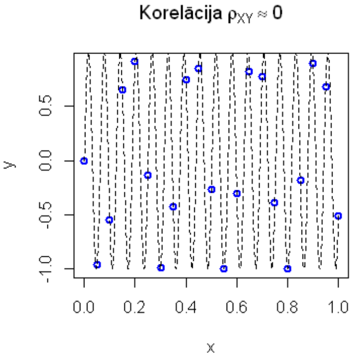




Uzmanību!

Pīrsona korelācijas koeficients ir lineārs saistības jeb atkarības mērs!





No korelācijas neseko cēloņsakarība (*causality*)

1.Piemērs

Novērots, ka gulēšana ar kurpēm stipri korelē ar pamošanos ar galvassāpēm. *Secinājums*: gulēšana ar kurpēm izraisa galvassāpes.

No korelācijas neseko cēloņsakarība (*causality*)

1.Piemērs

Novērots, ka gulēšana ar kurpēm stipri korelē ar pamošanos ar galvassāpēm. *Secinājums*: gulēšana ar kurpēm izraisa galvassāpes.

Atbilde

Secinājums ir nepareizs. Ir citi mainīgie, kas ietekmē dotos mainīgos, kurus mēs bieži nevaram apjaust. Piemēram, gulēšana ar kurpēm varētu notikt cita mainīgā: alkohola lietošanas rezultātā, kas arī varētu izsaukt galvassāpes!

No korelācijas neseko cēloņsakarība (*causality*)

2.Piemērs

Novērots, ka saldējuma pārdošanas pieaugums stipri un pozitīvi korelē ar noslīkšanas gadījumu straujų pieaugumu. *Secinājums*: saldējuma patēriņš izraisa noslīkšanu (vai otrādāk!).

No korelācijas neseko cēloņsakarība (*causality*)

2.Piemērs

Novērots, ka saldējuma pārdošanas pieaugums stipri un pozitīvi korelē ar noslīkšanas gadījumu strauju pieaugumu. *Secinājums*: saldējuma patēriņš izraisa noslīkšanu (vai otrādāk!).

Atbilde

Šajā gadījumā arī secinājums ir nepareizs. Karstā laikā cilvēki ēd saldējumu un arī daudz vairāk peldas nekā aukstākā laikā. Tāpēc arī slīcēju skaits palielinās.

No korelācijas neseko cēloņsakarība (*causality*)

3.Piemērs

Aplūkojot cilvēkus uz ielas, novērots, ka lietus līšana stipri pozitīvi korelē ar lietussarga paņemšanu līdzi no mājām. *Secinājums*: lietussarga paņemšana izraisa lietus līšanu!

No korelācijas neseko cēloņsakarība (*causality*)

3.Piemērs

Aplūkojot cilvēkus uz ielas, novērots, ka lietus līšana stipri pozitīvi korelē ar lietussarga paņemšanu līdzi no mājām. *Secinājums*: lietussarga paņemšana izraisa lietus līšanu!

Atbilde

Arī šajā gadījumā arī secinājums ir nepareizs. Cilvēkiem bija lietussargi, jo viņi jau no laika ziņām zināja, ka būs lietus. Kādu eksperimentu jāveic, lai izdarītu pareizus secinājumus?

No korelācijas neseko cēloņsakarība (*causality*)

3.Piemērs

Aplūkojot cilvēkus uz ielas, novērots, ka lietus līšana stipri pozitīvi korelē ar lietussarga paņemšanu līdzi no mājām. *Secinājums*: lietussarga paņemšana izraisa lietus līšanu!

Atbilde

Arī šajā gadījumā arī secinājums ir nepareizs. Cilvēkiem bija lietussargi, jo viņi jau no laika ziņām zināja, ka būs lietus. Kādu eksperimentu jāveic, lai izdarītu pareizus secinājumus?

Atbilde

Šajā gadījumā pareizi būtu veikt eksperimentu: izmetam monētu, vai ņemsim lietussargu vai nē, tad pētām interesējošo sakarību!

Eksperimenta veikšana vai novērojumu analīze?

Experiment

An **experiment** is a study in which, when collecting the data, the researcher controls the values of the predictor variables.

Observational study

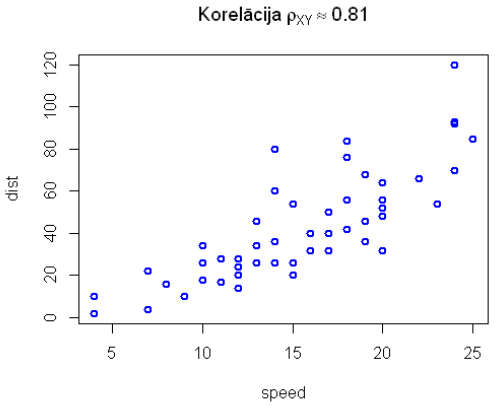
An **observational study** is a study in which, when collecting the data, the researcher merely observes and records the values of the predictor variables as they happen.

Eksperimenta veikšana vai novērojumu analīze?

- The primary advantage of conducting experiments is that one can typically conclude that differences in the predictor values is what caused the changes in the response values. This is not the case for observational studies.
- Unfortunately, most data used in regression analyses arise from observational studies. Therefore, you should be careful not to overstate your conclusions, as well as be cognizant that others may be overstating their conclusions.

Piemērs iebūvēto datu masīvam `cars`

Datu apraksts: Iebūvētie dati `cars` satur novērojumus par mašīnu ātumiem (mainīgais `speed`) un bremzēšanas distancēm (mainīgais `dist`).



Secinājums: pastāv acīmredzama, cieša lineāra saistība jeb atkarība starp mašīnu ātrumiem un bremzēšanas distancēm.

Piemērs iebūvēto datu masīvam `mtcars`

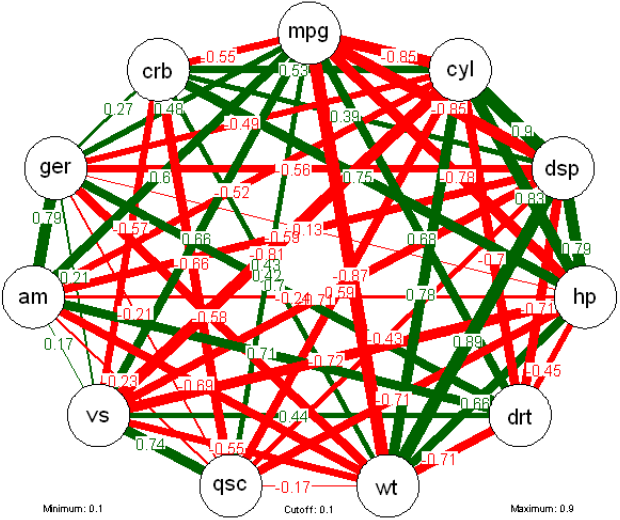
Datu apraksts: Iebūvētie dati `mtcars` satur novērojumus par 32 mašīnu dažādiem raksturlielumiem (1973-74 modeļi)

1	mpg	Miles/(US) gallon
2	cyl	Number of cylinders
3	disp	Displacement (cu.in.)
4	hp	Gross horsepower
5	drat	Rear axle ratio
6	wt	Weight (1000 lbs)
7	qsec	1/4 mile time
8	vs	V/S
9	am	Transmission (0 = automatic, 1 = manual)
10	gear	Number of forward gears
11	carb	Number of carburetors

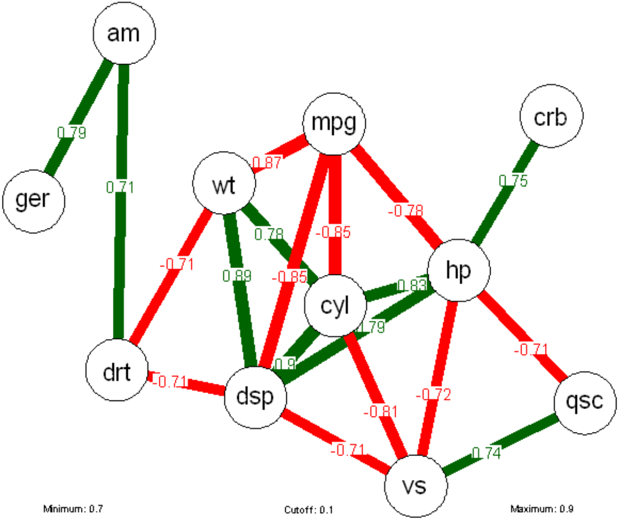
Piemērs iebūvēto datu masīvam `mtcars`

Datu apraksts: Iebūvētie dati `mtcars` satur novērojumus par 32 mašīnu dažādiem raksturlielumiem (1973-74 modeļi). Pirmie 17 novērojumi:

```
> mtcars
                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4          21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag      21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710         22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive     21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout  18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant            18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360         14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D          24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230           22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280           19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C          17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE         16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL         17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC        15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4  8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial  14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
```


mpg
crb
cyl
ger
dsp
am
hp
vs
drt
qsc
wt
-0.55
-0.85
0.53
0.27
0.48
-0.49
0.39
-0.85
0.9
0.6
-0.56
0.75
-0.78
-0.52
0.83
0.79
-0.57
0.13
0.68
0.79
0.66
-0.81
-0.66
0.43
0.42
-0.87
-0.7
0.21
0.78
-0.59
-0.24
-0.58
-0.71
-0.21
-0.43
0.89
0.17
0.71
-0.72
-0.45
-0.23
-0.69
-0.71
0.66
0.44
0.74
-0.55
-0.71
-0.17
Minimum: 0.1
Cutoff: 0.1
Maximum: 0.9

Korelācijas datu masīvam `mtcars`



Korelāciju statistiskā nozīmīguma tests

Statistiskais tests

Hipotēžu pārbaudē

$$H_0 : \rho_{XY} = 0 \text{ pret } H_1 : \rho_{XY} \neq 0$$

lieto statistiku

$$t = \frac{\rho_{XY}\sqrt{n-2}}{\sqrt{1-\rho_{XY}^2}} \rightarrow_d t_{n-2}$$

kurai ir t_{n-2} robežsadalījums, ja nulles hipotēze ir spēkā.

```
> cor.test(speed,dist)

        Pearson's product-moment correlation

data:  speed and dist
t = 9.464, df = 48, p-value = 1.49e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6816422 0.8862036
sample estimates:
      cor
0.8068949
```

Korelāciju statistiskā nozīmīguma tests: komentāri

- Ja p-vērtība < 0.05, tad saka, ka korelācijas koeficients ir statistiski nozīmīgs (pie nozīmības līmeņa $\alpha = 0.05$);
- Korelācijas koeficients var nebūt nozīmīgs, bet var viņa vērtība var būt liela un otrādāk!
- Ja $(1-\alpha)100\%$ ticamības intervāli nesatur 0, tad korelācijas koeficients statistiski nozīmīgs.

Citi saistību jeb atkarību mēri: Spīrmena koeficients

Spīrmena rangu korelācijas koeficients

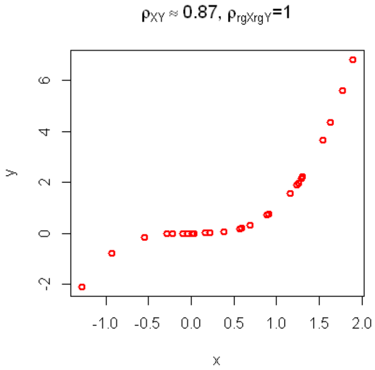
Ja X un Y ir divi gadījuma lielumi, tad Spīrmena korelācijas koeficients

$$\rho_{rg_X rg_Y} = \frac{cov(rg_X, rg_Y)}{\sqrt{Drg_X}\sqrt{Drg_Y}},$$

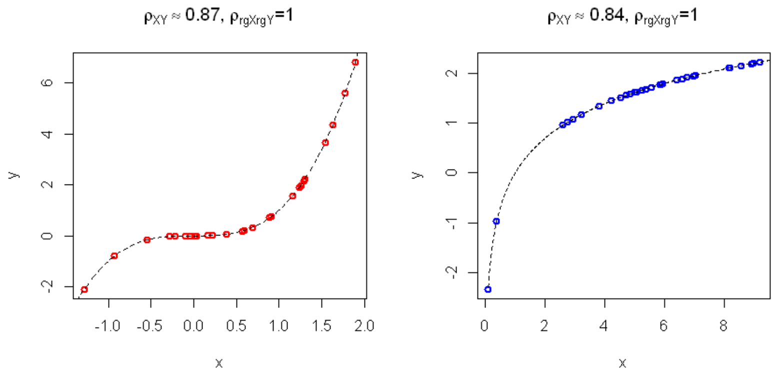
kur rg_X un rg_Y apzīmē X un Y rangus.

- Ja X un Y saista kāda monotona funkcija (augoša vai dilstoša), tad $\rho_{rg_X rg_Y}$ ir vai nu 1 vai nu -1 (nav obligāti lineāra!);
- Šis koeficients pēta korelāciju starp rangiem, tāpēc mazāk jūtīgs pret izlecējiem (salīdzinot ar Pīrsona korelācijas koeficientu)!
- Ja datu mākonis ir *eliptisks*, abi koeficienti dod līdzīgus rezultātus;
- Koeficients ir *robusts* un *neparametrisks* (pārrunāt klasē)!

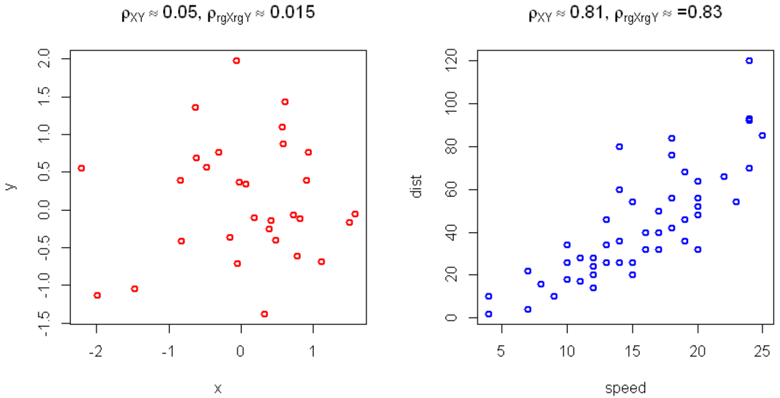
Ja X un Y saista kāda augoša funkcija, tad $\rho_{rg_X rg_Y} = 1$. Šajā gadījumā parasti Pīrsona korelācijas koeficients arī būs liels, tomēr atšķirīgs no Spīrmena koeficienta.



Ja X un Y saista kāda augoša funkcija, tad $\rho_{rg_X rg_Y} = 1$. Šajā gadījumā parasti Pīrsona korelācijas koeficients arī būs liels, tomēr atšķirīgs no Spīrmena koeficienta.



Ja X un Y ir eliptiski datu mākoņi, tad koeficientu vērtības daudz neatšķirsies! (ja datos nav daudz izlecēju).



Citi saistību jeb atkarību mēri: Kendala korelācijas koeficients

Kendala rangu korelācijas koeficients

Dotiem novērojumiem (x_1, y_1), (x_2, y_2),..., (x_n, y_n) Kendala rangu korelācijas koeficients

$$\tau = \frac{\text{(konkordu pāru skatis)} - \text{(diskonkordu pāru skaits)}}{n(n-1)/2}.$$

Konkords vai diskonkords novērojumu pāris (x_i, y_i) un (x_j, y_j)

Novērojumu pāri (x_i, y_i) un (x_j, y_j), kur $i \neq j$ sauc par *konkordu*, ja rangi abiem elementiem uzvedas līdzīgi: ja $x_i > x_j$, tad $y_i > y_j$ un otrādi. Citādi to sauc par *diskonkordu* novērojumu pāri.

Piezīme. Var lietot ordināliem (tas ir sakārtotiem kategorāliem) novērojumiem!

Doti gadījumu lielumu pāri (X_1, Y_1),..., (X_n, Y_n) un to novērojumi (x_1, y_1), (x_2, y_2),..., (x_n, y_n).

Vienkāršā lineārā regresija

Vienkāršās lineārās regresijas modelis

$$Y_i = a + bx_i + \epsilon_i,$$

kur $a \in \mathbb{R}, b \in \mathbb{R}$ ir regresijas koeficienti, ϵ_i ir gadījuma lielums ar $E(\epsilon_i) = 0$, kurš raksturo kļūdu (rezidiji, atlikumi).

Vienkāršā lineārā regresija

- Regresija raksturo vidējo Y uzvedību pie fiksētas $X = x$ vērtības;
- Y - atkarīgais jeb atbildes mainīgais, X - neatkarīgais jeb skaidrojošais mainīgais (prediktors);
- Vispārējā formā regresija tiek uzdota kā nosacītā matemātiskā cerība $E(Y|X = x)$. Vienkāršās lineārās regresijas gadījumā

$$E(Y|X = x) = a + bx.$$

Vienkāršā lineārā regresija: mērķi

1. Novērtēt parametrus a un b un pārbaudīt to statistisko nozīmīgumu (tad noraidām hipotēzi $H_0 : b = 0$)!
2. Aprēķināt, cik liela variācija tiek izskaidrota ar vienkāršās lineārās regresijas palīdzību (determinācijs koeficients);
3. Noskaidrot, vai regresija ir statistiski nozīmīga (ANOVA F-tests);
4. Veikt prognozi Y citām X interesējošām vērtībām;
5. Ticamības joslas regresijas taisnei.

Vienkāršā lineārā regresija: nosacījumi

1. $\epsilon_1, \ldots, \epsilon_n$ ir neatkarīgi;
2. $\epsilon_i \sim N(0, \sigma^2_{\epsilon_i})$ visiem $i = 1, \ldots, n$;
3. $\sigma^2_{\epsilon_i} = \sigma^2$, ko sauc par *homoscedasticity* (dažkārt par homogenitāti jeb dispersiju viendabīgumu).

Piezīme. Pirmais nosacījums par atlikumu neatkarību parasti tiek pārbaudīts jebkuram modelim. Kļūdu haotiskums (kad neveidojas iekšēja atkarība) liecina par modeļa pielāgošanos datiem.

Otrais un trešais nosacījums par normalitāti un homogenitāti nepieciešami dažādiem statistiskiem testiem (piemēram, par koeficientu nozīmību) kā arī mazāko kvadrātu metodes novērtējumu BLUE (*Best linear unbiased estimators*) īpašības pamatojumam.

Vienkāršā lineārā regresija: parametru novērtēšana

Mazāko kvadrātu metode

Ideja: atrast tādus novērtējumus $\hat{a}$ un $\hat{b}$, kas minimizē izteiksmi

$$\sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

MKM novērtejumi

$$\hat{a} = \bar{y} - \hat{b}\bar{x},$$

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

Vienkāršā lineārā regresija: novērtējumu kvalitāte

Atlikumu kvadrātu summa RSS un atlikumi

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2,$$

kur $\hat{\epsilon}$ apzīmē atlikumus (rezidijus).

Prognoze $\hat{y}$

$$\hat{y} = \hat{a} + \hat{b}x.$$

Apgalvojums

$\hat{a}$ un $\hat{b}$ ir BLUE (*Best Linear Unbiased Estimators*), tas ir, nenovirzīti novērtējumi ar mazāko iespējamo dispersiju!

Determinācijas koeficients R^2

$$R^2 = \frac{RegSS}{SST} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \rho_{XY}^2,$$

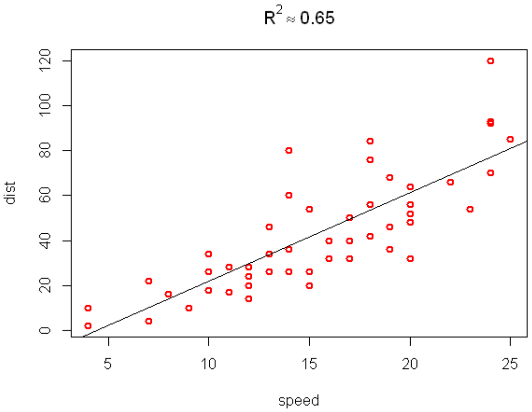
kur

$$SST = RegSS + RSS.$$

- **Interpretācija**. R^2 mēra kopējās variācijas (dispersijas) proporciju, kas tiek izskaidrota ar novērtētās lineārās taisnes palīdzību.
- **Īpašības**. $0 \leq R^2 \leq 1$. Ja $y_1, \ldots, y_n$ sakrīt ar taisnes vērtībām, tad $y_i = \hat{y}_i$ visiem $i = 1, \ldots, n$ un $R^2 = 1$. Savukārt, ja $y_1, \ldots, y_n$ ir tālu no taisnes, tad R^2 būs tuvu 0.

Determinācijas koeficients R^2

Iebūvēto datu piemērs `cars`: mainīgie `speed` un `dist` ir savā starpā pozitīvi korelēti ar $\hat{\rho}_{XY} = 0.80$.



Interpretācija: aptuveni 65% no kopējās dispersijas izskaidrots ar vienkāršās lineārās regresijas palīdzību!

Koeficientu un regresijas statistiskais nozīmīgums

Vai regresija ir statistiski nozīmīga?

- Hipotēžu pārbaude

$$H_0 : b = 0 \text{ pret} H_1 : b \neq 0$$

  Ja pie nozīmības līmeņa $\alpha = 0.05$ nulles hipotēze tiek noraidīta, tas nozīmē, ka koeficients ir statistiski nozīmīgs.
- Ekvivalents tests ir ANOVA F-tests!

Determinācijas koeficients R^2

Iebūvēto datu piemērs `cars`: mainīgie `speed` un `dist` ir savā starpā pozitīvi korelēti ar $\hat{\rho}_{XY} = 0.80$.

```
Call:
lm(formula = dist ~ speed)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Interpretācija: regresijas koeficienti nozīmīgi, arī pats F-tests norāda uz nozīmīgumu!

Citi prognorzējoši modeļi

- Daudzfaktoru lineārā regresija;
- Loģistiskā regresija;
- Vispārinātie lineārie modeļi un citas regresijas.
  `https://rpubs.com/jt_rpubs/279278`