

Curs 13

Cristian Niculescu

1 Intervale de încredere pentru media datelor nenormale

1.1 Scopurile învățării

1. Să poată deduce formula pentru intervale de încredere normale conservatoare pentru proporția θ din date Bernoulli.
2. Să poată calcula intervale de încredere 95% cu regula degetului mare pentru proporția θ a unei repartiții Bernoulli.
3. Să poată calcula intervale de încredere de selecție mare pentru media unei repartiții generale.

1.2 Introducere

Până acum, ne-am concentrat pe construirea intervalor de încredere pentru date provenite dintr-o repartiție normală. Acum vom învăța despre intervale de încredere pentru medie când datele nu sunt neapărat normale.

Întâi vom studia estimarea probabilității θ de succes când datele provin din o repartiție Bernoulli(θ) - reamintim că θ este de asemenea media repartiției Bernoulli.

Apoi vom considera cazul unei selecții mari dintr-o repartiție necunoscută; în acest caz putem apela la teorema limită centrală pentru a justifica intervalele de încredere z .

1.3 Datele Bernoulli și votarea

O utilizare obișnuită a intervalor de încredere este pentru estimarea proporției θ dintr-o repartiție Bernoulli(θ). De exemplu, presupunem că vrem să folosim un sondaj politic pentru a estima proporția din populație care susține candidatul A, sau echivalent, probabilitatea θ ca o persoană aleatoare să susțină candidatul A. În acest caz avem o regulă a degetului mare care ne permite să calculăm repede un interval de încredere.

1.3.1 Intervale de încredere normale conservatoare

Presupunem că avem datele i.i.d. x_1, x_2, \dots, x_n toate provenind dintr-o repartiție Bernoulli(θ). Atunci un interval de încredere $(1 - \alpha)$ **normal conservator** pentru θ este dat de

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{1}{2\sqrt{n}}. \quad (1)$$

Demonstrația dată mai jos folosește teorema limită centrală și observația că $\sigma = \sqrt{\theta(1 - \theta)} \leq 1/2$.

Veți vedea de asemenea în deducerea de mai jos că această formulă este conservatoare, dând un interval de încredere "cel puțin $(1 - \alpha)$ ".

Exemplul 1. Un sondaj întreabă 196 de persoane dacă preferă candidatul A candidatului B și află că 120 preferă A și 76 preferă B. Aflați intervalul de încredere normal conservator pentru θ , proporția din populație care preferă A.

Răspuns. Avem $\bar{x} = 120/196 \approx 0.612$, $\alpha = 0.05$ și $z_{0.025} \approx 1.96$. Formula spune că un interval de încredere 95% este

$$I \approx 0.612 \pm \frac{1.96}{2 \cdot 14} = 0.612 \pm 0.07 = [0.542, 0.682].$$

1.3.2 Demonstrația formulei (1)

Demonstrația formulei (1) se va baza pe următoarea leamnă.

Lemă. Deviația standard a unei repartiții Bernoulli(θ) este cel mult 0.5.

Demonstrația lemei. Notăm această deviație standard cu σ_θ pentru a sublinia dependența ei de θ . Atunci dispersia este $\sigma_\theta^2 = \theta(1 - \theta)$. Este ușor de văzut folosind analiza matematică sau proprietățile funcției de gradul 2 sau făcând graficul parabolei că maximul apare când $\theta = 1/2$. De aceea dispersia maximă este $1/4$, ceea ce implică faptul că deviația standard σ este cel mult $\sqrt{1/4} = 1/2$.

Demonstrația formulei (1). Demonstrația se bazează pe teorema limită centrală care spune că (pentru n mare) repartiția lui \bar{x} este aproximativ normală cu media θ și deviația standard σ_θ/\sqrt{n} . Pentru date normale avem intervalul de încredere $z(1 - \alpha)$

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma_\theta}{\sqrt{n}}.$$

Acum trucul este să înlocuim σ_θ cu $\frac{1}{2}$: deoarece $\sigma_\theta \leq \frac{1}{2}$, intervalul rezultat în jurul lui \bar{x}

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{1}{2\sqrt{n}}$$

este cel puțin la fel de lung ca intervalul folosind $\pm\sigma_\theta/\sqrt{n}$. Un interval mai lung este mai probabil să conțină adevărata valoare a lui θ , deci avem un interval de încredere $(1 - \alpha)$ ”conservator” pentru θ .

Din nou, numim acesta conservator deoarece $\frac{1}{2\sqrt{n}}$ supraestimează deviația standard a lui \bar{x} , rezultând un interval mai lung decât este necesar pentru a atinge un nivel de încredere de $(1 - \alpha)$.

1.3.3 Cum sunt raportate sondajele politice

Sondajele politice sunt adesea raportate ca o valoare cu o marjă de eroare. De exemplu puteți auzi

52% favorizează candidatul A cu o marjă de eroare de $\pm 5\%$.

Sensul precis al acestui fapt este

dacă θ este proporția din populație care sprijină pe A, atunci estimarea punctuală pentru θ este 52% și intervalul de încredere 95% este $52\% \pm 5\%$. Observați că reporterii sondajelor la știri nu menționează încrederea 95%. Doar va trebui să știi că asta este ceea ce fac sondajele.

Intervalul de încredere cu regula degetului mare 95%.

Reamintim că intervalul de încredere normal conservator $(1 - \alpha)$ este

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{1}{2\sqrt{n}}.$$

Dacă folosim aproximarea standard $z_{0.025} \approx 2$ (în loc de 1.96) obținem **intervalul de încredere 95% cu regula degetului mare** pentru θ :

$$\bar{x} \pm \frac{1}{\sqrt{n}}.$$

Exemplul 2. Votare. Presupunem că în curând vor fi alegeri locale între candidatul A și candidatul B. Presupunem că fracția din populația cu drept de vot care-l susține pe A este θ .

2 organizații de sondaj întreabă votanții pe cine preferă.

1. Firma *Rapid și primul* sondează 40 de votanți aleatori și află că 22 sprijină pe A.

2. Firma *Rapid, dar precaut* sondează 400 de votanți aleatori și află că 190 sprijină pe A.

Aflați estimările punctuale și intervalele de încredere cu regula degetului mare 95%. Explicați cum statisticile reflectă intuiția că sondajul a 400 de votanți este mai exact.

Răspuns. Pentru sondajul 1 avem

Estimare punctuală: $\bar{x} = 22/40 = 0.55$.

Interval de încredere: $\bar{x} \pm \frac{1}{\sqrt{n}} = 0.55 \pm \frac{1}{\sqrt{40}} = 0.55 \pm 0.16 = 55\% \pm 16\%$.

Pentru sondajul 2 avem

Estimare punctuală: $\bar{x} = 190/400 = 0.475$.

Interval de încredere: $\bar{x} \pm \frac{1}{\sqrt{n}} = 0.475 \pm \frac{1}{\sqrt{400}} = 0.475 \pm 0.05 = 47.5\% \pm 5\%$.

Precizia mai mare a sondajului cu 400 de votanți este reflectată de marginea mai mică a erorii, i.e. 5% pentru sondajul cu 400 de votanți vs. 16% pentru sondajul cu 40 de votanți.

Intervale de încredere pentru proporția binomială

Sunt multe metode de a produce intervale de încredere pentru proporția p a unei repartiții binomiale(n, p). Pentru un număr de abordări uzuale, vezi:

http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval.

1.4 Intervale de încredere de selecție mare

Un scop tipic în statistică este să estimăm media unei repartiții. Când datele au o repartiție normală am putea folosi intervale de încredere bazate pe statistici standardizate pentru a estima media.

Dar presupunem că datele x_1, x_2, \dots, x_n provin dintr-o repartiție cu pmf sau pdf $f(x)$ care poate să nu fie normală sau chiar parametrică. Dacă repartiția are medie și dispersie finite și dacă n este suficient de mare, atunci următoarea versiune a teoremei limită centrală ne arată că încă putem folosi o statistică standardizată.

Teorema limită centrală. Pentru n mare, repartiția de selecție a mediei Studentizate este aproximativ normală standard: $\frac{\bar{x} - \mu}{s/\sqrt{n}} \approx N(0, 1)$.

Deci, pentru n mare, intervalul de încredere $(1 - \alpha)$ pentru μ este aproximativ

$$\left[\bar{x} - \frac{s}{\sqrt{n}} \cdot z_{\alpha/2}, \bar{x} + \frac{s}{\sqrt{n}} \cdot z_{\alpha/2} \right],$$

unde $z_{\alpha/2}$ este valoarea critică $\alpha/2$ pentru $N(0, 1)$. Acesta este numit [intervalul de încredere de selecție mare](#).

Exemplul 3. Cât de mare trebuie să fie n ?

Reamintim că o eroare CI de tipul 1 apare când intervalul de încredere nu conține adevărata valoare a parametrului, în acest caz media. Să numim valoarea $(1 - \alpha)$ nivelul de încredere *nominal*. Spunem nominal deoarece dacă n nu este mare, nu ar trebui să ne așteptăm ca adevărata rată a erorii CI de tipul 1 să fie α .

Putem face simulări numerice pentru a aproxima adevăratul nivel de încredere. Ne așteptăm că, atunci când n devine mai mare, adevăratul nivel de încredere al intervalului de încredere de dispersie mare să converge la valoarea nominală.

Facem astfel de simulări pentru x provenite de la repartiția exponențială $\exp(1)$ (care este departe de normală). Pentru câteva valori ale lui n și nivelul de încredere nominal c facem 100000 de încercări. Fiecare încercare constă din următorii pași:

1. Extragem n date din $\exp(1)$.
2. Calculăm media de selecție \bar{x} și deviația standard de selecție s .
3. Construim intervalul de încredere c de selecție mare: $\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$.
4. Verificăm pentru o eroare CI de tipul 1, i.e. vedem dacă adevărata medie $\mu = 1$ nu este în interval.

Cu 100000 de încercări, nivelul de încredere empiric ar trebui să aproximeze bine adevăratul nivel.

Pentru comparație facem aceleași teste pe date provenite dintr-o repartiție normală standard. Iată rezultatele:

n	nominal conf. $1 - \alpha$	simulated conf.
20	0.95	0.905
20	0.90	0.856
20	0.80	0.762
50	0.95	0.930
50	0.90	0.879
50	0.80	0.784
100	0.95	0.938
100	0.90	0.889
100	0.80	0.792
400	0.95	0.947
400	0.90	0.897
400	0.80	0.798

Simulări pentru $\exp(1)$

n	nominal conf. $1 - \alpha$	simulated conf.
20	0.95	0.936
20	0.90	0.885
20	0.80	0.785
50	0.95	0.944
50	0.90	0.894
50	0.80	0.796
100	0.95	0.947
100	0.900	0.896
100	0.800	0.797
400	0.950	0.949
400	0.900	0.898
400	0.800	0.798

Simulări pentru $N(0, 1)$

Pentru repartiția $\exp(1)$ vedem că pentru $n = 20$ încrederea simulată a intervalului de încredere de selecție mare este mai mică decât încrederea nominală $(1 - \alpha)$. Dar pentru $n = 100$ încrederea simulată și încrederea nominală sunt foarte aproape. Deci, pentru $\exp(1)$, n undeva între 50 și 100 este destul de mare pentru cele mai multe scopuri.

Gândiți. Pentru $n = 20$ de ce încrederea simulată pentru repartiția $N(0, 1)$ este mai mică decât încrederea nominală?

Aceasta este deoarece am folosit $z_{\alpha/2}$ în loc de $t_{\alpha/2}$. Pentru n mare acestea sunt foarte aproape, dar pentru $n = 20$ există o diferență sesizabilă, de exemplu $z_{0.025} \approx 1.96$ și $t_{0.025} \approx 2.09$.

2 Intervale de încredere bootstrap

2.1 Scopurile învățării

1. Să poată construi și selecta din repartiția empirică a datelor.
2. Să poată explica principiul bootstrap.
3. Să poată proiecta și folosi un bootstrap empiric pentru a calcula intervale de încredere.
4. Să poată proiecta și folosi un bootstrap parametric pentru a calcula intervale de încredere.

2.2 Introducere

Bootstrap-ul empiric este o tehnică statistică popularizată de Bradley Efron în 1979. Cu toate că este remarcabil de simplu de aplicat, bootstrap-ul n-ar fi posibil fără puterea de calcul modernă. Ideea cheie este să facem calcule asupra datelor pentru a estima variația statisticilor care sunt calculate din aceleași date. "Bootstrap" = "curea de cizmă". (O căutare pe google a "by ones own bootstrap" vă va da o etimologie a acestei metafore.) Astfel de tehnici existau înainte de 1979, dar Efron a lărgit aplicabilitatea lor și a demonstrat modul de aplicare al bootstrap-ului folosind efectiv calculatoarele. El a inventat de asemenea termenul "bootstrap".

Aplicația noastră principală a bootstrap-ului va fi estimarea variației estimărilor punctuale; adică, estimarea intervalelor de încredere.

Exemplul 1. Presupunem că avem datele

$$x_1, x_2, \dots, x_n.$$

Dacă am ști că datele provin din $N(\mu, \sigma^2)$ cu medie necunoscută μ și dispersie cunoscută σ^2 , atunci am văzut că

$$\left[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right],$$

este un interval de încredere 95% pentru μ .

Acum presupunem că datele provin dintr-o repartiție complet necunoscută. Pentru a avea un nume, vom numi această repartiție F și media ei (necunoscută) μ . Putem folosi încă media de selecție \bar{x} ca o **estimare punctuală** a lui μ . Dar cum putem afla un interval de încredere pentru μ în jurul lui \bar{x} ? Răspunsul nostru va fi să folosim bootstrap-ul!

De fapt, bootstrap-ul tratează alte statistici la fel de ușor cum tratează media. De exemplu: mediana sau alte percentile. Acestea sunt statistici unde,

chiar pentru repartițiile normale, poate fi dificil să calculăm un interval de încredere doar din teorie.

2.3 Selectarea

În statistică, a **selecta** dintr-o mulțime înseamnă a alege elemente din acea mulțime. Într-o selecție aleatoare, elementele sunt alese aleator. Sunt 2 metode uzuale pentru selectare aleatoare.

Selectare fără înlocuire

Presupunem că tragem 10 cărți de joc dintr-un pachet de 52 de cărți fără a pune una din cărți înapoi în pachet între trageri. Aceasta este numită **selectare fără înlocuire** sau **selectare aleatoare simplă**. Cu această metodă de selectare selecția noastră de 10 cărți nu va avea cărți duplicate.

Selectare cu înlocuire

Acum presupunem că tragem 10 cărți aleator din pachet, dar după fiecare extragere punem cartea la loc în pachet și amestecăm cărțile. Aceasta este numită **selectare cu înlocuire**. Cu această metodă, selecția de 10 cărți poate avea duplicate. Este chiar posibil să tragem 6 de cupă de 10 ori.

Gândiți: Care este probabilitatea să tragem 6 de cupă de 10 ori la rând?

Exemplul 2. Putem vedea aruncarea repetată a unui zar cu 8 fețe ca selectarea cu înlocuire din mulțimea $\{1, 2, 3, 4, 5, 6, 7, 8\}$. Deoarece fiecare număr este egal probabil, spunem că selectăm uniform din date. Aici este o subtilitate: fiecare dată este egal probabilă, dar dacă sunt valori repetate în date, acele valori vor avea o probabilitate mai mare de a fi alese.

Observație. În practică, dacă luăm un număr mic dintr-o mulțime foarte mare, atunci nu contează dacă selectăm cu sau fără înlocuire. De exemplu, dacă selecăm aleator 400 din cei 300 de milioane de locuitori ai SUA, atunci este atât de improbabil că aceeași persoană va fi aleasă de 2 ori încât nu este o diferență reală între selectarea cu sau fără înlocuire.

2.4 Repartiția empirică a datelor

Repartiția empirică a datelor este pur și simplu repartiția pe care o vedeți în date.

Exemplul 3. Presupunem că aruncăm un zar cu 8 fețe de 10 ori și obținem următoarele date, scrise în ordine crescătoare:

1, 1, 2, 3, 3, 3, 3, 4, 7, 7.

Imaginați-vă că scriem aceste valori pe 10 bucățele de hârtie, le punem într-o pălărie și tragem una aleator. Atunci, de exemplu, probabilitatea de a trage

un 3 este $4/10$ și probabilitatea de a trage un 4 este $1/10$. Repartiția empirică completă poate fi pusă într-un tabel de probabilitate ("value"="valoarea").

value x	1	2	3	4	7
$p(x)$	$2/10$	$1/10$	$4/10$	$1/10$	$2/10$

Notație. Dacă notăm adevărata repartiție din care provin datele cu F , atunci vom nota repartiția empirică a datelor cu F^* . Dacă avem destule date, atunci legea numerelor mari ne spune că F^* ar trebui să fie o bună aproximare a lui F .

Exemplul 4. În exemplul cu zaruri de mai sus, repartițiile adevărată și empirică sunt:

value x	1	2	3	4	5	5	7	8
true $p(x)$	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$
empirical $p(x)$	$2/10$	$1/10$	$4/10$	$1/10$	0	0	$2/10$	0

Adevărata repartiție F și repartiția empirică F^* a unui zar cu 8 fețe. Pe prima linie, în loc de al 2-lea "5" se va citi "6".

Deoarece F^* este obținută strict din date, o numim **repartiția empirică** a datelor. O vom numi de asemenea **repartiția de reselectare**. Totdeauna știm F^* explicit. În particular media lui F^* este chiar media de selecție \bar{x} .

2.5 Reselectarea

Bootstrap-ul empiric începe prin reselectarea datelor. Continuăm exemplul cu zaruri de mai sus.

Exemplul 5. Presupunem că avem 10 date, date în ordine crescătoare:

1, 1, 2, 3, 3, 3, 3, 4, 7, 7.

Vedem aceasta ca o **selecție** luată dintr-o repartiție subiacentă. A **reselecta** este a selecta cu înlocuire din repartiția empirică, de exemplu a pune aceste 10 numere într-o pălărie și a trage unul aleator. Apoi puneți numărul înapoi în pălărie și trageți iar. Trageți atâtea numere ca mărimea dorită a reselectiei. Pentru a ajunge un pic mai aproape de a implementa aceasta pe un calculator, reformulăm aceasta în modul următor. Notăm cele 10 date cu x_1, x_2, \dots, x_{10} . A reselecta este a trage un număr j din repartiția uniformă pe $\{1, 2, \dots, 10\}$ și a lua x_j ca valoarea noastră reselectată. În acest caz am putea face aceasta aruncând un zar cu 10 fețe. De exemplu, dacă aruncăm un 6, atunci valoarea noastră reselectată este 3, al 6-lea element din lista noastră. Dacă vrem o mulțime de date reselectate de mărime 5, atunci vom arunca

zarul cu 10 fețe de 5 ori și vom alege elementele corespunzătoare din lista datelor. Dacă cele 5 aruncări sunt

$$5, 3, 6, 6, 1,$$

atunci reselectia este

$$3, 2, 3, 3, 1.$$

Observații. 1. Deoarece selectăm cu înlocuire, aceeași dată poate apărea de mai multe ori când reselectăm.

2. De asemenea, deoarece selectăm cu înlocuire, putem avea o mulțime de date reselectate de orice mărime vrem, de exemplu am putea reselecta de 1000 de ori.

Desigur, în practică se folosește un pachet software ca **R** pentru a face reselectarea.

2.5.1 Notăția cu stelută

Dacă avem o selecție de mărime n

$$x_1, x_2, \dots, x_n,$$

atunci notăm o [reselectie de mărime \$m\$](#) adăugând o stelută simbolurilor

$$x_1^*, x_2^*, \dots, x_m^*.$$

Similar, la fel cum \bar{x} este media datelor originale, scriem [\$\bar{x}^*\$ pentru media datelor reselectate](#).

2.6 Bootstrap-ul empiric

Presupunem că avem n date

$$x_1, x_2, \dots, x_n,$$

provenite dintr-o repartiție F . O [selecție bootstrap empirică](#) este o reselectie de [aceeași mărime \$n\$](#) :

$$x_1^*, x_2^*, \dots, x_n^*.$$

Ar trebui să gândiți ultima ca o selecție de mărime n extrasă din repartiția empirică F^* . Pentru orice statistică v calculată din datele selecției originale, putem defini o statistică v^* cu aceeași formulă, dar calculată folosind datele reselectate în locul celor originale. Cu această notație putem formula principiul bootstrap.

2.6.1 Principiul bootstrap

Schema bootstrap este după cum urmează:

1. x_1, x_2, \dots, x_n este o selecție de date provenite dintr-o repartiție F .
2. O statistică u este calculată din selecție.
3. F^* este repartiția empirică a datelor (repartiția de reselectare).
4. $x_1^*, x_2^*, \dots, x_n^*$ este o reselectie a datelor de aceeași mărime ca selecția originală.
5. u^* este statistica calculată din reselectie.

Atunci principiul bootstrap spune că

1. $F^* \approx F$.
2. Variația lui u este bine aproximată de variația lui u^* .

Interesul nostru real este în punctul 2: putem aproxima variația lui u prin variația lui u^* . Vom exploata aceasta pentru a estima mărimea intervalelor de încredere.

2.6.2 De ce reselectia are aceeași mărime ca selecția originală?

Aceasta este direct: variația statisticii u va depinde de mărimea selecției. Dacă vrem să aproximăm această variație trebuie să folosim reselectii de aceeași mărime.

2.6.3 Exemplu didactic de un interval de încredere bootstrap empiric

Exemplul 6. Exemplu didactic. Începem cu o mulțime inventată de date care este destul de mică pentru a arăta fiecare pas explicit. Datele selecției sunt

30, 37, 36, 43, 42, 43, 43, 46, 41, 42.

Problema: Estimați media μ a repartiției subiacente și dați un interval de încredere bootstrap 80%.

Răspuns. Media de selecție este $\bar{x}=40.3$. Folosim aceasta ca o estimare a adevăratei medii μ a repartiției subiacente. Ca în exemplul 1, pentru a face intervalul de încredere trebuie să știm cât de mult repartiția lui \bar{x} variază în jurul lui μ . Adică, ne-ar plăcea să știm repartiția lui

$$\delta = \bar{x} - \mu.$$

Dacă am ști această repartiție, am putea afla $\delta_{.1}$ și $\delta_{.9}$, valorile critice 0.1 și 0.9 ale lui δ . Atunci am avea

$$P(\delta_{.9} \leq \bar{x} - \mu \leq \delta_{.1} | \mu) = 0.8 \iff P(\bar{x} - \delta_{.9} \geq \mu \geq \bar{x} - \delta_{.1} | \mu) = 0.8,$$

ceea ce dă intervalul de încredere 80%

$$[\bar{x} - \delta_{.1}, \bar{x} - \delta_{.9}].$$

Ca întotdeauna la intervalele de încredere, ne-am grăbit să subliniem că probabilitățile calculate mai sus sunt probabilitățile privind statistica \bar{x} **dat fiind că adevărata medie este μ** .

Principiul bootstrap dă o abordare practică a estimării distribuției lui $\delta = \bar{x} - \mu$. Spune că o putem aproxima prin repartiția lui

$$\delta^* = \bar{x}^* - \bar{x},$$

unde \bar{x}^* este media unei selecții bootstrap empirice.

Iată cheia frumoasă pentru aceasta: deoarece δ^* este calculată reselectând datele originale, putem simula pe calculator δ^* de câte ori vrem. Prin urmare, din legea numerelor mari, putem estima repartiția lui δ^* cu mare precizie.

Acum să ne întoarcem la datele de selecție cu 10 puncte. Am folosit R pentru a genera 20 de selecții bootstrap, fiecare de mărime 10. Fiecare dintre cele 20 de coloane ale următorului tabel este o selecție bootstrap.

43	36	46	30	43	43	43	37	42	42	43	37	36	42	43	43	42	43	42	43
43	41	37	37	43	43	46	36	41	43	43	42	41	43	46	36	43	43	43	42
42	43	37	43	46	37	36	41	36	43	41	36	37	30	46	46	42	36	36	43
37	42	43	41	41	42	36	42	42	43	42	43	41	43	36	43	43	41	42	46
42	36	43	43	42	37	42	42	42	46	30	43	36	43	43	42	37	36	42	30
36	36	42	42	36	36	43	41	30	42	37	43	41	41	43	43	42	46	43	37
43	37	41	43	41	42	43	46	46	36	43	42	43	30	41	46	43	46	30	43
41	42	30	42	37	43	43	42	43	43	46	43	30	42	30	42	30	43	43	42
46	42	42	43	41	42	30	37	30	42	43	42	43	37	37	37	42	43	43	46
42	43	43	41	42	36	43	30	37	43	42	43	41	36	37	41	43	42	43	43

Apoi calculăm $\delta^* = \bar{x}^* - \bar{x}$ pentru fiecare selecție bootstrap (i.e. fiecare coloană) și le sortăm de la cea mai mică la cea mai mare:

-1.6, -1.4, -1.4, -0.9, -0.5, -0.2, -0.1, 0.1, 0.2, 0.2, 0.4, 0.4, 0.7, 0.9, 1.1, 1.2, 1.2, 1.6, 1.6, 2.

Vom aproxima valorile critice $\delta_{.1}$ și $\delta_{.9}$ prin $\delta_{.1}^*$ și $\delta_{.9}^*$. Deoarece $\delta_{.1}^*$ este la a 90-a percentilă, alegem al 18-lea element din listă, i.e. 1.6. Analog, deoarece $\delta_{.9}^*$ este la a 10-a percentilă, alegem al 2-lea element din listă, i.e. -1.4.

De aceea intervalul nostru de încredere 80% bootstrap pentru μ este

$$[\bar{x} - \delta_{.1}, \bar{x} - \delta_{.9}] = [40.3 - 1.6, 40.3 + 1.4] = [38.7, 41.7].$$

În acest exemplu am generat doar 20 de selecții bootstrap pentru ca să încapă pe pagină. Folosind R, am putea genera cel puțin 10000 de selecții bootstrap pentru a obține o estimare foarte precisă pentru $\delta_{.1}^*$ și $\delta_{.9}^*$.

2.6.4 Justificarea pentru principiul bootstrap

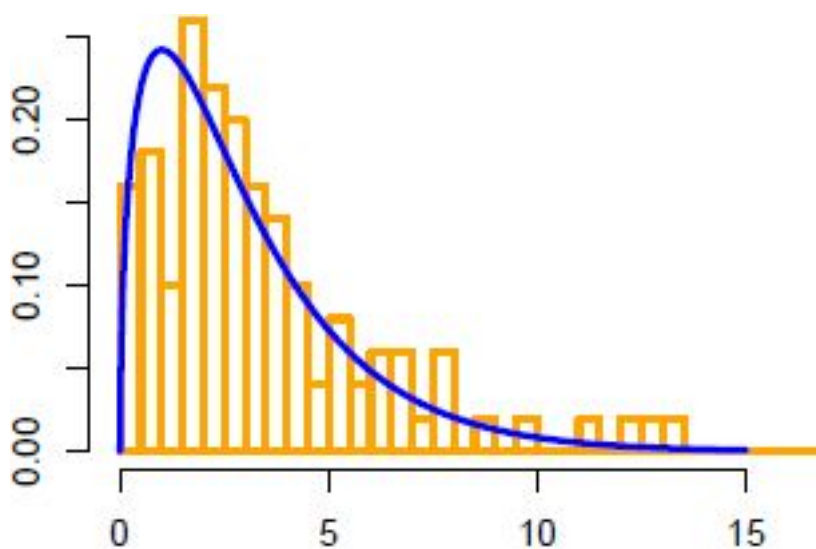
Bootstrap-ul este remarcabil pentru că reselectarea ne dă o estimare decentă a cum poate varia estimarea punctuală. Bootstrap-ul este bazat pe legea numerelor mari, care spune, pe scurt, că având destule date, repartiția empirică va fi o bună aproximare a adevăratei repartiții. Vizual, spune că histograma datelor ar aproxima densitatea adevăratei repartiții.

Reselectarea nu poate îmbunătăți estimarea noastră punctuală. De exemplu, dacă estimăm media μ prin \bar{x} , atunci în bootstrap calculăm \bar{x}^* pentru multe reselectări ale datelor. Dacă luăm media tuturor \bar{x}^* , ne-am aștepta să fie foarte aproape de \bar{x} . Aceasta nu ne-ar spune nimic nou despre adevărata valoare a lui μ .

Chiar cu o cantitate corectă de date, potrivirea dintre repartițiile adevărată și empirică nu este perfectă, deci va fi o eroare în estimarea mediei (sau oricărei alte valori). Dar cantitatea de variație a estimărilor este mult mai puțin sensibilă la diferențe între densitate și histogramă. Atât timp cât sunt rezonabil de aproape, atât repartiția adevărată, cât și cea empirică vor avea cantități similare de variație. Deci, în general principiul bootstrap este mai robust când aproximăm repartiția variației relative decât când aproximăm repartiții absolute.

Repartiția (peste diferite seturi de date experimentale) a lui \bar{x} este "centrată" în μ și repartiția lui \bar{x}^* este centrată în \bar{x} . Dacă este o separare semnificativă între \bar{x} și μ , atunci aceste 2 repartiții vor diferi de asemenea semnificativ. Pe de altă parte, repartiția lui $\delta = \bar{x} - \mu$ descrie variația lui \bar{x} în jurul centrului lui. Analog, repartiția lui $\delta^* = \bar{x}^* - \bar{x}$ descrie variația lui \bar{x}^* în jurul centrului lui. Deci chiar dacă cele 2 centre sunt destul de diferite, cele 2 variații în jurul centrelor pot fi aproximativ egale.

Figura de mai jos ilustrează cum repartiția empirică aproximează adevărata repartiție. Pentru a face figura au fost generate 100 de valori aleatoare dintr-o repartiție χ^2 cu 3 grade de libertate. Figura arată pdf-ul adevăratei repartiții ca o linie albastră și histograma repartiției empirice cu portocaliu.



Repartițiile adevărată și empirică sunt aproximativ egale.

2.7 Alte statistici

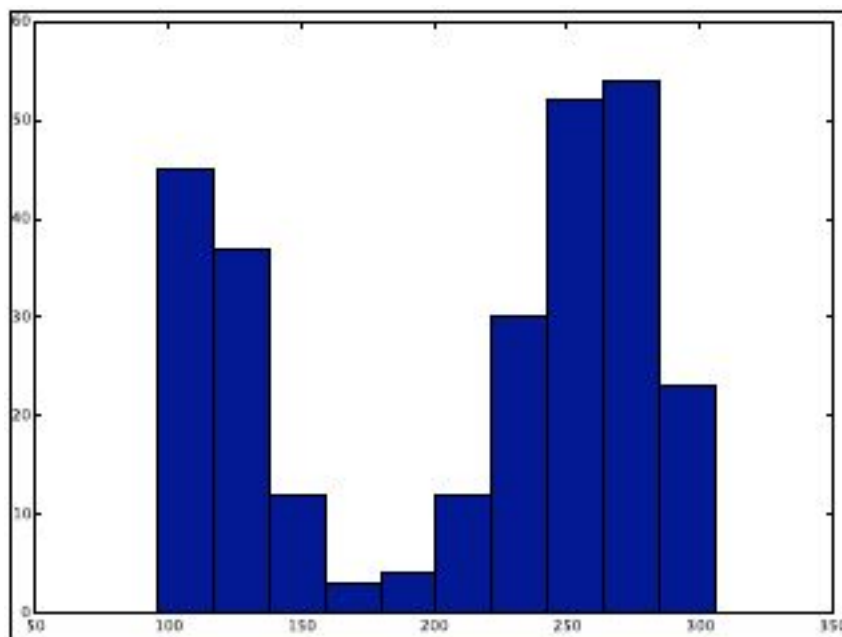
Până acum am evitat intervalele de încredere pentru mediană și alte statistici deoarece repartițiile lor de selecție sunt greu de descris teoretic. Bootstrap-ul nu are această problemă. De fapt, pentru a trata mediana, tot ce avem de făcut este să schimbăm "mean" în "median" în codul R de la exemplul 6.

Exemplul 7. Old Faithful: intervale de încredere pentru mediană

Old Faithful este un gheizer în parcul național Yellowstone din Wyoming:

http://en.wikipedia.org/wiki/Old_Faithful.

Există un set de date care dă duratele a 272 de erupții consecutive. Iată histograma datelor.



Întrebare. Estimați lungimea mediană a unei erupții și dați un interval de încredere de 90% pentru mediană.

Răspuns. Facem un rezumat al pașilor necesari pentru a răspunde întrebării.

1. Datele: x_1, \dots, x_{272} .
2. Mediana datelor: $x_{\text{median}} = 240$.
3. Află mediana x_{median}^* a unei selecții bootstrap x_1^*, \dots, x_{272}^* . Repetă de 1000 de ori.
4. Calculează diferențele bootstrap

$$\delta^* = x_{\text{median}}^* - x_{\text{median}}.$$

Pune aceste 1000 de valori în ordine crescătoare și alege valorile critice .95 și .05, i.e. a 50-a și a 950-a valoare. Numește aceste valori $\delta_{.95}^*$ și $\delta_{.05}^*$.

5. Principiul bootstrap spune că putem folosi $\delta_{.95}^*$ și $\delta_{.05}^*$ ca estimări pentru $\delta_{.95}$ și $\delta_{.05}$. Deci intervalul nostru de încredere bootstrap 90% pentru mediană este

$$[x_{\text{median}} - \delta_{.05}^*, x_{\text{median}} - \delta_{.95}^*].$$

CI (intervalul de încredere) 90% bootstrap aflat pentru datele Old Faithful a fost [235,250]. Deoarece s-au folosit 1000 de selecții bootstrap, o nouă simulare plecând de la aceleași date de selecție va produce un interval similar. Dacă în pasul 3 creștem numărul de selecții bootstrap la 10000, atunci intervalele produse prin simulare vor varia chiar mai puțin. O strategie uzuală este să creștem numărul de selecții bootstrap până simulările rezultate produc intervale care variază mai puțin decât un nivel acceptabil.

Exemplul 8. Folosind datele Old Faithful, estimați $P(|\bar{x} - \mu| > 5|\mu)$.

Răspuns. Procedăm exact ca în exemplul precedent folosind media în locul medianei.

1. Datele: x_1, \dots, x_{272} .
2. Media datelor: $\bar{x} = 209.27$.
3. Află media \bar{x}^* a 1000 de selecții bootstrap empirice: x_1^*, \dots, x_{272}^* .
4. Calculează diferențele bootstrap

$$\delta^* = \bar{x}^* - \bar{x}.$$

5. Principiul bootstrap spune că putem folosi repartiția lui δ^* ca o aproximație pentru repartiția lui $\delta = \bar{x} - \mu$. De aici,

$$P(|\bar{x} - \mu| > 5|\mu) = P(|\delta| > 5|\mu) \approx P(|\delta^*| > 5).$$

Simularea Bootstrap pentru datele Old Faithful a dat 0.225 pentru această probabilitate.

2.8 Bootstrap parametric

Exemplele din secțiunea anterioară au folosit toate bootstrap-ul empiric, care nu face nicio presupunere despre repartiția subiacentă și extrage selecții bootstrap prin reselectarea datelor. În această secțiune vom aborda [bootstrap-ul parametric](#). Singura diferență dintre bootstrapul parametric și empiric este sursa selecției bootstrap. Pentru bootstrap-ul parametric, generăm selecția bootstrap dintr-o repartiție parametrizată.

Iată elementele folosirii bootstrap-ului parametric pentru a estima un interval de încredere pentru un parametru:

0. Date: x_1, \dots, x_n extrase dintr-o repartiție $F(\theta)$ cu parametru necunoscut θ .
1. O statistică $\hat{\theta}$ care estimează pe θ .
2. Selecțiile noastre bootstrap sunt extrase din $F(\hat{\theta})$.
3. Pentru fiecare selecție bootstrap

$$x_1^*, \dots, x_n^*$$

calculăm $\hat{\theta}^*$ și diferența bootstrap $\delta^* = \hat{\theta}^* - \hat{\theta}$.

4. Principiul bootstrap spune că repartiția lui δ^* aproximează repartiția lui $\delta = \hat{\theta} - \theta$.
5. Folosiți diferențele bootstrap pentru a face un interval de încredere bootstrap pentru θ .

Exemplul 9. Presupunem că datele x_1, \dots, x_{300} sunt extrase dintr-o repartiție

$\exp(\lambda)$. Presupunem de asemenea că media datelor $\bar{x} = 2$. Estimați λ și dați un interval de încredere bootstrap parametric 95% pentru λ .

Răspuns. Cel mai ușor este să explicăm soluția folosind codul R comentat.

```
# Bootstrap parametric
# Sunt 300 de date cu media 2.
# Presupunem că datele sunt exp(lambda).
# PROBLEMA: Calculați un interval de încredere bootstrap parametric
95% pentru lambda.
# Se dau numărul datelor și media.
n=300
xbar=2
# MLE pentru lambda este 1/xbar
lambdahat=1/xbar
# Generăm selecțiile bootstrap.
# Fiecare coloană este o selecție bootstrap (de 300 de valori reselectionate).
nboot=1000
#Iată diferența cheie față de bootstrap-ul empiric:
# Extragem selecția bootstrap din exponențiala(lambdahat).
x=rexp(n*nboot,lambdahat)
bootstrapsample=matrix(x,nrow=n,ncol=nboot)
# Calculăm lambdastar bootstrap.
lambdastar=1/colMeans(bootstrapsample)
# Calculăm diferențele.
deltastar=lambdastar-lambdahat
# Aflăm cuantilele 0.05 și 0.95 pentru deltaxstar.
d=quantile(deltastar, c(0.05,0.95))
# Calculăm intervalul de încredere 95% pentru lambda.
ci=lambdahat-c(d[2],d[1])
# Următoarele linii de cod sunt doar un mod de a formata textul de
ieșire.
# R are și alte moduri de a face aceasta.
s=sprintf("Interval de incredere pentru lambda:  [%.3f, %.3f]", ci[1],
ci[2])
cat(s)
```


2.9 Transcrieri R adnotate

2.9.1 Folosirea lui R pentru a genera un interval de încredere bootstrap empiric

Acest cod generează doar 20 de selecții bootstrap. În practica reală ar fi generate mult mai multe selecții bootstrap. Este făcut un interval de încredere pentru medie.

```
# Date pentru exemplul 6
x=c(30,37,36,43,42,43,43,46,41,42)
n=length(x)
# Media de selecție
xbar=mean(x)
nboot=20
# Generăm 20 de selecții bootstrap, i.e. o matrice n x 20 de reselectii
# aleatoare din x.
tmpdata=sample(x,n*nboot,replace=TRUE)
bootstrapsample=matrix(tmpdata,nrow=n,ncol=nboot)
# Calculăm mediile  $\bar{x}^*$ .
bsmeans=colMeans(bootstrapsample)
# Calculăm  $\delta^*$  pentru fiecare selecție bootstrap.
deltastar=bsmeans-xbar
# Sortăm rezultatele.
sorteddeltastar=sort(deltastar)
# Aflăm valorile critice .1 și .9 ale lui deltaxstar.
d9=sorteddeltastar[2]
d1=sorteddeltastar[18]
# Calculăm intervalul de încredere 80% pentru medie.
ci=xbar-c(d1,d9)
cat('Intervalul de incredere: ',ci, '\n')
```

3 Regresia liniară

3.1 Scopurile învățării

1. Să poată folosi metoda celor mai mici pătrate pentru a potrivi o dreaptă cu date bivariate.
2. Să poată da o formulă pentru eroarea pătratică totală la potrivirea oricărui tip de curbă cu datele.
3. Să poată spune cuvintele homoscedasticitate și heteroscedasticitate.

3.2 Introducere

Presupunem că avem colectate date bivariate $(x_i, y_i), i = 1, \dots, n$. Scopul regresiei liniare este modelarea relației dintre x și y prin aflarea unei funcții $y = f(x)$ care dă o potrivire apropiată cu datele. Presupunerile modelării pe care le vom folosi sunt că x_i nu sunt aleatoare și că y_i este o funcție de x_i plus un zgomot aleator. Cu aceste presupuneri, x este numită **variabilă independentă** sau **predictor** și y este numită variabilă **dependentă** sau **răspuns**.

Exemplul 1. Costul unui timbru de clasa întâi în dolari de-a lungul timpului este dat în lista următoare:

.05 (1963)	.06 (1968)	.08 (1971)	.10 (1974)	.13 (1975)	.15 (1978)	.20 (1981)	.22 (1985)
.25 (1988)	.29 (1991)	.32 (1995)	.33 (1999)	.34 (2001)	.37 (2002)	.39 (2006)	.41 (2007)
.42 (2008)	.44 (2009)	.45 (2012)	.46 (2013)	.49 (2014)			

Folosind codul R:

```
x=c(3,8,11,14,15,18,21,25,28,31,35,39,41,42,46,47,48,49,52,53,54)
```

```
y=c(5,6,8,10,13,15,20,22,25,29,32,33,34,37,39,41,42,44,45,46,49)
```

```
lm(y~x),
```

obținem:

Call:

```
lm(formula = y ~ x)
```

Coefficients:

```
(Intercept) x
```

```
-0.1324 0.8791.
```

Am aflat că dreapta care dă ”potrivirea celor mai mici pătrate” cu aceste date (dreapta de regresie) este

$$y = -0.1324 + 0.8791x,$$

unde x este numărul de ani de la 1960, iar y este în cenți.

Folosind acest rezultat ”prezicem” că în 2023 ($x = 63$), costul unui timbru va fi 55 de cenți (deoarece $-0.1324 + 0.8791 \cdot 63 = 55.2509$).

Folosind codul R (în continuarea celui de mai sus):

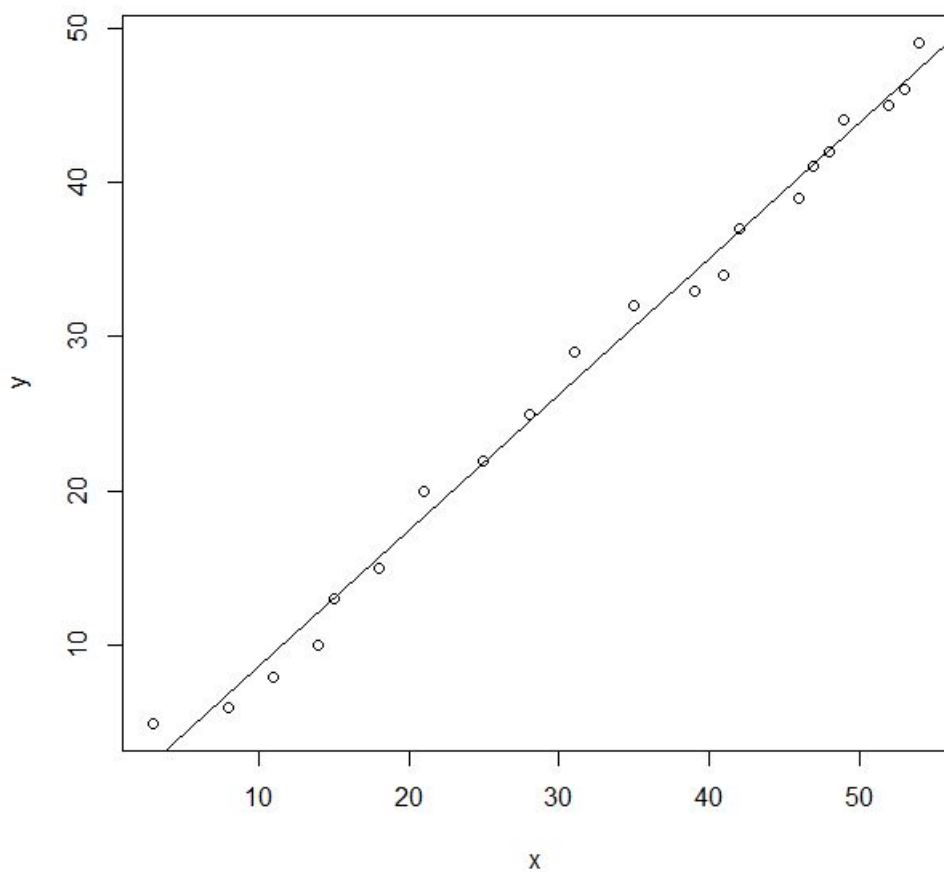
```
a=-0.1324
```

```
b=0.8791
```

```
plot(x,y)
```

```
abline(a,b)
```

obținem:



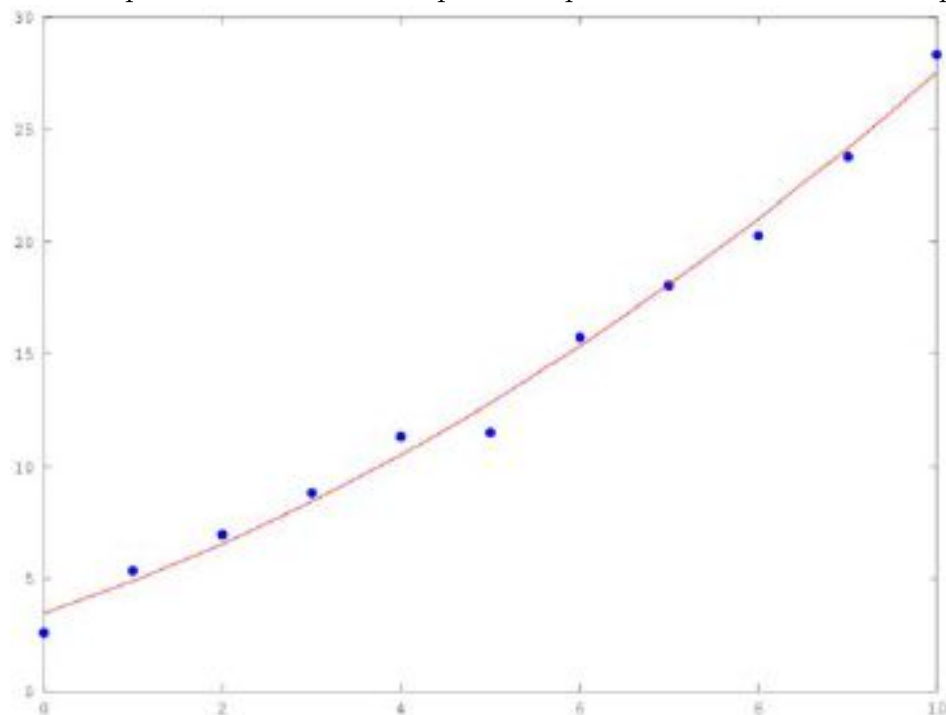
Costul timbrului (cenți) vs. timp (ani din 1960)

Niciuna din date nu se află chiar pe dreaptă. Mai degrabă această dreaptă are "cea mai bună potrivire" în raport cu [toate datele](#), cu o mică eroare pentru fiecare dată.

Exemplul 2. Presupunem că avem n perechi de tați și fii adulți. Fie x_i și y_i înălțimile celui de-al i -lea tată, respectiv fiu. Dreapta celor mai mici pătrate

pentru aceste date poate fi folosită pentru a prezice înălțimea de adult a unui băiat tânăr din cea a tatălui lui.

Exemplul 3. Nu suntem limitați la drepte cu cea mai bună potrivire. $\forall d \in \mathbb{N}^*$, metoda celor mai mici pătrate poate fi folosită pentru a afla un polinom de grad d cu "cea mai bună potrivire cu datele". Iată o figură arătând potrivirea datelor cu o parabolă prin metoda celor mai mici pătrate:



Potrivirea unei parabole, $y = b_2x^2 + b_1x + b_0$ cu datele

3.3 Potrivirea unei drepte folosind cele mai mici pătrate

Presupunem că avem datele (x_i, y_i) ca mai sus. Scopul este să aflăm dreapta

$$y = \beta_1x + \beta_0,$$

care "se potrivește cel mai bine" cu datele. Modelul nostru spune că fiecare y_i este prezis de x_i până la o eroare ϵ_i :

$$y_i = \beta_1x_i + \beta_0 + \epsilon_i.$$

Deci,

$$\epsilon_i = y_i - \beta_1x_i - \beta_0.$$

Metoda celor mai mici pătrate află valorile $\hat{\beta}_0$ și $\hat{\beta}_1$ ale lui β_0 și β_1 care minimizează suma pătratelor erorilor:

$$S(\beta_0, \beta_1) = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0)^2.$$

Folosind analiza matematică (detalii în adaos), aflăm

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad (2)$$

unde

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s_{xx} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Aici, \bar{x} este media de selecție a lui x , \bar{y} este media de selecție a lui y , s_{xx} este dispersia de selecție a lui x și s_{xy} este covarianța de selecție a lui x și y .

Exemplul 4. Folosiți cele mai mici pătrate pentru a potrivi cu o dreaptă următoarele date: (0,1), (2,1), (3,4).

Răspuns. În cazul nostru, $(x_1, y_1) = (0, 1)$, $(x_2, y_2) = (2, 1)$ și $(x_3, y_3) = (3, 4)$. Deci

$$\bar{x} = \frac{1}{3}(x_1 + x_2 + x_3) = \frac{1}{3}(0 + 2 + 3) = \frac{5}{3},$$

$$\bar{y} = \frac{1}{3}(y_1 + y_2 + y_3) = \frac{1}{3}(1 + 1 + 4) = \frac{6}{3} = 2,$$

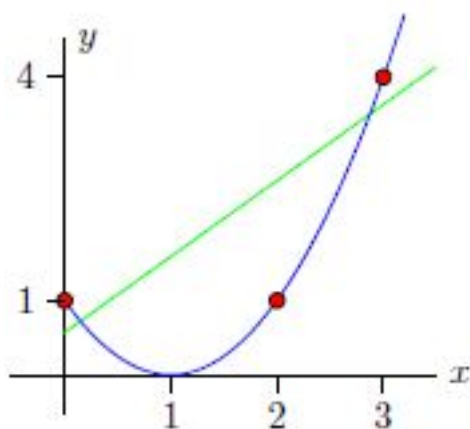
$$s_{xx} = \frac{1}{3-1} \sum_{i=1}^3 (x_i - \bar{x})^2 = \frac{1}{2} \left[\left(0 - \frac{5}{3}\right)^2 + \left(2 - \frac{5}{3}\right)^2 + \left(3 - \frac{5}{3}\right)^2 \right] = \frac{7}{3},$$

$$s_{xy} = \frac{1}{2} \left[\left(0 - \frac{5}{3}\right)(1 - 2) + \left(2 - \frac{5}{3}\right)(1 - 2) + \left(3 - \frac{5}{3}\right)(4 - 2) \right] = 2;$$

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{2}{\frac{7}{3}} = \frac{6}{7};$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2 - \frac{6}{7} \cdot \frac{5}{3} = 2 - \frac{10}{7} = \frac{4}{7}.$$

Deci dreapta de regresie a celor mai mici pătrate are ecuația $y = \frac{4}{7} + \frac{6}{7}x$. Aceasta este arătată ca dreapta verde din figura următoare.



Potrivirea celor mai mici pătrate a unei drepte (verde) și a unei parabole (albastru)

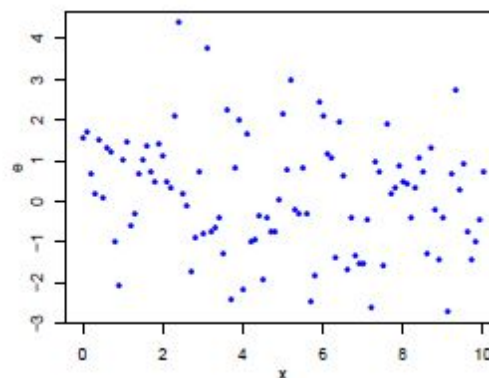
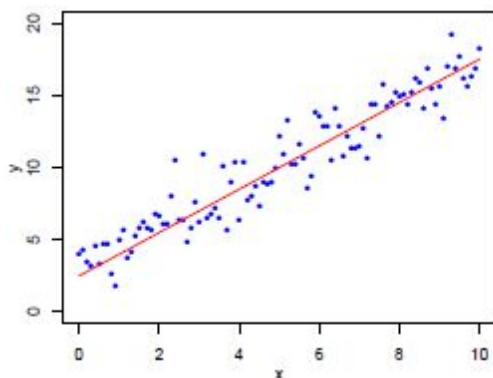
Regresie liniară simplă: Este puțin confuz, dar cuvântul "liniară" din "regresie liniară" nu se referă la potrivirea unei drepte. Totuși, cea mai uzuală curbă pentru potrivire este o dreaptă. Potrivirea unei drepte la date bivariate este numită [regresie liniară simplă](#).

3.3.1 Reziduuri

Pentru o dreaptă, modelul este

$$y_i = \hat{\beta}_1 x_i + \hat{\beta}_0 + \epsilon_i.$$

Gândim $\hat{\beta}_1 x_i + \hat{\beta}_0$ ca prezicând sau explicând y_i . Termenul rămas ϵ_i este numit [reziduul](#), pe care-l gândim ca pe un zgomot aleator sau o eroare de măsurare. O verificare vizuală folositoare a modelului de regresie liniară este reprezentarea reziduurilor. Datele ar trebui să fie lângă dreapta de regresie. Reziduurile ar trebui să arate cam la fel de-a lungul domeniului lui x .

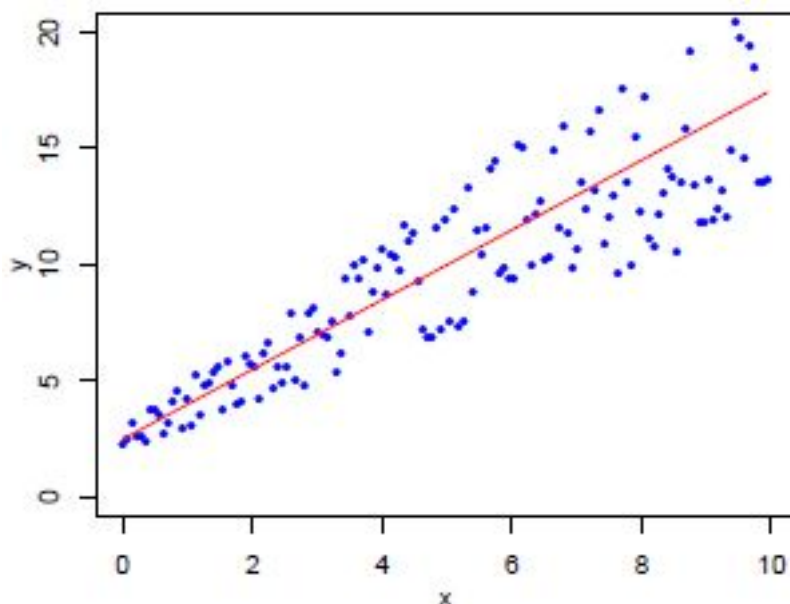


Date cu dreapta de regresie (stânga) și reziduuri (dreapta). Observați homoscedasticitatea.

3.3.2 Homoscedasticitatea

O presupunere importantă a modelului de regresie liniară este că reziduurile ϵ_i au aceeași dispersie $\forall i$. Această presupunere este numită **homoscedasticitate**. Puteți vedea aceasta în cazul ambelor figuri de mai sus. Datele sunt în banda de lățime fixă în jurul dreptei de regresie și la fiecare x reziduurile au cam aceeași împrăștiere verticală.

Mai jos este o figură arătând date **heteroscedastice**. Împrăștierea verticală a datelor crește când x crește. Înainte de a folosi cele mai mici pătrate pe aceste date ar trebui să transformăm datele pentru a fi homoscedastice.



Date heteroscedastice

3.4 Regresie liniară pentru potrivirea polinoamelor

Potrivirea unei drepte la date este numită **regresie liniară simplă**. Putem de asemenea folosi regresia liniară pentru a potrivi polinoame cu datele. Folosirea cuvântului "liniară" în ambele cazuri poate părea confuză. Aceasta este deoarece cuvântul "liniară" din "regresia liniară" nu se referă la potrivirea unei drepte. Mai degrabă se referă la ecuațiile algebrice liniare pentru parametrii necunoscuți β_i , i.e. fiecare β_i are exponentul 1.

Exemplul 5. Luați aceleași date ca în exemplul 4 și folosiți cele mai mici pătrate pentru a afla parabola cu cea mai bună potrivire pentru date.

Răspuns. O parabolă are formula $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Eroarea pătratică este

$$S(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^3 (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2.$$

După substituirea valorilor date pentru x_i și y_i putem folosi analiza matematică (egalăm derivatele parțiale în raport cu $\beta_0, \beta_1, \beta_2$ cu 0, obținând un sistem de 3 ecuații liniare cu 3 necunoscute) pentru a afla tripletul $(\beta_0, \beta_1, \beta_2)$ care minimizează S . Sau putem folosi codul R

```
x=c(0,2,3)
y=c(1,1,4)
C=cbind(1,x,x^ 2)
solve(t(C)%*%C,t(C)%*%y).
```

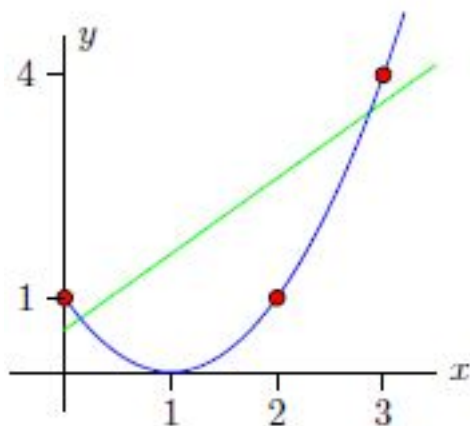
Sau codul R

```
x=c(0,2,3)
y=c(1,1,4)
x1=x
x2=x^ 2
lm(y~x1+x2)
```

Cu aceste date, parabola celor mai mici pătrate are ecuația

$$y = 1 - 2x + x^2.$$

Pentru 3 puncte, potrivirea pătratică este perfectă.



Potrivirea celor mai mici pătrate a unei drepte (verde) și a unei parabole (albastru)

Exemplul 6. Perechile (x_i, y_i) pot da vârsta și mărimea vocabularului a n copii. Deoarece copiii mici dobândesc cuvinte noi într-un ritm accelerat, putem ghici că un polinom de grad mai mare poate fi cea mai bună potrivire

pentru date.

Exemplul 7. (Transformarea datelor). Uneori este necesar să transformăm datele înainte de a folosi regresia liniară. De exemplu, presupunem că relația este exponențială, i.e. $y = ce^{ax}$. Atunci

$$\ln(y) = ax + \ln(c).$$

Deci putem folosi regresia liniară simplă pentru a obține un model

$$\ln(y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

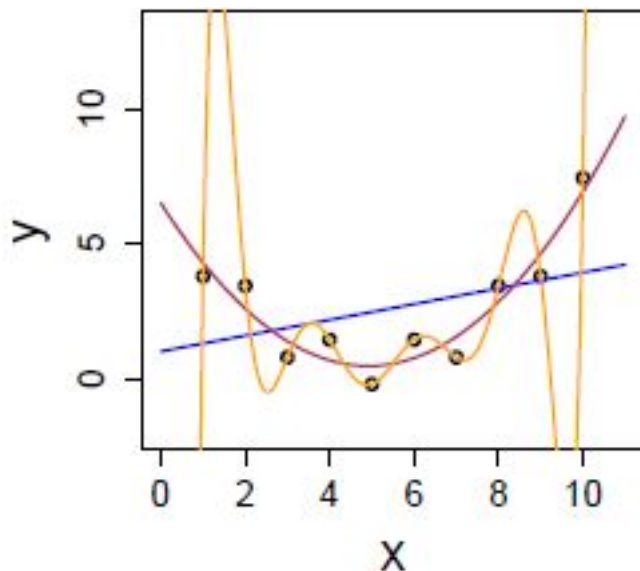
și apoi obținem modelul exponențial

$$y_i = e^{\hat{\beta}_0} e^{\hat{\beta}_1 x_i}.$$

3.4.1 Suprapotrivirea

Putem totdeauna obține o potrivire mai bună folosind un polinom de ordin mai mare. De exemplu, date fiind 6 date bivariate (cu x_i distincte) se poate totdeauna afla un polinom de grad 5 care trece prin toate. Aceasta poate duce la [suprapotrivire](#). Adică, potrivirea zgomotului la fel de bine ca adevărata relație între x și y . Un model suprapotrivit va potrivi datele originale mai bine, dar va prezice mai puțin bine y pentru noi valori ale lui x . O provocare a modelării statistice este echilibrarea potrivirii modelului cu complexitatea modelului.

Exemplul 8. În reprezentarea de mai jos potrivim polinoame de gradul 1, 2 și 9 la 10 date bivariate. Modelul de gradul 2 (maro) dă o potrivire semnificativ mai bună decât modelul de gradul 1 (albastru). Modelul de gradul 9 (portocaliu) dă o potrivire exactă cu datele, dar dintr-o privire am ghici că este suprapotrivit. Adică, nu ne așteptăm să potrivească bine următoarea dată bivariată pe care o vedem. De fapt, datele au fost generate folosind un model pătratic, deci modelul de gradul 2 va tinde să facă cea mai bună potrivire cu date noi.



3.4.2 Funcția R `lm`

Nu facem regresia liniară cu mâna. Regresia liniară se reduce la rezolvarea sistemelor de ecuații liniare, i.e. la calcul matriceal. Funcția R `lm` poate fi folosită la potrivirea unui polinom de orice grad cu datele. ([lm înseamnă model liniar](#)). De fapt, `lm` poate potrivi multe tipuri de funcții, exceptând polinoamele, după cum puteți explora folosind ajutorul lui R sau google.

3.5 Regresie liniară multiplă

Datele nu sunt totdeauna bivariante. Pot fi trivariate sau chiar de o dimensiune mai mare. Presupunem că avem datele de forma

$$(y_i, x_{1i}, x_{2i}, \dots, x_{mi}).$$

Putem analiza aceste date într-o manieră foarte similară cu datele bivariante. Adică, putem folosi cele mai mici pătrate pentru a potrivi modelul

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m.$$

Aici fiecare x_j este o variabilă predictor și y este variabila de răspuns. De exemplu, putem fi interesați de cum variază o populație de pești în funcție nivelele măsurate ale câtorva poluanți, sau am vrea să prezicem înălțimea de adult a unui fiu pe baza înălțimilor tatălui și a mamei.

3.6 Cele mai mici pătrate ca un model statistic

Modelul de regresie liniară pentru potrivirea unei drepte spune că valoarea y_i din perechea (x_i, y_i) este extrasă dintr-o variabilă aleatoare

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

unde termenii de "eroare" ε_i sunt variabile aleatoare independente cu media 0 și deviația standard σ . Presupunerea standard este că ε_i sunt i.i.d. cu repartiția $N(0, \sigma^2)$. În orice caz, media lui Y_i este dată de:

$$E(Y_i) = \beta_0 + \beta_1 x_i + E(\varepsilon_i) = \beta_0 + \beta_1 x_i.$$

Din această perspectivă, metoda celor mai mici pătrate alege valorile lui β_0 și β_1 care minimizează dispersia de selecție din jurul dreptei.

De fapt, estimarea celor mai mici pătrate $(\hat{\beta}_0, \hat{\beta}_1)$ coincide cu estimarea de verosimilitate maximă pentru parametrii (β_0, β_1) ; adică, dintre toți coeficienții posibili, $(\hat{\beta}_0, \hat{\beta}_1)$ sunt cei care fac datele observate cele mai probabile.

3.7 Regresia la medie

Motivul termenului "regresie" este că variabila de răspuns prezisă y va tinde să fie "mai aproape" de (i.e. să regreseze la) media ei decât variabila predictor x este față de media ei. Aici "mai aproape" este în ghilimele deoarece trebuie să controlăm scala (i.e. deviația standard) fiecărei variabile. Modul de a controla scala este mai întâi să standardizăm fiecare variabilă.

$$u_i = \frac{x_i - \bar{x}}{\sqrt{s_{xx}}}, \quad v_i = \frac{y_i - \bar{y}}{\sqrt{s_{yy}}}.$$

Standardizarea schimbă media în 0 și dispersia în 1:

$$\bar{u} = \bar{v} = 0, \quad s_{uu} = s_{vv} = 1.$$

Proprietățile algebrice ale covarianței arată că

$$s_{uv} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \rho,$$

coeficientul de corelație. Astfel, potrivirea celor mai mici pătrate pentru $v = \beta_0 + \beta_1 u$ are

$$\hat{\beta}_1 = \frac{s_{uv}}{s_{uu}} = \rho \text{ și } \hat{\beta}_0 = \bar{v} - \hat{\beta}_1 \bar{u} = 0.$$

Deci dreapta celor mai mici pătrate este $v = \rho u$. Deoarece ρ este coeficientul de corelație, el este între -1 și 1. Presupunem că este pozitiv și mai mic ca 1 (i.e., x și y sunt pozitiv, dar nu perfect corelate). Atunci formula $v = \rho u$ înseamnă că, dacă u este pozitiv, atunci valoarea prezisă a lui v este mai mică decât u . Adică, v este mai aproape de 0 decât u . Echivalent,

$$\frac{y - \bar{y}}{\sqrt{s_{yy}}} < \frac{x - \bar{x}}{\sqrt{s_{xx}}},$$

i.e., y regresează la \bar{y} . Standardizarea are grijă de scală.

Considerăm cazul extrem al corelației 0 între x și y . Atunci, indiferent de valoarea lui x , valoarea prezisă a lui y este totdeauna \bar{y} . Adică, y a regresat până la media lui.

Dreapta de regresie trece totdeauna prin punctul (\bar{x}, \bar{y}) .

Exemplul 9. Regresia la medie este importantă în studiile longitudinale. Rice (*Mathematical Statistics and Data Analysis*) dă următorul exemplu. Presupunând că li se dau copiilor un test IQ la vârsta de 4 ani și altul la vârsta de 5 ani, ne așteptăm ca rezultatele să fie pozitiv corelate. Analiza de mai sus spune că, în medie, acei copii care au făcut slab la primul test tind să arate îmbunătățire (i.e. regresează la medie) în al 2-lea test. Astfel, o intervenție inutilă poate fi interpretată greșit ca utilă deoarece pare a îmbunătăți scorurile.

Exemplul 10. Alt exemplu cu consecințe practice este recompensa și pedeapsa. Imaginați-vă o școală unde performanța înaltă la un examen este recompensată și performanța slabă este pedepsită. Regresia la medie ne spune că (în medie) studenții foarte performanți vor face puțin mai slab la următorul examen și studenții puțin performanți vor face puțin mai bine. O viziune ne-sofisticată a datelor va face să pară că pedeapsa a îmbunătățit performanța și recompensa de fapt a scăzut performanța. Sunt reale consecințe dacă cei cu autoritate acționează după această idee.

3.8 Adaos

3.8.1 Demonstrația formulei pentru potrivirea celor mai mici pătrate a unei drepte

Cea mai directă demonstrație este cu analiza matematică. Suma erorilor pătrate este

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0)^2.$$

Luând derivatele parțiale (și reamintind că x_i și y_i sunt date, deci constante)

$$\frac{\partial S}{\partial \beta_0} = \sum_{i=1}^n -2(y_i - \beta_1 x_i - \beta_0) = 0$$

$$\frac{\partial S}{\partial \beta_1} = \sum_{i=1}^n -2x_i(y_i - \beta_1 x_i - \beta_0) = 0.$$

Se obține următorul sistem de 2 ecuații liniare în necunoscutele β_0 și β_1 :

$$\left(\sum_{i=1}^n x_i\right) \beta_1 + n\beta_0 = \sum_{i=1}^n y_i$$

$$\left(\sum_{i=1}^n x_i^2\right) \beta_1 + \left(\sum_{i=1}^n x_i\right) \beta_0 = \sum_{i=1}^n x_i y_i$$

Rezolvând sistemul obținem formulele (2).

Pentru multe aplicații între discipline vezi:

http://en.wikipedia.org/wiki/Linear_regression#Applications_of_linear_regression.

3.8.2 Măsurarea potrivirii

Odată ce se calculează coeficienții de regresie, este important să verificăm cât de bine modelul de regresie se potrivește cu datele (i.e., cât de aproape cea mai potrivită dreaptă urmărește datele). O măsură uzuală dar brută a ”bunătății de potrivire” este **coeficientul de determinare**, notat R^2 . **Suma totală a pătratelor** este dată de:

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2.$$

Suma reziduală a pătratelor este dată de suma pătratelor reziduurilor. Când potrivim o dreaptă, aceasta este:

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

RSS este porțiunea ”neexplicată” a sumei totale a pătratelor, i.e. neexplicată de ecuația de regresie. Diferența TSS–RSS este porțiunea ”explicată” a sumei totale a pătratelor. **Coeficientul de determinare** R^2 este raportul dintre porțiunea ”explicată” și suma totală a pătratelor:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}.$$

Cu alte cuvinte, R^2 măsoară proporția variabilității datelor care este contabilizată pentru modelul de regresie. O valoare aproape de 1 indică o potrivire bună, în timp ce o valoare aproape de 0 indică o potrivire slabă. În cazul regresiei liniare simple, R^2 este pur și simplu pătratul coeficientului de corelație dintre valorile observate y_i și valorile prezise $\beta_0 + \beta_1 x_i$.

Exemplul 11. În exemplul 8 de suprapotrivire, valorile lui R^2 sunt ("degree"="grad"):

degree	R^2
1	0.3968
2	0.9455
9	1.0000

Măsura bunăității potrivirii crește când n (gradul) crește. Potrivirea este mai bună, dar modelul devine de asemenea mai complex, deoarece este nevoie de mai mulți coeficienți pentru a descrie polinoame de grad mai mare.