Curs 13

Cristian Niculescu

1 Intervale de încredere pentru media datelor nenormale

1.1 Scopurile învăţării

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

1.2 Introducere

Până acum, ne-am concentrat pe construirea intervalelor de încredere pentru date provenite dintr-o repartiţie normală. Acum vom învăţa despre intervale de încredere pentru medie când datele nu sunt neapărat normale.
Întâi vom studia estimarea probabilităţii θ de succes când datele provin din o repartiţie Bernoulli(θ) - reamintim că θ este de asemenea media repartiţiei Bernoulli.
Apoi vom considera cazul unei selecţii mari dintr-o repartiţie necunoscută; în acest caz putem apela la teorema limită centrală pentru a justifica intervalele de încredere z.

1.3 Datele Bernoulli şi votarea

O utilizare obişnuită a intervalelor de încredere este pentru estimarea proporţiei θ dintr-o repartiţie Bernoulli(θ). De exemplu, presupunem că vrem să folosim un sondaj politic pentru a estima proporţia din populaţie care susţine candidatul A, sau echivalent, probabilitatea θ ca o persoană aleatoare să susţină 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. $x_1, x_2, ..., x_n$ toate provenind dintr-o repartiţie Bernoulli(θ). Atunci un interval de încredere $(1-\alpha)$ normal conservator pentru θ este dat de

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

Demonstraţia dată mai jos foloseşte teorema limită centrală şi observaţia că $\sigma = \sqrt{\theta(1-\theta)} \leq 1/2$.
Veţi vedea de asemenea în deducerea de mai jos că această formulă este conservatoare, dând un interval de încredere "cel puţin $(1-\alpha)$".
Exemplul 1. Un sondaj întreabă 196 de persoane dacă preferă candidatul A candidatului B şi află că 120 preferă A şi 76 preferă B. Aflaţi intervalul de încredere normal conservator pentru θ, proporţia din populaţie care preferă A.
Răspuns. Avem $\overline{x} = 120/196 \approx 0.612$, $\alpha = 0.05$ şi $z_{0.025} \approx 1.96$. 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 Demonstraţia formulei (1)

Demonstraţia formulei (1) se va baza pe următoarea lemă.
Lemă. Deviaţia standard a unei repartiţii Bernoulli(θ) este cel mult 0.5.
Demonstraţia lemei. Notăm această deviaţie standard cu σ_θ pentru a sublinia dependenţa ei de θ. Atunci dispersia este $\sigma_\theta^2 = \theta(1-\theta)$. Este uşor de văzut folosind analiza matematică sau proprietăţile funcţiei 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 deviaţia standard σ este cel mult $\sqrt{1/4} = 1/2$.
Demonstraţia formulei (1). Demonstraţia se bazează pe teorema limită centrală care spune că (pentru n mare) repartiţia lui $\overline{x}$ este aproximativ normală cu media θ şi deviaţia standard $\sigma_\theta/\sqrt{n}$. Pentru date normale avem intervalul de încredere z $(1-\alpha)$

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

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

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

este cel puţin la fel de lung ca intervalul folosind $\pm\sigma_\theta/\sqrt{n}$. Un interval mai lung este mai probabil să conţină 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 deviaţia standard a lui $\overline{x}$, 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 puteţi auzi

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

Sensul precis al acestui fapt este
dacă θ este proporţia din populaţie care sprijină pe A, atunci estimarea punctuală pentru θ este 52% şi intervalul de încredere 95% este $52\% \pm 5\%$.
Observaţi că reporterii sondajelor la ştiri nu menţionează încrederea 95%. Doar va trebui să ştii 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

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

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

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

Exemplul 2. Votare. Presupunem că în curând vor fi alegeri locale între candidatul A şi candidatul B. Presupunem că fracţia din populaţia cu drept de vot care-l susţine pe A este θ.
2 organizaţii de sondaj întreabă votanţii pe cine preferă.
1. Firma *Rapid şi primul* sondează 40 de votanţi aleatori şi află că 22 sprijină pe A.
2. Firma *Rapid, dar precaut* sondează 400 de votanţi aleatori şi află că 190 sprijină pe A.
Aflaţi estimările punctuale şi intervalele de încredere cu regula degetului mare 95%. Explicaţi cum statisticile reflectă intuiţia că sondajul a 400 de votanţi este mai exact.
Răspuns. Pentru sondajul 1 avem

Estimare punctuală: $\overline{x} = 22/40 = 0.55$.
Interval de încredere: $\overline{x} \pm \frac{1}{\sqrt{n}} = 0.55 \pm \frac{1}{\sqrt{40}} = 0.55 \pm 0.16 = 55\% \pm 16\%$.
Pentru sondajul 2 avem
Estimare punctuală: $\overline{x} = 190/400 = 0.475$.
Interval de încredere: $\overline{x} \pm \frac{1}{\sqrt{n}} = 0.475 \pm \frac{1}{\sqrt{400}} = 0.475 \pm 0.05 = 47.5\% \pm 5\%$.
Precizia mai mare a sondajului cu 400 de votanţi este reflectată de marginea mai mică a erorii, i.e. 5% pentru sondajul cu 400 de votanţi vs. 16% pentru sondajul cu 40 de votanţi.

Intervale de încredere pentru proporţia binomială

Sunt multe metode de a produce intervale de încredere pentru proporţia p a unei repartiţii 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 selecţie mare

Un scop tipic în statistică este să estimăm media unei repartiţii. Când datele au o repartiţie normală am putea folosi intervale de încredere bazate pe statistici standardizate pentru a estima media.
Dar presupunem că datele $x_1, x_2, ..., x_n$ provin dintr-o repartiţie cu pmf sau pdf $f(x)$ care poate să nu fie normală sau chiar parametrică. Dacă repartiţia are medie şi dispersie finite şi 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, repartiţia de selecţie a mediei Studentizate este aproximativ normală standard: $\frac{\overline{x}-\mu}{s/\sqrt{n}} \approx N(0,1)$.
Deci, pentru n mare, intervalul de încredere $(1-\alpha)$ pentru μ este aproximativ

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

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

Exemplul 3. Cât de mare trebuie să fie n?

Reamintim că o eroare CI de tipul 1 apare când intervalul de încredere nu conţine 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 aşteptăm ca adevărata rată a erorii CI de tipul 1 să fie α.
Putem face simulări numerice pentru a aproxima adevăratul nivel de încredere. Ne aşteptăm că, atunci când n devine mai mare, adevăratul nivel de încredere al intervalului de încredere de dispersie mare să conveargă la valoarea nominală.

Facem astfel de simulări pentru x provenite de la repartiţia exponenţială exp(1) (care este departe de normală). Pentru câteva valori ale lui n şi nivelul de încredere nominal c facem 100000 de încercări. Fiecare încercare constă din următorii paşi:
1. Extragem n date din exp(1).
2. Calculăm media de selecţie $\overline{x}$ şi deviaţia standard de selecţie s.
3. Construim intervalul de încredere c de selecţie mare: $\overline{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$.
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 comparaţie facem aceleaşi teste pe date provenite dintr-o repartiţie 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 repartiţia exp(1) vedem că pentru $n = 20$ încrederea simulată a intervalului de încredere de selecţie mare este mai mică decât încrederea nominală $(1-\alpha)$. Dar pentru $n = 100$ încrederea simulată şi încrederea nominală sunt foarte aproape. Deci, pentru exp(1), n undeva între 50 şi 100 este destul de mare pentru cele mai multe scopuri.
Gândiţi. Pentru $n = 20$ de ce încrederea simulată pentru repartiţia $N(0,1)$ 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 diferenţă sesizabilă, de exemplu $z_{0.025} \approx 1.96$ şi $t_{0.025} \approx 2.09$.

2 Intervale de încredere bootstrap

2.1 Scopurile învăţării

1. Să poată construi şi selecta din repartiţia empirică a datelor.
2. Să poată explica principiul bootstrap.
3. Să poată proiecta şi folosi un bootstrap empiric pentru a calcula intervale de încredere.
4. Să poată proiecta şi 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 variaţia statisticilor care sunt calculate din aceleaşi 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 şi a demonstrat modul de aplicare al bootstrap-ului folosind efectiv calculatoarele. El a inventat de asemenea termenul "bootstrap".
Aplicaţia noastră principală a bootstrap-ului va fi estimarea variaţiei estimărilor punctuale; adică, estimarea intervalelor de încredere.
Exemplul 1. Presupunem că avem datele

$$x_1, x_2, ..., x_n.$$

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

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

este un interval de încredere 95% pentru μ.
Acum presupunem că datele provin dintr-o repartiţie complet necunoscută. Pentru a avea un nume, vom numi această repartiţie F şi media ei (necunoscută) μ. Putem folosi încă media de selecţie $\overline{x}$ ca o estimare punctuală a lui μ. Dar cum putem afla un interval de încredere pentru μ în jurul lui $\overline{x}$? Răspunsul nostru va fi să folosim bootstrap-ul!
De fapt, bootstrap-ul tratează alte statistici la fel de uşor cum tratează media. De exemplu: mediana sau alte percentile. Acestea sunt statistici unde,

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

2.3 Selectarea

În statistică, a selecta dintr-o mulţime însemnă a alege elemente din acea mulţime. Într-o selecţie aleatoare, elementele sunt alese aleator. Sunt 2 metode uzuale pentru selectare aleatoare.

Selectare fără înlocuire

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

Selectare cu înlocuire

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

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

Exemplul 2. Putem vedea aruncarea repetată a unui zar cu 8 feţe ca selectarea cu înlocuire din mulţimea $\{1, 2, 3, 4, 5, 6, 7, 8\}$. 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.

Observaţie. În practică, dacă luăm un număr mic dintr-o mulţime 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 aceeaşi persoană va fi aleasă de 2 ori încât nu este o diferenţă reală între selectarea cu sau fără înlocuire.

2.4 Repartiţia empirică a datelor

Repartiţia empirică a datelor este pur şi simplu repartiţia pe care o vedeţi în date.

Exemplul 3. Presupunem că aruncăm un zar cu 8 feţe de 10 ori şi obţinem următoarele date, scrise în ordine crescătoare:

$$1,\ 1,\ 2,\ 3,\ 3,\ 3,\ 3,\ 4,\ 7,\ 7.$$

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

un 3 este 4/10 şi probabilitatea de a trage un 4 este 1/10. Repartiţia 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

Notaţie. Dacă notăm adevărata repartiţie din care provin datele cu F, atunci vom nota repartiţia 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, repartiţiile adevărată şi 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 repartiţie F şi repartiţia empirică F^* a unui zar cu 8 feţe. Pe prima linie, în loc de al 2-lea "5" se va citi "6".

Deoarece F^* este obţinută strict din date, o numim repartiţia empirică a datelor. O vom numi de asemenea repartiţia de reselectare. Totdeauna ştim F^* explicit. În particular media lui F^* este chiar media de selecţie $\overline{x}$.

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 selecţie luată dintr-o repartiţie subiacentă. A reselecta este a selecta cu înlocuire din repartiţia empirică, de exemplu a pune aceste 10 numere într-o pălărie şi a trage unul aleator. Apoi puneţi numărul înapoi în pălărie şi trageţi iar. Trageţi atâtea numere ca mărimea dorită a reselecţiei. Pentru a ajunge un pic mai aproape de a implementa aceasta pe un calculator, reformulăm aceasta în modul următor. Notăm cele 10 date cu $x_1, x_2, ..., x_{10}$. A reselecta este a trage un număr j din repartiţia uniformă pe $\{1, 2, ..., 10\}$ şi a lua x_j ca valoarea noastră reselecţionată. În acest caz am putea face aceasta aruncând un zar cu 10 feţe. De exemplu, dacă aruncăm un 6, atunci valoarea noastră reselectată este 3, al 6-lea element din lista noastră.

Dacă vrem o mulţime de date reselectate de mărime 5, atunci vom arunca

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

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

atunci reselecţia este

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

Observaţii. 1. Deoarece selectăm cu înlocuire, aceeaşi dată poate apărea de mai multe ori când reselectăm.
2. De asemenea, deoarece selectăm cu înlocuire, putem avea o mulţime de date reselectate de orice mărime vrem, de exemplu am putea reselecta de 1000 de ori.
Desigur, în practică se foloseşte un pachet software ca `R` pentru a face reselectarea.

2.5.1 Notaţia cu steluţă

Dacă avem o selecţie de mărime n

$$x_1, x_2, ..., x_n,$$

atunci notăm o reselecţie de mărime m adăugând o steluţă simbolurilor

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

Similar, la fel cum $\overline{x}$ este media datelor originale, scriem $\overline{x}^*$ pentru media datelor reselecţionate.

2.6 Bootstrap-ul empiric

Presupunem că avem n date

$$x_1, x_2, ..., x_n,$$

provenite dintr-o repartiţie F. O selecţie bootstrap empirică este o reselecţie de aceeaşi mărime n:

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

Ar trebui să gândiţi ultima ca o selecţie de mărime n extrasă din repartiţia empirică F^*. Pentru orice statistică v calculată din datele selecţiei originale, putem defini o statistică v^* cu aceeaşi formulă, dar calculată folosind datele reselectate în locul celor originale. Cu această notaţie putem formula principiul bootstrap.

2.6.1 Principiul bootstrap

Schema bootstrap este după cum urmează:
1. $x_1, x_2, ..., x_n$ este o selecţie de date provenite dintr-o repartiţie F.
2. O statistică u este calculată din selecţie.
3. F^* este repartiţia empirică a datelor (repartiţia de reselectare).
4. $x_1^*, x_2^*, ..., x_n^*$ este o reselecţie a datelor de aceeaşi mărime ca selecţia originală.
5. u^* este statistica calculată din reselecţie.
Atunci principiul bootstrap spune că
1. $F^* \approx F$.
2. Variaţia lui u este bine aproximată de variaţia lui u^*.
Interesul nostru real este în punctul 2: putem aproxima variaţia lui u prin variaţia lui u^*. Vom exploata aceasta pentru a estima mărimea intervalelor de încredere.

2.6.2 De ce reselecţia are aceeaşi mărime ca selecţia originală?

Aceasta este direct: variaţia statisticii u va depinde de mărimea selecţiei. Dacă vrem să aproximăm această variaţie trebuie să folosim reselecţii de aceeaşi mărime.

2.6.3 Exemplu didactic de un interval de încredere bootstrap empiric

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

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

Problema: Estimaţi media μ a repartiţiei subiacente şi daţi un interval de încredere bootstrap 80%.
Răspuns. Media de selecţie este $\overline{x}$=40.3. Folosim aceasta ca o estimare a adevăratei medii μ a repartiţiei subiacente. Ca în exemplul 1, pentru a face intervalul de încredere trebuie să ştim cât de mult repartiţia lui $\overline{x}$ variază în jurul lui μ. Adică, ne-ar plăcea să ştim repartiţia lui

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

Dacă am şti această repartiţie, am putea afla $\delta_{.1}$ şi $\delta_{.9}$, valorile critice 0.1 şi 0.9 ale lui δ. Atunci am avea

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

ceea ce dă intervalul de încredere 80%

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

Ca întotdeauna la intervalele de încredere, ne-am grăbit să subliniem că probabilităţile calculate mai sus sunt probabilităţile privind statistica $\overline{x}$ dat fiind că adevărata medie este μ.
Principiul bootstrap dă o abordare practică a estimării distribuţiei lui $\delta = \overline{x} - \mu$. Spune că o putem aproxima prin repartiţia lui

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

unde $\overline{x}^*$ este media unei selecţii 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 repartiţia lui δ^* cu mare precizie.
Acum să ne întoarcem la datele de selecţie cu 10 puncte. Am folosit `R` pentru a genera 20 de selecţii bootstrap, fiecare de mărime 10. Fiecare dintre cele 20 de coloane ale următorului tabel este o selecţie 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 $\delta^* = \overline{x}^* - \overline{x}$ pentru fiecare selecţie bootstrap (i.e. fiecare coloană) şi le sortăm de la cea mai mică la cea mai mare:

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

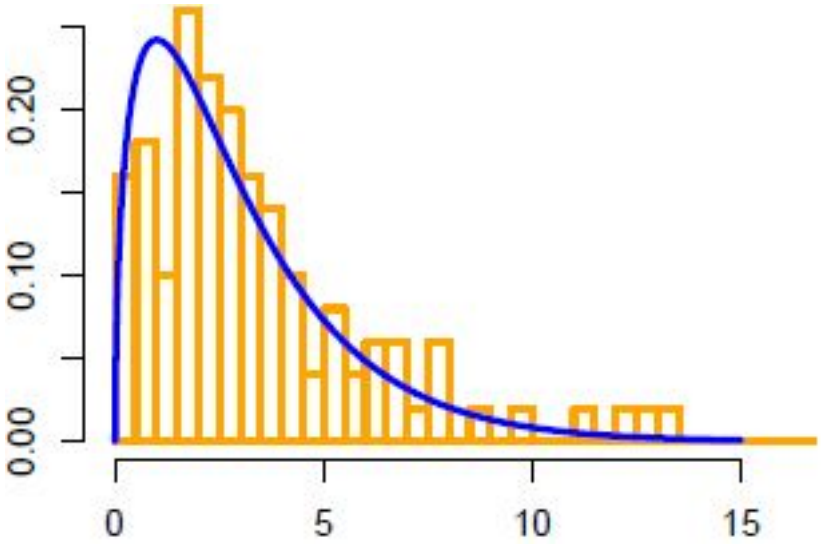
Vom aproxima valorile critice $\delta_{.1}$ şi $\delta_{.9}$ prin $\delta^*_{.1}$ şi $\delta^*_{.9}$. Deoarece $\delta^*_{.1}$ este la a 90-a percentilă, alegem al 18-lea element din listă, i.e. 1.6. Analog, deoarece $\delta^*_{.9}$ este la a 10-a percentilă, alegem al 2-lea element din listă, i.e. -1.4.
De aceea intervalul nostru de încredere 80% bootstrap pentru μ este

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

În acest exemplu am generat doar 20 de selecţii bootstrap pentru ca să încapă pe pagină. Folosind `R`, am putea genera cel puţin 10000 de selecţii bootstrap pentru a obţine o estimare foarte precisă pentru $\delta^*_{.1}$ şi $\delta^*_{.9}$.

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, repartiţia empirică va fi o bună aproximare a adevăratei repartiţii. Vizual, spune că histograma datelor ar aproxima densitatea adevăratei repartiţii.
Reselectarea nu poate îmbunătăţi estimarea noastră punctuală. De exemplu, dacă estimăm media μ prin $\overline{x}$, atunci în bootstrap calculăm $\overline{x}^*$ pentru multe reselectări ale datelor. Dacă luăm media tuturor $\overline{x}^*$, ne-am aştepta să fie foarte aproape de $\overline{x}$. Aceasta nu ne-ar spune nimic nou despre adevărata valoare a lui μ.
Chiar cu o cantitate corectă de date, potrivirea dintre repartiţiile adevărată şi empirică nu este perfectă, deci va fi o eroare în estimarea mediei (sau oricărei alte valori). Dar cantitatea de variaţie a estimărilor este mult mai puţin sensibilă la diferenţe între densitate şi histogramă. Atât timp cât sunt rezonabil de aproape, atât repartiţia adevărată, cât şi cea empirică vor avea cantităţi similare de variaţie. Deci, în general principiul bootstrap este mai robust când aproximăm repartiţia variaţiei relative decât când aproximăm repartiţii absolute.
Repartiţia (peste diferite seturi de date experimentale) a lui $\overline{x}$ este "centrată" în μ şi repartiţia lui $\overline{x}^*$ este centrată în $\overline{x}$. Dacă este o separare semnificativă între $\overline{x}$ şi μ, atunci aceste 2 repartiţii vor diferi de asemenea semnificativ. Pe de altă parte, repartiţia lui $\delta = \overline{x} - \mu$ descrie variaţia lui $\overline{x}$ în jurul centrului lui. Analog, repartiţia lui $\delta^* = \overline{x}^* - \overline{x}$ descrie variaţia lui $\overline{x}^*$ în jurul centrului lui. Deci chiar dacă cele 2 centre sunt destul de diferite, cele 2 variaţii în jurul centrelor pot fi aproximativ egale.
Figura de mai jos ilustrează cum repartiţia empirică aproximează adevărata repartiţie. Pentru a face figura au fost generate 100 de valori aleatoare dintr-o repartiţie χ^2 cu 3 grade de libertate. Figura arată pdf-ul adevăratei repartiţii ca o linie albastră şi histograma repartiţiei empirice cu portocaliu.

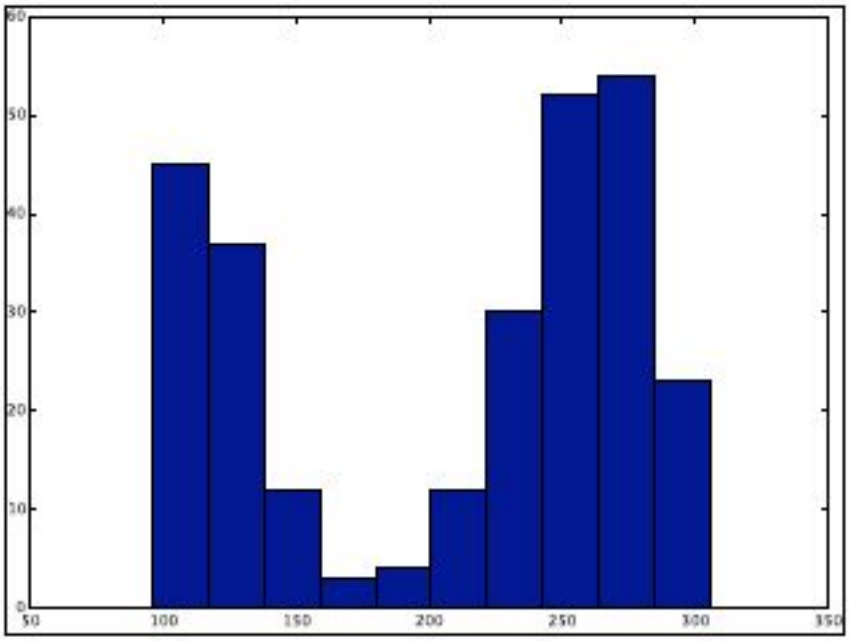


Repartiţiile adevărată şi empirică sunt aproximativ egale.

2.7 Alte statistici

Până acum am evitat intervalele de încredere pentru mediană şi alte statistici deoarece repartiţiile lor de selecţie 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 naţional Yellowsone din Wyoming: `http://en.wikipedia.org/wiki/Old_Faithful`.
Există un set de date care dă duratele a 272 de erupţii consecutive. Iată histograma datelor.



Întrebare. Estimaţi lungimea mediană a unei erupţii şi daţi un interval de încredere de 90% pentru mediană.

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

1. Datele: $x_1, ..., x_{272}$.
2. Mediana datelor: $x_{\text{median}} = 240$.
3. Află mediana x^*_{median} a unei selecţii bootstrap $x^*_1, ..., x^*_{272}$. Repetă de 1000 de ori.
4. Calculează diferenţele bootstrap

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

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

5. Principiul bootstrap spune că putem folosi $\delta^*_{.95}$ şi $\delta^*_{.05}$ ca estimări pentru $\delta_{.95}$ şi $\delta_{.05}$. 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 selecţii bootstrap, o nouă simulare plecând de la aceleaşi date de selecţie va produce un interval similar. Dacă în pasul 3 creştem numărul de selecţii bootstrap la 10000, atunci intervalele produse prin simulare vor varia chiar mai puţin. O strategie uzuală este să creştem numărul de selecţii bootstrap până simulările rezultate produc intervale care variază mai puţin decât un nivel acceptabil.

Exemplul 8. Folosind datele Old Faithful, estimaţi $P(|\overline{x} - \mu| > 5|\mu)$.
Răspuns. Procedăm exact ca în exemplul precedent folosind media în locul medianei.
1. Datele: $x_1, ..., x_{272}$.
2. Media datelor: $\overline{x} = 209.27$.
3. Află media $\overline{x}^*$ a 1000 de selecţii bootstrap empirice: $x_1^*, ..., x_{272}^*$.
4. Calculează diferenţele bootstrap

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

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

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

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

2.8 Bootstrap parametric

Exemplele din secţiunea anterioară au folosit toate bootstrap-ul empiric, care nu face nicio presupunere despre repartiţia subiacentă şi extrage selecţii bootstrap prin reselectarea datelor. În această secţiune vom aborda bootstrap-ul parametric. Singura diferenţă dintre bootstrapul parametric şi empiric este sursa selecţiei bootstrap. Pentru bootstrap-ul parametric, generăm selecţia bootstrap dintr-o repartiţie parametrizată.
Iată elementele folosirii bootstrap-ului parametric pentru a estima un interval de încredere pentru un parametru:
0. Date: $x_1, ..., x_n$ extrase dintr-o repartiţie $F(\theta)$ cu parametru necunoscut θ.
1. O statistică $\hat{\theta}$ care estimează pe θ.
2. Selecţiile noastre bootstrap sunt extrase din $F(\hat{\theta})$.
3. Pentru fiecare selecţie bootstrap

$$x_1^*, ..., x_n^*$$

calculăm $\hat{\theta}^*$ şi diferenţa bootstrap $\delta^* = \hat{\theta}^* - \hat{\theta}$.
4. Principiul bootstrap spune că repartiţia lui δ^* aproximează repartiţia lui $\delta = \hat{\theta} - \theta$.
5. Folosiţi diferenţele bootstrap pentru a face un interval de încredere bootstrap pentru θ.
Exemplul 9. Presupunem că datele $x_1, ..., x_{300}$ sunt extrase dintr-o repartiţie

exp(λ). Presupunem de asemenea că media datelor $\overline{x} = 2$. Estimaţi λ şi daţi un interval de încredere bootstrap parametric 95% pentru λ.

Răspuns. Cel mai uşor este să explicăm soluţia folosind codul `R` comentat.

```
# Bootstrap parametric
# Sunt 300 de date cu media 2.
# Presupunem că datele sunt exp(lambda).
# PROBLEMA: Calculaţi un interval de încredere bootstrap parametric
95% pentru lambda.
# Se dau numărul datelor şi media.
n=300
xbar=2
# MLE pentru lambda este 1/xbar
lambdahat=1/xbar
# Generăm selecţiile bootstrap.
# Fiecare coloană este o selecţie bootstrap (de 300 de valori reselecţionate).
nboot=1000
#Iată diferenţa cheie faţă de bootstrap-ul empiric:
# Extragem selecţia bootstrap din exponenţiala(lambdahat).
x=rexp(n*nboot,lambdahat)
bootstrapsample=matrix(x,nrow=n,ncol=nboot)
# Calculăm lambdastar bootstrap.
lambdastar=1/colMeans(bootstrapsample)
# Calculăm diferenţele.
deltastar=lambdastar-lambdahat
# Aflăm cuantilele 0.05 şi 0.95 pentru deltastar.
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
ieşire.
# R are şi 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 selecţii bootstrap. În practica reală ar fi generate mult mai multe selecţii 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 selecţie
xbar=mean(x)
nboot=20
# Generăm 20 de selecţii bootstrap, i.e.  o matrice n x 20 de reselecţii
aleatoare din x.
tmpdata=sample(x,n*nboot,replace=TRUE)
bootstrapsample=matrix(tmpdata,nrow=n,ncol=nboot)
# Calculăm mediile x*.
bsmeans=colMeans(bootstrapsample)
# Calculăm δ* pentru fiecare selecţie bootstrap.
deltastar=bsmeans-xbar
# Sortăm rezultatele.
sorteddeltastar=sort(deltastar)
# Aflăm valorile critice .1 şi .9 ale lui deltastar.
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 învăţării

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 şi heteroscedasticitate.

3.2 Introducere

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

obţinem:

```
Call:
lm(formula = y ∼ x)
Coefficients:
(Intercept) x
-0.1324 0.8791.
```

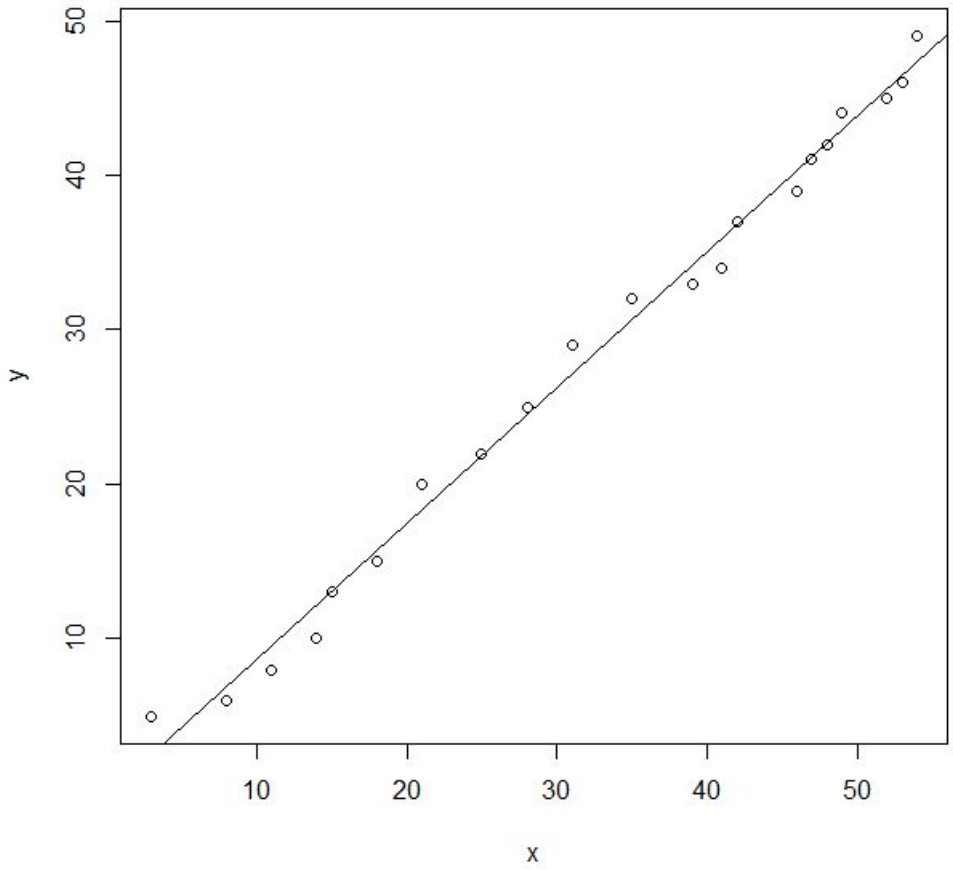
Am aflat că dreapta care dă "potrivirea celor mai mici pătrate" cu aceste date (dreapta de regresie) este

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

unde x este numărul de ani de la 1960, iar y este în cenţi.
Folosind acest rezultat "prezicem" că în 2023 ($x = 63$), costul unui timbru va fi 55 de cenţi (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)
```

obţinem:



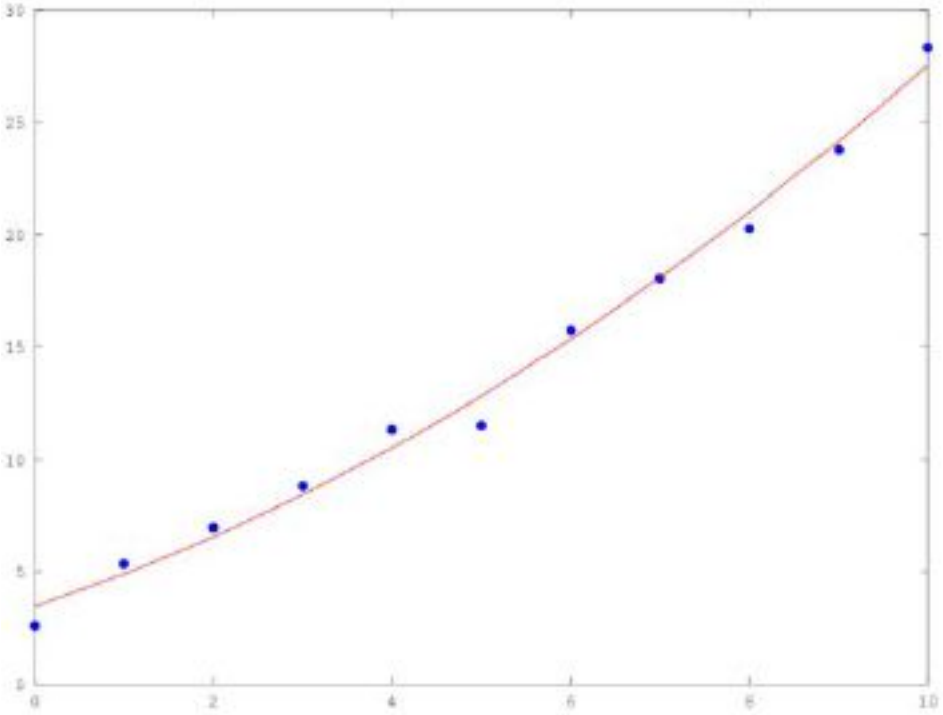
Costul timbrului (cenţi) vs. timp (ani din 1960)

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 taţi şi fii adulţi. Fie x_i şi y_i înălţimile celui de-al i-lea tată, respectiv fiu. Dreapta celor mai mici pătrate

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

Exemplul 3. Nu suntem limitaţi 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

$$y = \beta_1 x + \beta_0,$$

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

$$y_i = \beta_1 x_i + \beta_0 + \epsilon_i.$$

Deci,

$$\epsilon_i = y_i - \beta_1 x_i - \beta_0.$$

Metoda celor mai mici pătrate află valorile $\hat{\beta}_0$ şi $\hat{\beta}_1$ ale lui β_0 şi 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

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}},\ \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \tag{2}$$

unde

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

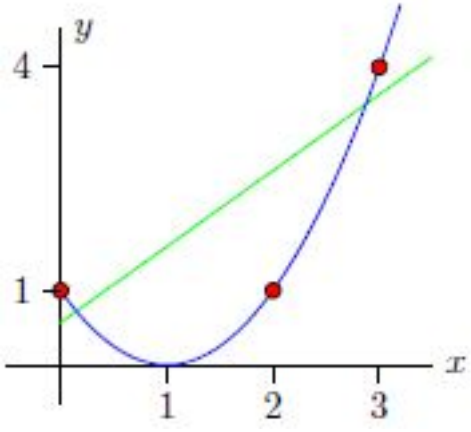
Aici, $\overline{x}$ este media de selecţie a lui x, $\overline{y}$ este media de selecţie a lui y, s_{xx} este dispersia de selecţie a lui x şi s_{xy} este covarianţa de selecţie a lui x şi y.

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

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

$$\overline{x} = \frac{1}{3}(x_1 + x_2 + x_3) = \frac{1}{3}(0 + 2 + 3) = \frac{5}{3},$$
$$\overline{y} = \frac{1}{3}(y_1 + y_2 + y_3) = \frac{1}{3}(1 + 1 + 4) = \frac{6}{3} = 2,$$
$$s_{xx} = \frac{1}{3-1}\sum_{i=1}^{3} (x_i - \overline{x})^2 = \frac{1}{2}\left[\left(0 - \frac{5}{3}\right)^2 + \left(2 - \frac{5}{3}\right)^2 + \left(3 - \frac{5}{3}\right)^2\right] = \frac{7}{3},$$
$$s_{xy} = \frac{1}{2}\left[\left(0 - \frac{5}{3}\right)(1 - 2) + \left(2 - \frac{5}{3}\right)(1 - 2) + \left(3 - \frac{5}{3}\right)(4 - 2)\right] = 2;$$
$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{2}{\frac{7}{3}} = \frac{6}{7};$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 2 - \frac{6}{7} \cdot \frac{5}{3} = 2 - \frac{10}{7} = \frac{4}{7}.$$

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



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

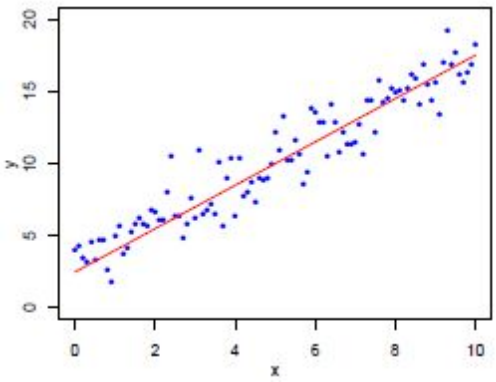
Regresie liniară simplă: Este puţin confuz, dar cuvântul "liniară" din "regresie liniară" nu se referă la potrivirea unei drepte. Totuşi, cea mai uzuală curbă pentru potrivire este o dreaptă. Potrivirea unei drepte la date bivariate este numită regresie liniară simplă.

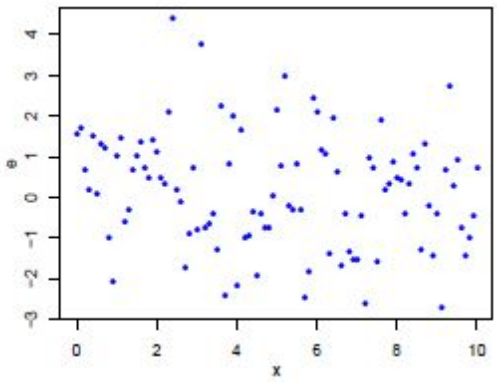
3.3.1 Reziduuri

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.

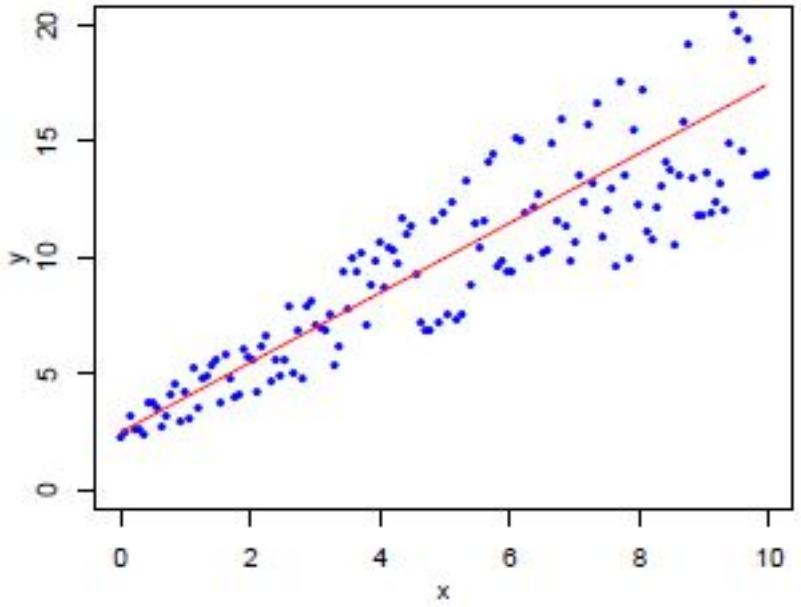




Date cu dreapta de regresie (stânga) şi reziduuri (dreapta). Observaţi homoscedasticitatea.

3.3.2 Homoscedasticitatea

O presupunere importantă a modelului de regresie liniară este că reziduurile ϵ_i au aceeaşi dispersie $\forall i$. Această presupunere este numită homoscedasticitate. Puteţi vedea aceasta în cazul ambelor figuri de mai sus. Datele sunt în banda de lăţime fixă în jurul dreptei de regresie şi la fiecare x reziduurile au cam aceeaşi împrăştiere verticală.
Mai jos este o figură arătând date heteroscedastice. Împrăştierea verticală a datelor creşte când x creşte. 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 ecuaţiile algebrice liniare pentru parametrii necunoscuţi β_i, i.e. fiecare β_i are exponentul 1.
Exemplul 5. Luaţi aceleaşi date ca în exemplul 4 şi folosiţi 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 şi y_i putem folosi analiza matematică (egalăm derivatele parţiale în raport cu $\beta_0, \beta_1, \beta_2$ cu 0, obţinând un sistem de 3 ecuaţii 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).
```

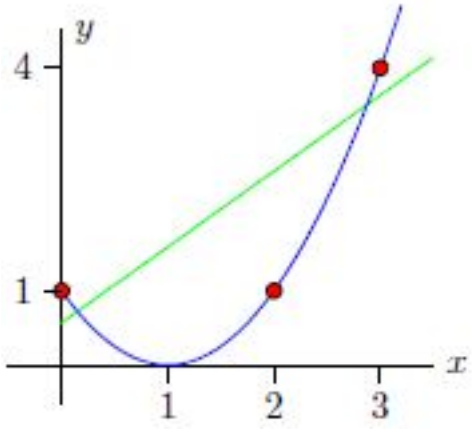
Sau codul `R`

```
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 ecuaţia

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

Pentru 3 puncte, potrivirea pătratică este perfectă.



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

Exemplul 6. Perechile (x_i, y_i) pot da vârsta şi 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 relaţia este exponenţială, i.e. $y = ce^{ax}$. Atunci

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

Deci putem folosi regresia liniară simplă pentru a obţine un model

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

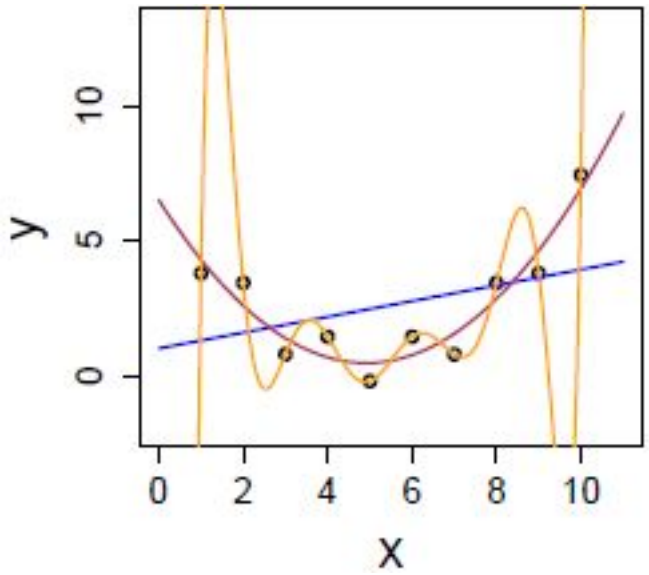
şi apoi obţinem modelul exponenţial

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

3.4.1 Suprapotrivirea

Putem totdeauna obţine 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 relaţie între x şi y. Un model suprapotrivit va potrivi datele originale mai bine, dar va prezice mai puţin 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 şi 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 aşteptăm 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 Funcţia `R lm`

Nu facem regresia liniară cu mâna. Regresia liniară se reduce la rezolovarea sistemelor de ecuaţii liniare, i.e. la calcul matriceal. Funcţia `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 funţii, exceptând polinoamele, după cum puteţi explora folosind ajutorul lui `R` sau google.

3.5 Regresie liniară multiplă

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

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

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

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

Aici fiecare x_j este o variabilă predictor şi y este variabila de răspuns. De exemplu, putem fi interesaţi de cum variază o populaţie de peşti în funcţie nivelele măsurate ale câtorva poluanţi, sau am vrea să prezicem înălţimea de adult a unui fiu pe baza înălţimilor tatălui şi 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 şi deviaţia standard σ. Presupunerea standard este că ε_i sunt i.i.d. cu repartiţia $N(0, \sigma^2)$. 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 şi β_1 care minimizează dispersia de selecţie 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 toţi coeficienţii 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 faţă de media ei. Aici "mai aproape" este în ghilimele deoarece trebuie să controlăm scala (i.e deviaţia standard) fiecărei variabile. Modul de a controla scala este mai întâi să standardizăm fiecare variabilă.

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

Standardizarea schimbă media în 0 şi dispersia în 1:

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

Proprietăţile algebrice ale covarianţei arată că

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

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

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

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

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

i.e., y regresează la $\overline{y}$. Standardizarea are grijă de scală.
Considerăm cazul extrem al corelaţiei 0 între x şi y. Atunci, indiferent de valoarea lui x, valoarea prezisă a lui y este totdeauna $\overline{y}$. Adică, y a regresat până la media lui.
Dreapta de regresie trece totdeauna prin punctul $(\overline{x}, \overline{y})$.
Exemplul 9. Regresia la medie este importantă în studiile longitudinale. Rice (*Mathematical Statistics and Data Analysis*) dă următoarul exemplu. Presupunând că li se dau copiilor un test IQ la vârsta de 4 ani şi altul la vârsta de 5 ani, ne aşteptăm 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 îmbunătăţire (i.e. regresează la medie) în al 2-lea test. Astfel, o intervenţie inutilă poate fi interpretată greşit ca utilă deoarece pare a îmbunătăţi scorurile.
Exemplul 10. Alt exemplu cu consecinţe practice este recompensa şi pedeapsa. Imaginaţi-vă o şcoală unde performanţa înaltă la un examen este recompensată şi performanţa slabă este pedepsită. Regresia la medie ne spune că (în medie) studenţii foarte performanţi vor face puţin mai slab la următorul examen şi studenţii puţin performanţi vor face puţin mai bine. O viziune nesofisticată a datelor va face să pară că pedeapsa a îmbunătăţit performanţa şi recompensa de fapt a scăzut performanţa. Sunt reale consecinţe dacă cei cu autoritate acţionează după această idee.

3.8 Adaos

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

Cea mai directă demonstraţie 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 parţiale (şi reamintind că x_i şi 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 obţine următorul sistem de 2 ecuaţii liniare în necunoscutele β_0 şi 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 obţinem formulele (2).
Pentru multe aplicaţii între discipline vezi:
`http://en.wikipedia.org/wiki/Linear_regression#Applications_of_linear_regression`.

3.8.2 Măsurarea potrivirii

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

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

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 porţiunea "neexplicată" a sumei totale a pătratelor, i.e. neexplicată de ecuaţia de regresie. Diferenţa TSS−RSS este porţiunea "explicată" a sumei totale a pătratelor. Coeficientul de determianre R^2 este raportul dintre porţiunea "explicată" şi suma totală a pătratelor:

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

Cu alte cuvinte, R^2 măsoară proporţia variabilităţii 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 şi simplu pătratul coeficientului de corelaţie dintre valorile observate y_i şi 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 bunătăţii potrivirii creşte când n (gradul) creşte. Potrivirea este mai bună, dar modelul devine de asemenea mai complex, deoarece este nevoie de mai mulţi coeficienţi pentru a descrie polinoame de grad mai mare.