PROIECT PROBABILITĂȚI ȘI STATISTICĂ
STACKLOSS

EXERCIȚIUL 1

Din documentația din RStudio despre setul de date stackloss, am aflat că variabilele Air Flow, Water Temperature și Acid Concentration sunt variabile independente, în vreme ce Stack Loss este variabilă dependenă. Astfel, Stack Loss va fi caracteristica despre care o să realizăm o estimare în funcție de celelalte caracteristici.

1. Analizând box plot-ul fiecărei variabile am observat:
    - Variabilele Water Temperature și Acid Concentration nu au outliers ducând la concluzia că toate datele sunt situate în apropierea celorlalte.
    - Variabila Air Flow are 3 outliers: 80, 80, 75, acestea nefiind suficient de apropiate de celelalte date. Concluzionăm că ele au fost observate printr-o greșeală / eroare voită.
    - Variabila Stack Loss are 3 outliers: 42, 37, 37, acestea nefiind suficient de apropiate de celelalte date. Concluzionăm că ele au fost observate printr-o greșeală / eroare voită.
2. Analizand scatter plot-ul pentru fiecare relație liniară dintre variabilele independente (Air Flow, Water Temperature, Acid Concentration) și variabila independent (Stack Loss), am observat:
    - Punctele din scatter plot-ul definit de relația Acid Concentration și Stack Loss sunt mult prea împrăștiate pentru a se forma o regresie liniară.
    - Punctele din scatter plot-ul definit de relația Air Flow și Stack Loss realizează o formă suficient de dreaptă pentru a se justifica utilizarea unei regresii liniare.
    - Punctele din scatter plot-ul definit de relația Water Temperature și Stack Loss realizează o formă suficient de dreaptă pentru a se justifica utilizarea unei regresii liniare.
3. Analizând density plot-ul fiecărei variabile, am concluzionat că doar caracteristica Acid Concentration prezintă o distribuție similară cu o variabila aleatoare continuă distribuită normal.
4. Media, varianța și quartilele.

	Media	Varianța	Quartilele		
			Prima quartilă	A doua quartilă	A treia quartilă
Air Flow	60.428571	84.057143	56	58	62
Water Temp.	21.095238	9.990476	18	20	24
Acid Conc.	86.285714	28.714286	82	87	89
Stack Loss	17.523810	103.461905	11	15	19

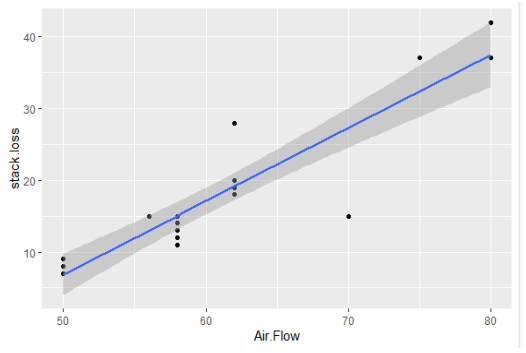
EXERCIȚIUL 2

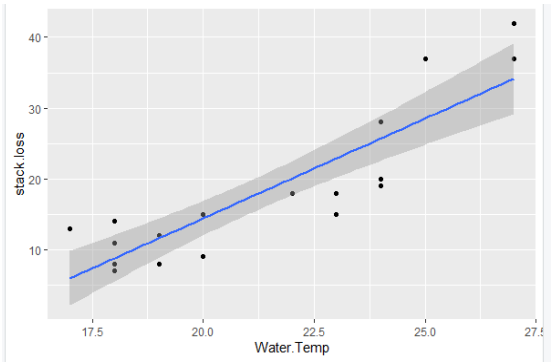
- REGRESIE LINIARĂ SIMPLĂ

Așa cum am arătat la exercițiul 1, la analiza scatter plot-ului, folosirea ca predictor a variabilei independente Acid Concentration nu este potrivită pentru realizarea unei regresii liniare simple cu semnificație statistică. Așadar, vom lua în considerare celelalte doua relații, anume Air Flow – Stack Loss și Water Temperature – Stack Loss.

Calculând corelațiile corespunzătoare celor două relații, am observat că corelația Air Flow – Stack Loss este 0.9196635 care este mai mare decât corelația Water Temperature – Stack Loss a cărei valoare e 0.8755044. Totodată Pr(>|t|) pentru intercept pentru ambele regresii este < 0.05, acest aspect neajutându – ne în a lua o decizie. Astfel, examinând valoarea R-squared din summary, am observat că valoarea pentru prima regresie Air Flow – Stack Loss este egală cu 0.84, care este mai mare decât a celei de-a doua regresii Water Temperature – Stack Loss a cărei valoare este 0.76. De asemenea, examinând Residual Standard Error, am observat că valoarea pentru prima regresie Air Flow – Stack Loss este 0.23, care e mai mica decât 0.28 care reprezintă valoarea celei de-a doua regresii Water Temperature – Stack Loss.

Concluzionăm că, din examinarea corelației, a valorii R-squared(cu cât e mai mare, cu atât mai bine) și a valorii Residual Standard Error(cu cât e mai mica, cu atât mai bine) , alegerea ca predictor – Air.Flow este cea mai bună. Astfel, vom scrie regresia liniară: *stack.loss = -44.13202 + 1.02031 * Air.Flow*





- REGRESIE LINIARĂ MULTIPLĂ

Vrem să construim un model de regresie liniară multiplă pentru a estima pierderea amoniacului de pe urma oxidării unei plante bazându-ne inițial pe fluxul de aer, temperatura apei și concentrația de acid.

Astfel, variabila răspuns este Stack Loss, iar variabilele predictor sunt Air Flow, Water Temperature și Acid Concentration.

Cu ajutorul funcției summary, am observant că variabila predictor Acid Concentration nu are o legătură seminificativă între ea și variabila de răspuns Stack Loss. Vom construi un model de regresie liniară multiplă, luând în calcul doar variabilele predictor Water Temperature și Air Flow.

```
             Estimate Std. Error    t value     Pr(>|t|)
(Intercept) -39.9196744 11.8959969 -3.3557234 3.750307e-03
Air.Flow      0.7156402  0.1348582  5.3066130 5.799025e-05
Water.Temp    1.2952861  0.3680243  3.5195672 2.630054e-03
Acid.Conc.   -0.1521225  0.1562940 -0.9733098 3.440461e-01
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -50.3588401  5.1383281 -9.800628 1.216471e-08
Air.Flow      0.6711544  0.1266910  5.297568 4.897970e-05
Water.Temp    1.2953514  0.3674854  3.524905 2.419146e-03
```

Imaginile de mai sus arată că variabila Acid Concentration **NU** este un predictor bun. Astfel, vom scrie regresia liniară multiplă: *stack.loss = -50.3588 + 0.6712 * Air.Flow + 1.2954 * Water.Temp*

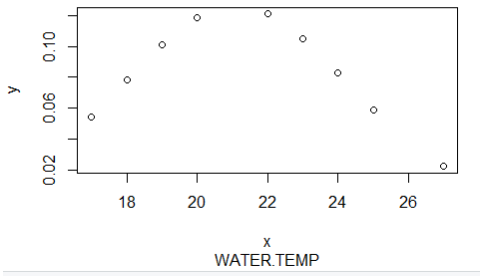
- INTERPRETARI DESPRE REZULTATELE OBTINUTE IN URMA EVALUARII CELOR DOUA MODELE DE REGRESIE

Comparând modelul de regresie liniară simplă construit anterior și modelul de regresie liniară multiplă construit anterior, constatăm că regresia liniară multiplă este mai potrivită pentru setul nostru de date deoarece:

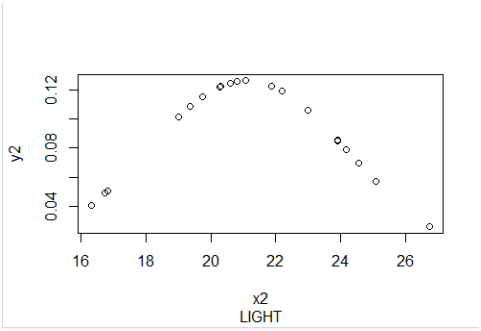
- Analizând R-squared (specificată în summary model): pentru regresia liniară simplă are un rezultat egal cu 0.8458, iar pentru regresia liniară multiplă are un rezultat egal cu 0.9088, care este mai mare decat 0.8458 => regresia liniară multiplă este mai potrivită pentru setul nostru de date.
- Analizând Residual Standard Error (dată prin formula: *sigma(model) / mean (stackloss$stack.loss))*: pentru regresia liniară simplă are un rezultat egal cu 0.23, iar pentru regresia liniară multiplă are un rezultat egal cu 0.18, care este mai mic decât 0.23 => regresia liniară multiplă este mai potrivită pentru setul nostru de date.
- Analizând Akaike's 'An Information Criterion' (AIC) : pentru regresia liniară simplă are un rezultat egal cu 122.7371, iar pentru regresia liniară multiplă are un rezultat egal cu 113.7144, care este mai mic decât 122.7371 => regresia liniară multiplă este mai potrivită pentru setul nostru de date.

- ADĂUGAREA UNEI NOI VARIABILE

Am decis să adăugăm o nouă variabila numită *light*, care reprezintă nivelul de luminozitate asupra plantei, ea potrivindu-se cu variabila Water Temperature. De pe urma analizei graficului densității variabilei Water Temperature se observă că acesta seamănă cu graficul densității distribuției normale.



Astfel vom genera variabila *light* folosind repartiția normală, având media și deviația standard a variabilei Water Temperature.

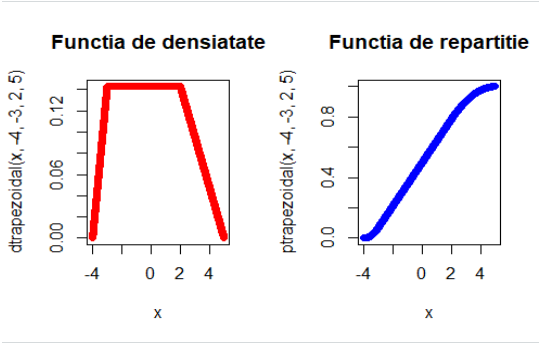


EXERCIȚIUL 3

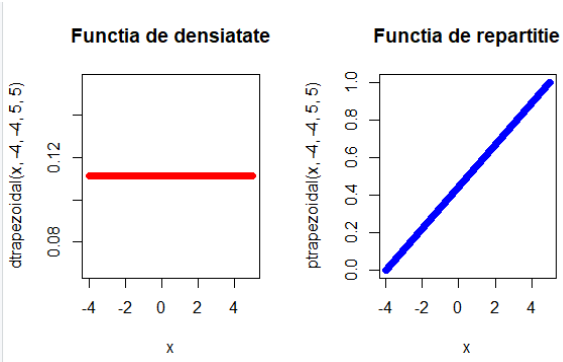
Pentru realizarea acestui exercițiu, am ales repartiția trapezoidală.

Repartiția trapezoidală este o repartiție de probabilitate continuă care prezintă următoarele proprietăți:

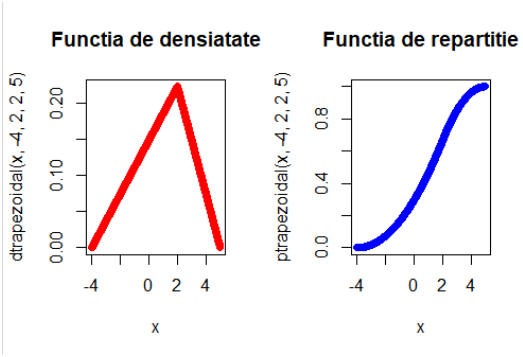
- Graficul realizat de funcția de densitate de probabilitate formează un trapez.
- Are o limită inferioară, notată *a*, și o limita superioară, notată *d*, în afara intervalului [a, d] neputând apărea alte evenimente. În plus, există 2 puncte non-diferențiabile și discontinue în funcția de densitate de probabilitate numite *b* și *c* care apar între *a* si *d* astfel încat $a \leq b \leq c \leq d$.
- Putem spune că *a, b, c, d* reprezintă coordonatele pe axa OX a "vârfurilor" trapezului dat de funcția de densitate.
- Functia de repartiție este crescătoare de la $-\infty$ la $+\infty$.
- Functia de repartiție este 0 pentru x < *a*, 1 pentru x >= *d*. Pentru *b* <= x < *c*, funcția de repartiție este liniară, în vreme ce pentru *a* <= x < *b* sau *c* <= x < *d* este cuadratică (x fiind un element din intervalul [*a, d*]).



- Atunci când *a* = *b* si *c* = *d*, funcția de repartiție și funcția de densitate de probabilitate specifice acestei repartiții vor deveni funcția de repartiție, respectiv funcția de densitate de probabilitate a repartiției uniforme.



- Atunci cand *b* = *c*, funcția de repartiție și funcția de densitate de probabilitate specifice acestei repartiții vor deveni funcția de repartiție, respectiv funcția de densitate de probabilitate a repartiției triunghulare.



Repartițiile trapezoidale sunt potrivite pentru modelarea duratei și formei unui fenomen care poate fi reprezentat prin trei etape. Prima etapă poate fi privită ca o etapă de creștere, a doua etapă corespunde uneia ce prezintă o stabilitate relativă, iar a treia ilustrează un regres.

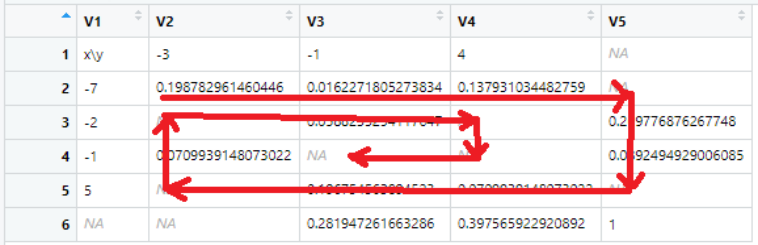
Un domeniu în care această repartiție este utilizată este cel al fizicii aplicate - mai exact, industria nucleară. Aici, repartiția trapezoidală impreună cu cea uniform sunt folosite ca modele de observare a calculelor aproximative a reducerii reactivității configurațiilor. Aceste repartiții sunt importante pentru a analiza în siguranță nivelul apei din reactoare. S-a arătat că distribuția uniformă este utilizată doar în cazul arderilor de intensitate mică, nu și atunci când acestea cresc. Astfel, repartițiile trapezoidale au fost recunoscute ca un pilon în industria nucleară datorită faptului ca au putut modela acest fenomen prezentat mai devreme care se desfășoară în trei etape.

BONUS

a) Pentru funcția *frepcomgen* am ales o strategie de generare a matricii ce folosește o parcurgere in spirală:

- Inițial generăm o matrice A care are ca elemente doar valori de 1 si -1. Valorile de 1 semnifica elementele pe care le cunoaștem, iar cele cu -1 reprezintă elementele care trebuiesc completate. Această metodă ne asigură faptul ca matricea poate fi completată, deoarece pentru a o rezolva este necesar doar să completăm elementele necunoscute in aceeași ordine în care le-am generat.
- Parcurgerea acestei matrice A pornește de la poziția (2,2) din matrice si parcurge in sensul acelor de ceasornic. Atunci când depașim limita matricei sau întalnim un element deja generat, schimbăm direcția si atribuim valoarea -1 elementului curent
- Generăm o alta matrice B care va conține în întregime(cu excepția primei linii și coloane) valori aleatoare cuprinse între 0 si 100.
- Calculăm suma tuturor valorilor lui B și ne generăm o noua matrice C care va fi matricea noastra finală. Valorile lui C vor fi valorile lui B împarțite la suma lor. Elemente care sunt -1 în matricea A vor fi cele necunoscute din matricea C.

	V1	V2	V3	V4	V5
1	x\y	-3	-1	4	NA
2	-7	0.198782961460446	0.0162271805273834	0.137931034482759	NA
3	-2	NA	0.0588235294117647	NA	0.269776876267748
4	-1	0.0709939148073022	NA	NA	0.0892494929006085
5	5	NA	0.196754563894523	0.0709939148073022	NA
6	NA	NA	0.281947261663286	0.397565922920892	1



b) Pentru funcția *fcomprepcom*:

- Apelez o structura repetitiva care are ca și condiție de terminare evenimentul în care nu mai există niciun element necompletat.
- Caut o linie sau o coloană care are toate elementele completate, mai puțin unul, pe care il completez in funcție de celelalte.

	V1	V2	V3	V4	V5
1	x\y	-3	-1	4	NA
2	-7	0.198782961460446	0.0162271805273834	0.137931034482759	0.352941176470588
3	-2	0.030425963488843	0.0588235294117647	0.180527383367139	0.269776876267748
4	-1	0.0709939148073022	0.0101419878296149	0.00811359026369141	0.0892494929006085
5	5	0.0202839756592308	0.196754563894523	0.0709939148073022	0.288032454361056
6	NA	0.320486815415822	0.281947261663286	0.397565922920892	1

c)

1) Aflu covarianța cu ajutorul formulei: Cov(X,Y)=E(XY)-E(X)E(Y). Determin X,Y și XY și calculez pentru fiecare media.

2) s1 = calculez suma tuturor elementelor din repartiția comună care respectă 0<X<3 și Y>2

s2 = calculez suma tuturor elementelor din repartiția comună care respectă Y>2 returnez s1/s2.

3) returnez suma tuturor elementelor care respectă X>6 și Y<7 .

d)

1) Pentru a verifica dacă repartițiile sunt independente, verific pentru fiecare element din repartitia comună daca acesta este egal cu produsul probabilităților corespunzatoare repartițiilor X și Y.

2) Pentru a verifica daca repartitiile sunt necorelate, calculez covarianța acestora cu ajutorul funcției de la punctul c) 1), iar dacă aceasta returnează 0 atunci repartițiile sunt necorelate.