Probabilități și statistică - Proiect final

Cîrstea Ionela-Mădălina, Cîrstea Natașa-Alexandra, Puiu Ana Maria

Februarie, 2020

1 Descrierea setului de date

Setul de date Trees pune la dispozitie informații despre diametrul, înălțimea și volumul de cherestea obținute în urma observațiilor efectuate pe un număr de 31 de cireși negri, etichetate Girth, Height și Volume. Diametrul a fost măsurat la 4 - 6 ft deasupra solului.

[1] Girth - diametrul arborelui (mai degrabă decât circumferința) măsurat în inch
[2] Height - înălțimea măsurată în ft
[3] Volume - volumul de cherestea măsurat în ft^3

2 Cerințe

1. Folosind setul de date X efectuaţi operaţii de statistică descriptivă pentru variabilele din acest set de date(medie, varianţa, quartile, boxplot, interpretări).

2. Folosind setul de date X construiţi două modele de regresie(o regresie simplă şi una multiplă) alegând după cum consideraţi potrivite variabila răspuns şi respectiv variabilele predictor. Adaugati la setul de date inițial una sau mai multe variabile pe care sa le consideraţi potrivite a fi incluse in cel puţin un model de regresie. Generaţi datele aferente variabilei nou adăugate conform unei repartiţii potrivite(folosiţi funcţiile din R care încep cu r: ex. pentru repartiţia normală rnorm). Justificaţi alegerile făcute şi interpretaţi rezultatele obținute În urma evaluării celor două modele de regresie. Care din cele două modele construite consideraţi că este mai potrivit pentru setul vostru de date? Daţi cel puţin două argumente pentru alegerea făcută.

3. Alegeţi o repartiţie diferită de cele studiate la laboratorul sau la cursul de Probabilităţi şi Statistică şi construiţi în două reprezentări alăturate funcţia de masă/densitatea de probabilitate(după cum e o repartiţie a unei variabile aleatoare discretă sau continuă) şi respectiv funcţia de repartiţie.

Indicaţi proprietăţile pe care le identificaţi la cele două funcţii şi precizaţi la ce este folosită repartiţia respectivă în practică(adică ce fel de fenomene poate modela).

3 R packages

1. trees - vom folosi setul de date trees, pentru a construi două modele de regresie liniară cu unul sau mai mulți predictori
2. ggplot2 - vom folosi acest pachet pentru a construi ploturi ale modelelor noastre
3. GGally - acest pachet extinde funcționalitatea ggplot2. O vom folosi pentru a crea o matrice grafică, parte a vizualizării datelor explorate inițial
4. scatterplot3d - vom folosi acest pachet pentru vizualizarea modelelor de regresie liniară mai complexe cu mai mulți predictori

4 Statistică descriptivă

Statistica descriptivă este utilizată pentru a descrie caracteristicile unui set de date, care poate reprezenta o întreagă populatie sau un eșantion din aceasta. Statistica descriptivă se împarte în măsurători de tendință centrală (medie, mediană) și măsuratori a variabilității (deviație standard, minim/maxim, asimetrie).

Operații de statistică descriptivă

- **Media** este o tendință centrală a datelor, mai exact un număr în jurul căruia sunt răspândite acestea și care poate estima valoarea întregului set. În R, ea se calculează prin intermediul funcției mean() care primește drept parametru un set de date.
- **Mediana** este valoarea care împarte un set de date în două părți egale, numărul de termeni din stânga fiind egal cu cel din dreapta când datele sunt aranjate în ordine crescătoare sau descrescătoare. Funcția corespondentă în R este median().
- **Varianța** reprezintă pătratul deviației standard (deviația medie a valorilor dintr-un set de date față de medie). Ea poate fi calculată cu ajutorul funcției var().
- **Quartilele** sunt valori care împart un set de date în sferturi. Ele sunt mediile primei (prima quartilă) și ultimei jumătăți (a treia quartilă), respectiv media întregului set (a doua quartilă). Se regăsesc la 25%, 50% și 75%. Funcția folosită este quantile().

- **Boxplot-ul** prezintă rezumatul de cinci numere a unui set de date: minimul, prima quartilă, media, a treia quartilă și maximul. Prin urmare, el este o măsură a cât de bine distribuite sunt datele într-un set.

Interpretări ale valorilor obținute

- **Medie și mediană**
  - **Girth:** Datele sunt ușor înclinate spre dreapta (distribuție pozitiv înclinată), media ($\simeq$ 13.25) fiind mai mare decât mediana ($\simeq$ 12.9).
  - **Height:** Datele sunt relativ echilibrate (distribuție simetrică), media (= 76) fiind egală cu mediana (= 76).
  - **Volume:** Datele sunt ușor înclinate spre dreapta (distribuție pozitiv înclinată), media ($\simeq$ 30.17) fiind mai mare decât mediana ($\simeq$ 24.2).
- **Varianță**
  - Varianța mică sugerează faptul că valorile sunt apropiate de medie si unele față de celelalte, în această ipostază situându-se **Girth** ($\simeq$9.85) și **Height** ($\simeq$40.6).
  - Varianța mare indică opusul: valorile sunt depărtate de medie și între ele - **Volume** ($\simeq$270.2).
- **Quartile**
  - **Girth:** Valorile quartilelor (11.05, 12.9 și 15.25) indică faptul că dispersia este mai mare între valorile mai mari ale setului decât între cele mai mici și că distribuția este ușor înclinată pozitiv.
  - **Height:** Întrucât mediana este media primei și ultimei quartile (72, 76, 80), dispersia valorilor este uniformă, fapt susținut și de egalitatea dintre medie și mediană.
  - **Volume:** Valorile quartilelor fiind 19.4, 24.2 și 37.3, se deduce o dispersie mai mare între valorile mai mari ale setului decât între cele mai mici și că distribuția este înclinată pozitiv.
- **Boxplot**
  - **Girth:** Boxplot-ul indică faptul că printre valori nu există outliers și că dispersia este mai mare între valorile mai mari.
  - **Height:** Nu se identifică niciun outlier în boxplot, iar dispersia este relativ simetrică.
  - **Volume:** Boxplot-ul sugereaza faptul ca exista un singur outlier și că dispersia este mai mare între valorile mai mari.

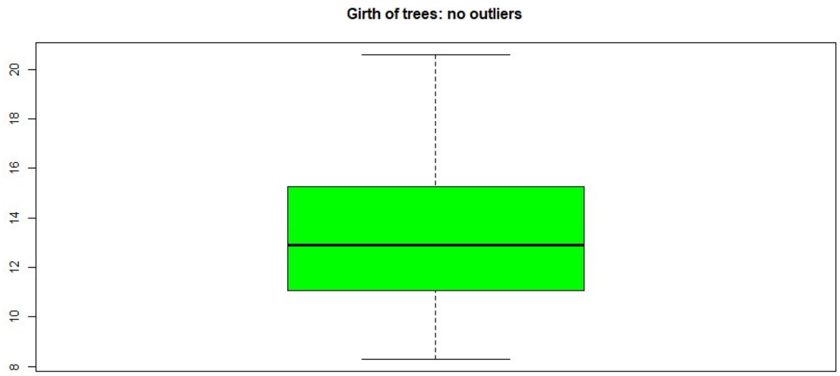

Figure 1: Boxplot - Girth

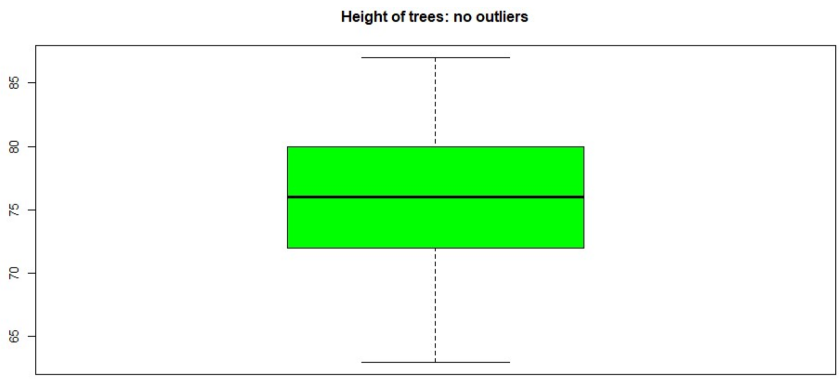

Figure 2: Boxplot - Height

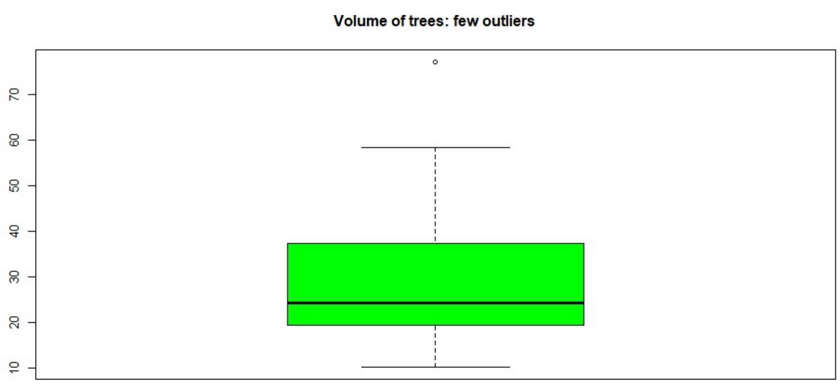

Figure 3: Boxplot - Volume

5 Regresie liniară simplă

Descrierea generală: Vrem să vedem dacă circumferința și înălțimea unui cireș negru influențează volumul de cherestea care poate fi obținut.

1. Pentru a verifica dacă setul de date este potrivit pentru a aplica regresia liniară simplă (cu alte cuvinte, pentru a decide dacă putem construi un model predictiv), calculăm corelația dintre circumferință și volum $\simeq$ 0,9671194 (circumferința va fi utilizată ca variabilă predictor și volumul ca variabilă răspuns), precum și corelația dintre înălțime și volum $\simeq$ 0,5982497. Se pare că există o relație mai puternică între circumferință și volum decât între volum și înălțime, deoarece coeficientul de corelație circumferință - volum este mai aproapiat de 1. Folosind funcția ggpairs() din pachetul GGally, creăm o matrice de tip plot pentru a vizualiza mai bine cum variabilele se interacționează între ele. De la analizarea outputului funcției ggpairs(), circumferința pare să fie cu siguranță legată de volum: coeficientul de corelație este aproapiat de 1, iar punctele par să aibă un model liniar.

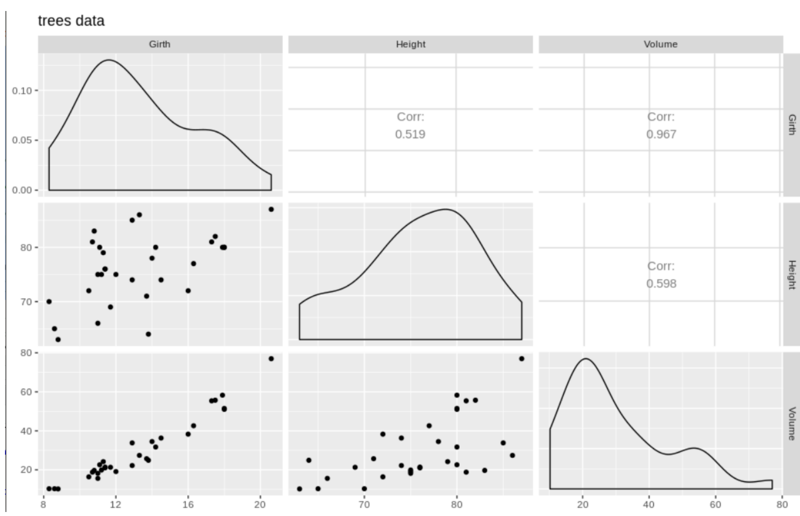

Figure 4: ggpairs() output

2. În continuare, vom împărți setul de date într-un eșantion 80:20 (training: test), vom folosi eșantionul 80% pentru a determina modelul liniar (training), iar eșantionul 20% pentru testare. Pe baza modelului construit vom prezice volumul, variabila dependentă, pe datele de testare. Analizând graficul pentru variabilele de circumferință și volum pe datele de antrenament, observăm o tendință ascendentă (cu cât este mai mare circumferința arborelui, cu atât volumul de cherestea crește).

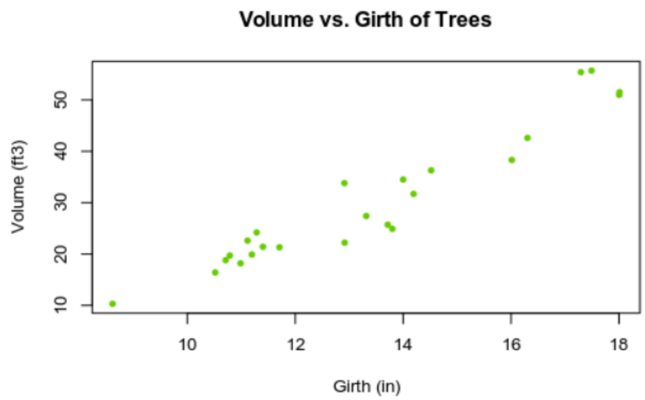

Figure 5: Volume vs. Girth

3. Apelând funcția lm() pe datele de training, construim un model liniar simplu în care volumul arborelui (variablia răspuns) depinde doar de circumferință (predictor, variabilă independentă). Funcția lm() determină o linie care să fie apropiată de toate cele 31 de observații. Mai precis, determinarea liniei se face astfel încât să se reducă suma diferenței pătratelor formate între puncte și linie; această metodă este cunoscută sub denumirea de "minimizarea celor mai mici pătrat".

```
Call:
lm(formula = Volume ~ Girth, data = trainingData)

Residuals:
   Min     1Q Median     3Q    Max 
-7.301 -1.545 -0.235  1.721  6.861 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -32.2183     3.9828  -8.089 4.91e-08 ***
Girth         4.6680     0.2925  15.960 1.40e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.759 on 22 degrees of freedom
Multiple R-squared:  0.9205,    Adjusted R-squared:  0.9169 
F-statistic: 254.7 on 1 and 22 DF,  p-value: 1.4e-13
```

Figure 6: Regresie liniară simplă - summary

4. Pentru a analiza acuratețea modelului creat, aplicăm funcția summary(). Oservăm astfel că pvalue este egal cu 1.4e-13, o valoare mai mică de 0.05 ce indică faptul că variabila predictor Girth este seminificativă pentru model. Mai mult, valoarea R^2 este o măsură a apropierii datelor de modelul de

regresie liniară. În cazul nostru, valoarea lui adjusted R^2 este 0.9169 (foarte apropiată de 1), ceea ce ne sugerează că am ales un model adecvat datelor. Din linia de regresie observăm o tendință ascendentă și deducem că un volum mai mare de cherestea este obținut de la un copac cu un diametru mai mare.

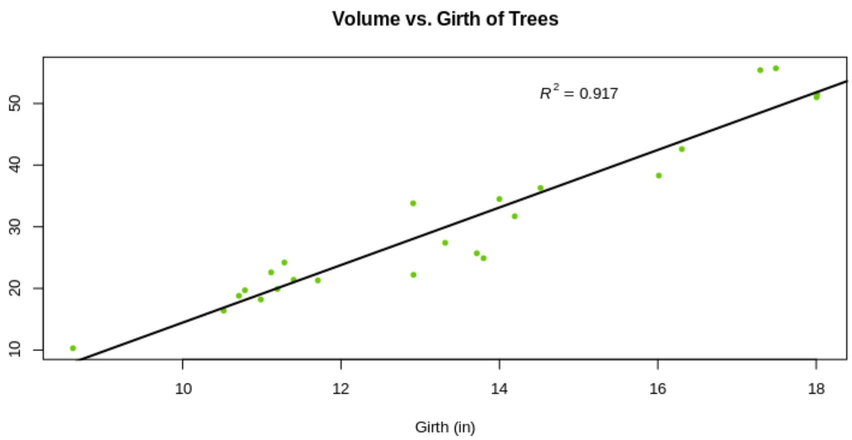

Figure 7: Regresie liniară simplă

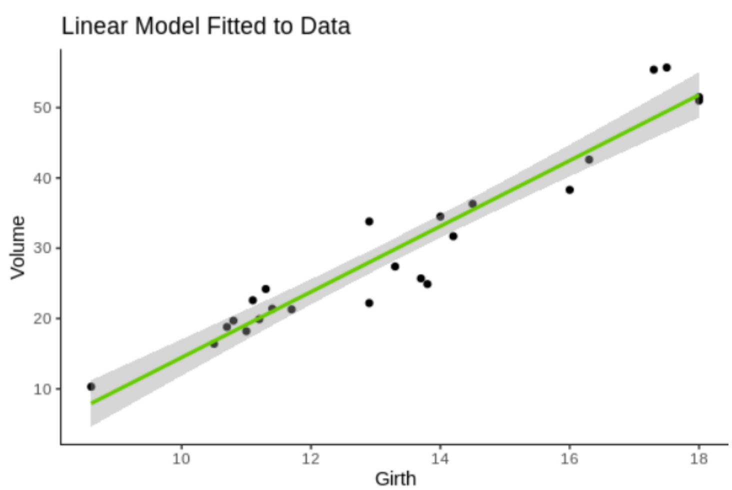

Figure 8: Intervalul de 95% încredere pentru linia de regresie

5. Umbra cenușie din jurul liniei de regresie reprezintă intervalul de 95% încredere; probabilitatea ca adevăratul model liniar pentru diametru - volum să se afle în această secțiune este deci foarte mare. Astfel, datele

noastre sunt suficient de puternice pentru crearea unui model capabil de a face predicții precise (pertinente).

```
> head(actuals_preds)
   actuals predicteds
1     10.3   6.526385
3     10.2   8.860404
7     15.6  19.130091
12    21.0  20.997306
15    19.1  23.798130
28    58.3  51.339562
> min_max_accuracy <- mean(apply(actuals_preds, 1, min) / apply(actuals_preds, 1, max))
> mape <- mean(abs((actuals_preds$predicteds - actuals_preds$actuals))/actuals_preds$actuals)
> min_max_accuracy
[1] 0.8330377
> mape
[1] 0.1798647
```

Figure 9: Predicții obținute pe setul de testare

6. Vom folosi funcția predict(), o funcție generică din R pentru a face predicții. predict() ia ca argument modelul nostru de regresie liniară și valorile setului de date de testare. Dorim să estimăm acuratețea cu care modelul construit face predicții pe baza măsurătorilor de precizie (min-max-accuracy $\simeq$ 0.8330377 - cu cât mai mare, cu atât mai bine) și rata de eroare (MAPE: media erorii procentuale absolute $\simeq$ 0.1798647 - cu cât mai mic, cu atât mai bine). Cum valoare ratei de eroare este foarte mică, iar valoarea min-max-accuracy este foarte apropiata de 1, cocluzionăm că acuratețea predicțiilor facute este considerabilă.

6 Regresie liniară multiplă

Descrierea generală: Ne întrebăm dacă putem îmbunătăți capacitatea de predicție a modelului nostru prin folosire tuturor informațiilor disponibile (circumferința și înălțimea) pentru a face predicții despre volumul arborelui.

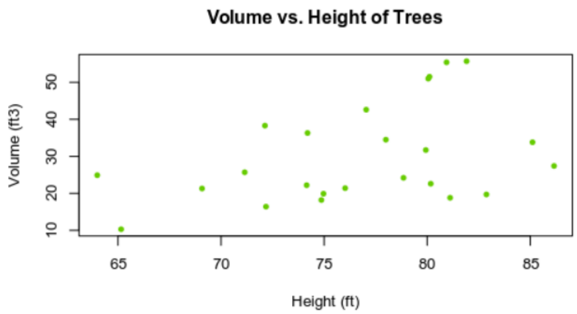

Figure 10: Volume vs. Height

1. După cum am arătat în modelul anterior, există o relație ascendentă între circumferință și volum. Prin același raționament, corelația dintre înălțime și volum nu poate fi ignorată, deoarece este suficient de mare pentru a fi luată în considerare ($\simeq$ 0.5982497). Interpretând plotul pentru înălțime - volum observăm o altă înclinație ascendentă, de această dată una mai mică decât cea dintre diametru - volum, însă și de această dată seminificativă. Astfel, bănuim că, în mare parte, pe măsură ce înălțimea arborelui crește, se mărește și volumul. Prin urmare, o soluție mai bună pentru problemă este construirea unui model liniar care să includă mai multe variabile predictori. Putem face acest lucru adăugând un coeficient de pantă pentru fiecare variabilă independentă suplimentară de interes pentru modelul nostru: circumferința și înălțimea.

$$volume = \beta_0 + \beta_1 \star (girth) + \beta_2 \star (height)$$

unde:

$$\beta_0, \beta_1, \beta_2$$

reprezintă: interceptul (= -56.9240), coeficientul de pantă pentru Girth (= 4.4520) și coeficientul de pantă pentru Height (= 0.3601)

2. Această pantă ne spune cât de mult se va schimba volumul dacă circumferința crește cu un inch sau înălțimea crește cu un ft.

3. Analizând ieșirea funcției summary() pentru modelul nou creat, putem vedea că atât circumferința, cât și înălțimea sunt semnificativ legate de volum și că modelul se potrivește bine datelor noastre. Valoarea obținută pentru adjusted R^2 a crescut față de cea obținută pentru modelul anterior (= 0.9386). De asemenea, valoarea lui pvalue este mai mică decât nivelul de semnificație statistică predeterminat (= 7.28e-14), așa că știm că avem un model statistic semnificativ. Deoarece avem două variabile predictor în acest model, avem nevoie de o a treia dimensiune pentru a vizualiza modelul. Putem crea un grafic de împrăștiere 3d folosind pachetul scatterplot3d. În primul rând, realizăm o grilă de valori pentru variabilele noastre predictor (în limita datelor noastre; ținand cont de faptul că dimensiunea setului de date este extrem de mică, trebuie să considerăm și posibilitatea ca modelul nostru să facă overfitting, adică să producă o analiză care să corespundă prea îndeaproape sau exact unui anumit set de date și, prin urmare, să nu poată fi adaptat pentru date suplimentare, eșuând astfel în a face predicții în mod fiabil). Funcția expand.grid() creează un frame de date ce conține toate combinațiile de variabile posibile.

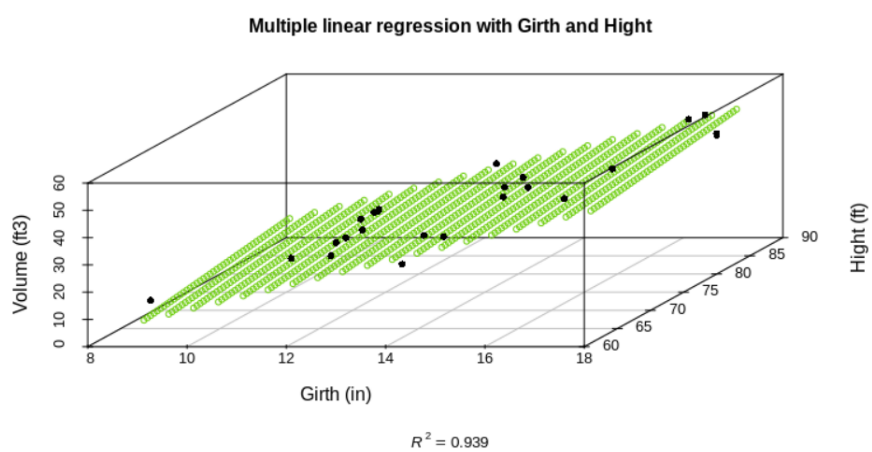

Figure 11: Regresie liniară multiplă cu Girth, Height și Volume

7 Regresie liniară multiplă cu interacțiuni

1. Deși am făcut îmbunătățiri, modelul pe care tocmai l-am construit încă nu reflectă cu exactitate realitatea, căci presupune că efectul circumferinței arborelui asupra volumului este independent de efectul înălțimii arborelui asupra volumului. Această ipoteză nu este validă, întrucât bănuim că înălțimea și circumferința arborilor sunt, de asemenea, relaționate.

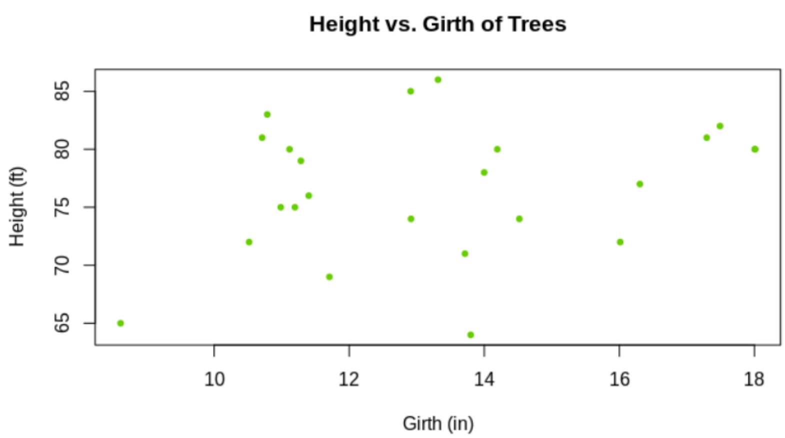

Figure 12: Girth vs. Height

2. Pe măsură ce calculăm corelația ($\simeq 0.5192801$), devine din ce în ce mai clar că există o legătură între circumferință și înălțime. Altfel spus, panta pentru circumferință ar trebui să crească pe măsură ce panta pentru înălțime crește. Pentru a ține cont de această neindependență a variabilelor predictoare din modelul nostru, putem specifica un termen de interacțiune, care este calculat ca produs al variabilelor predictor.

$$volume = \beta_0 + \beta_1 \star (girth) + \beta_2 \star (height) + \beta_3 \star (girth \star height)$$

3. După cum bănuiam, interacțiunea dintre circumferință și înălțime este semnificativă, ceea ce sugerează că decizia de a include termenul de interacțiune în modelul pe care îl utilizăm pentru a prezice volumul arborelui a fost corectă. Această concluzie este, de asemenea, susținută de valoarea lui adjusted R^2 mai apropiată de 1 și mai mare decât valoarea lui adjusted R^2 obținută în modelul anterior (= 0.9558), valoarea pvalue (= 2.564e-14) mai mică decât cea anterioară, valoarea min-max-accuracy (= 0,908044) mai mare decât cea menționată în modelul anterior și scăderea ratei de eroare (mape = 0,1002732); toate acestea sugerează că modelul curent (regresia liniară multiplă cu interacțiuni) este cel mai potrivit pentru setul de date ales dintre toate cele considerate.

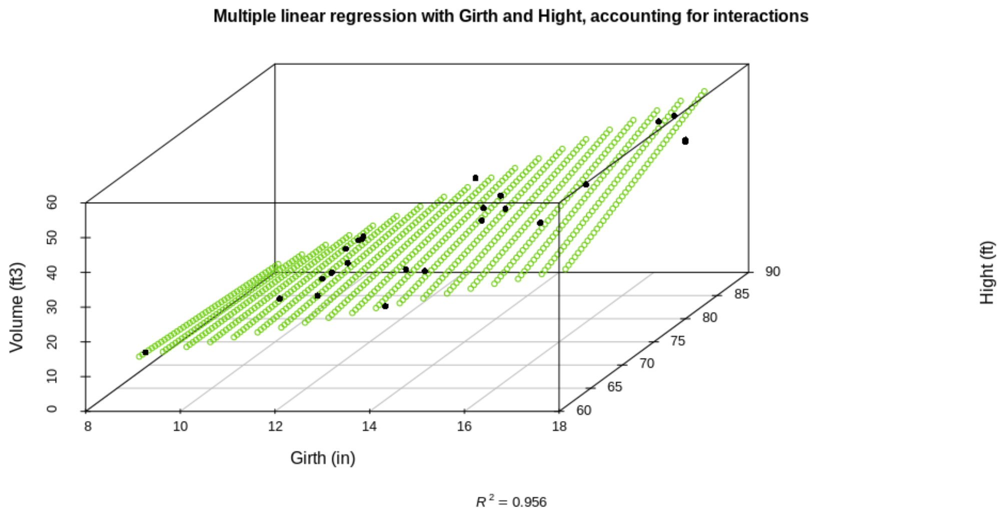

Figure 13: Regresie liniară multiplă cu interacțiuni

8 Adăugare de noi date în data frame

(a) Ne întrebăm dacă putem îmbunătăți capcacitatea predictivă a modelului prin adăugare unei noi date observațiilor deja existente în setul de date. Adăugăm o nouă coloană în data frame, coloană ce va conține procentul de lemn din copac ce nu poate fi folosit pentru obținerea de cherestea. Altfel spus, procentul ce reprezintă defectele apărute în mod natural pe suprafața unui cireș negru.

(b) Apariția defectelor este favorizată de numeroși factori: tipul și caracteristicile speciei, condițiile climatice, poziția geografică, acțiunea distructivă a florei, a faunei sau a activității umane și de multe alte procese. Prin urmare, totul se reduce la interacțiuni ale mai multor procese ascunse, la scară mică. Având în vedere că apariția defectelor rezultă din însumarea mai multor procese la scară mică, intuim ca distribuția acestora se apropie de cea a unei repartiții normale (specifică tiparelor naturii). Fluctuațiile individuale, la scară mică, provocate de fiecare proces contribuitor urmează rar curba Gaussiană. Dar, prin agregarea mai multor fluctuații parțial necorelate, fiecare la scară mică în raport cu agregatul, suma fluctuațiilor se netezește în curba Gaussiană. De asemenea, prin combinarea anumitor defecte, probabititatea reducerii volumului poate crește semnificativ. Prin urmare, variabila aleasă poate influența în mod semnificativ volumul de cherestea.

(c) Cu ajutorul repartiției normale (rnorm), generăm o serie de 31 de valori reprezentând defectele copacilor în procente, așa cum presupunem că apar și în natură. Cum datele pe care trebuie să le introducem în data frame nu pot fi doar niște valori random (pentru că verosimilitatea investigațiilor ar avea de suferit, întrucât nu am ține cont de interdependențele din natură), trebuie să stabilim o legătură între defecte și cel puțin o altă caracteristică a observațiilor din data frame (Height, Girth, Volume). Intuim că, pe măsură ce diametrul unui arbore crește, crește și timpul expunerii la factorii de mediu ce ar putea conduce la apariția defectelor. Prin urmare, alegem să corelăm procentajul de defecte cu diametrele deja cunoscute (stabilim valorile sunt corelate într-o proporție de 30%) folosind o funcție auxiliară[1].

(d) Construim un nou model de regresie liniară multiplă cu interacțiuni pentru a prezice volumul (variabila răspuns) folosindu-ne de diametru și procentajul de defecte (variabile predictor). Procedeul de construcție este asemănător cazului anterior.

(e) Ultimul model liniar are valoarea lui adjusted R^2 mai mică decât cea obținută în regresia multiplă girth - height cu interacțiuni, ceea ce este de așteptat având în vedere că am ales corelația dintre defects - girth (= 0.3) mai mică decât corelația existentă între height și girth (= 0.519).

(f) Conculzii: Cel mai bun model predictiv pentru setul nostru de observații, dintre toate cele construite mai sus, rămâne volume-girth-height, întrucât corelația height-volume ($\simeq 0.5982497$) este cu mult mai mare decât corelația defects-volume ($\simeq 0.1399002$), ținâd cont de faptul că am pornit în construirea modelului alegând să corelăm defectele și înălțimea în proporție de 30%. Acest rezultat verifică observațiile noastre intuitive: înălțimea influențează volumul într-o proporție mai mare decât defectele.

[1] https://stats.stackexchange.com/questions/15011/generate-a-random-variable-with-a-defined-correlation-to-an-e

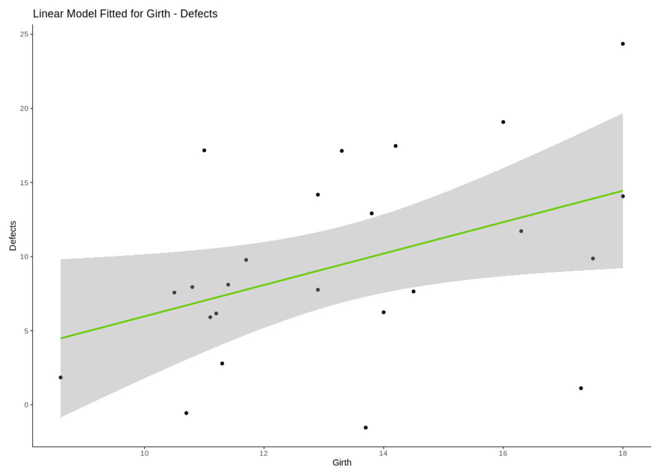

Figure 14: Date (Girth-Defects) corelate în proporție de 30%

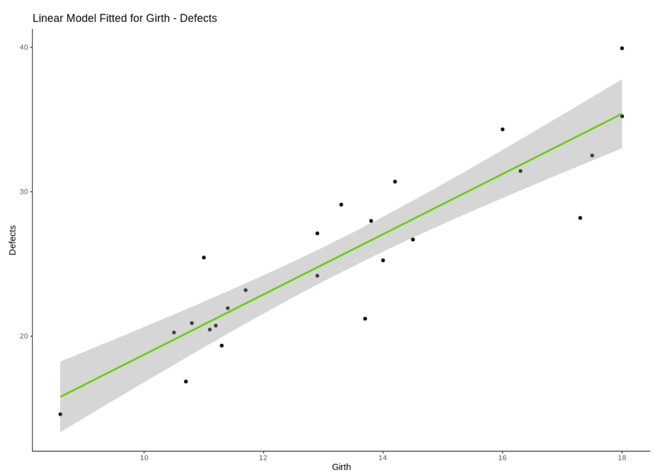

Figure 15: Date (Girth-Defects) corelate în proporție de 90%

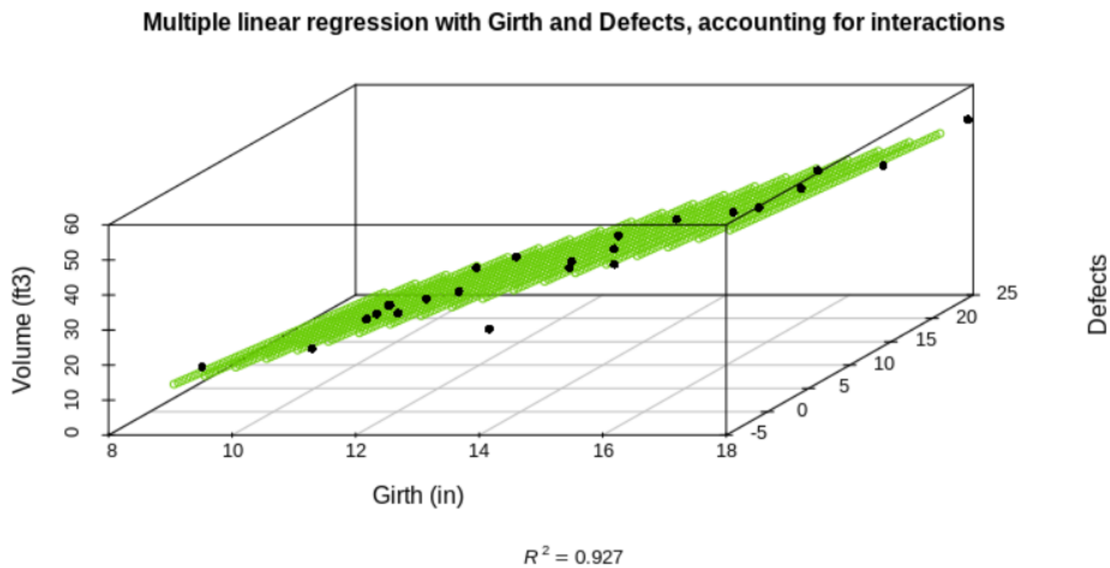

Figure 16: Regresie liniară multiplă cu interacțiuni

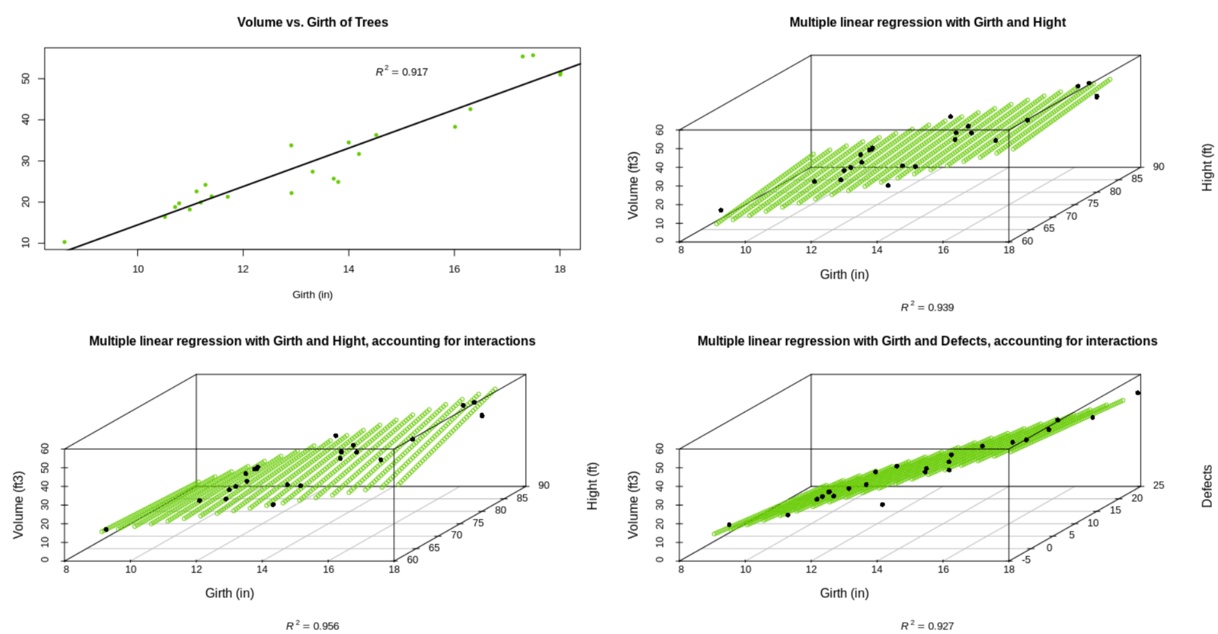

Figure 17: Concluzii: sumar regresii liniare

9 Repartiția Laplace

(a) În teoria probabilității și statisticilor, distribuția Laplace este o distribuție continuă a probabilităților numită după Pierre-Simon Laplace. Este, de asemenea, uneori numită distribuție exponențială dublă, deoarece poate fi gândită ca două distribuții exponențiale (cu un parametru suplimentar de locație) împărțite împreună înapoi.[2]

(b) **Definiții și proprietăți:** [2]

i. O variabilă aleatoare are o distribuție Laplace (μ,b) dacă funcția sa densitate de probabilitate este:

$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$
$$= \frac{1}{2b} \begin{cases} \exp\left(-\frac{\mu - x}{b}\right) & \text{if } x < \mu \\ \exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu \end{cases}$$

Funcția densității de probabilitate a distribuției Laplace este de asemenea o reminiscență a distribuției normale; cu toate acestea, în timp ce distribuția normală este exprimată în termenii diferenței pătrate față de media μ, densitatea Laplace este exprimată în termenii diferenței absolute față de medie. În consecință, distribuția Laplace are cozi mai grase decât distribuția normală.

ii. Distribuția Laplace este ușor de integrat (dacă se disting două cazuri simetrice) datorită utilizării funcției de valoare absolută. Funcția sa de repartiție este următoarea:

$$F(x) = \int_{-\infty}^{x} f(u)\, \mathrm{d}u = \begin{cases} \frac{1}{2} \exp\left(\frac{x-\mu}{b}\right) & \text{if } x < \mu \\ 1 - \frac{1}{2} \exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu \end{cases}$$
$$= \tfrac{1}{2} + \tfrac{1}{2} \operatorname{sgn}(x-\mu) \left(1 - \exp\left(-\frac{|x-\mu|}{b}\right)\right).$$

iii. Momente:

$$\mu_r' = \left(\frac{1}{2}\right) \sum_{k=0}^{r} \left[\frac{r!}{(r-k)!} b^k \mu^{(r-k)} \{1 + (-1)^k\}\right] = \frac{m^{n+1}}{2b} \left(e^{m/b} E_{-n}(m/b) - e^{-m/b} E_{-n}(-m/b)\right)$$

[2] https://en.wikipedia.org/wiki/Laplace_distribution

(c) **Observații:**

i. distribuția Laplace este o distribuție simetrică
ii. dispersia datelor în jurul mediei este mai mare decât aceea a unei distribuții normale
iii. o distribuție normală are cozi foarte subțiri, adică densitatea de probabilitate scade foarte repede pe măsură ce ne depărtăm de mijloc, ca exp(-x^2). Distribuția Laplace are cozi moderate [3], tinzând către zero ca exp(-$| x |$)[4]

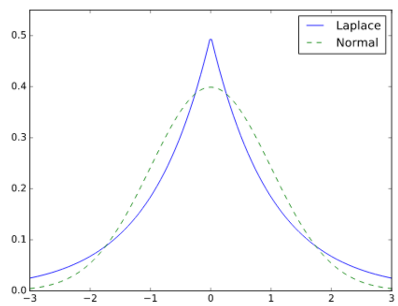

Figure 18: Comparație între repartiția Laplace și repartiția Normală

(d) **Aplicații:**

Distribuția Laplace este utilizată pentru modelarea procesării semnalului, a diverselor procese biologice, a finanțelor și a economiei. Exemple de evenimente care pot fi modelate de distribuția Laplace includ:

i. adăugarea zgomotului provenit dintr-o distribuție laplaceană, cu parametrul de scalare adecvat sensibilității unei funcții, la ieșirea unei interogări a bazelor de date statistice este cel mai comun mijloc de a oferi confidențialitate diferențială în bazele de date statistice
ii. în analiza regresiei, estimarea celor mai puține abateri absolute apare ca estimare a probabilității maxime dacă erorile au o distribuție Laplace
iii. în hidrologie, distribuția Laplace se aplică la evenimente extreme, precum precipitații maxime anuale de o zi și descărcări ale râurilor [5]

[3] Distribuția normală este exemplul canonic al unei distribuții cu coada subțire, în timp ce cozile exponențiale sunt în mod convențional limita dintre gros și subțire. Coada groasă" și coada subțire" reprezintă, de obicei, o coadă mai groasă decât exponențiala și, respectiv, o coadă mai subțire decât exponențiale.

[4] https://www.johndcook.com/blog/2019/02/05/normal-approximation-to-Laplace-distribution

[5] https://en.wikipedia.org/wiki/Laplace_distribution

iv. riscul de credit și opțiuni exotice în inginerie financiară

v. creanțe de asigurare

vi. modificări structurale în modelul regimului de comutare și filtrul Kalman[6]

(e) **Reprezentări vizuale:** ale funcției densitate de probabilitate și funcției de repartiție, folosind x = [-100, -99, ..., 99, 100] (vectorul răspunsurilor), μ = 10 (parametru locație) și b = 7 (diversitate , parametru de scară)

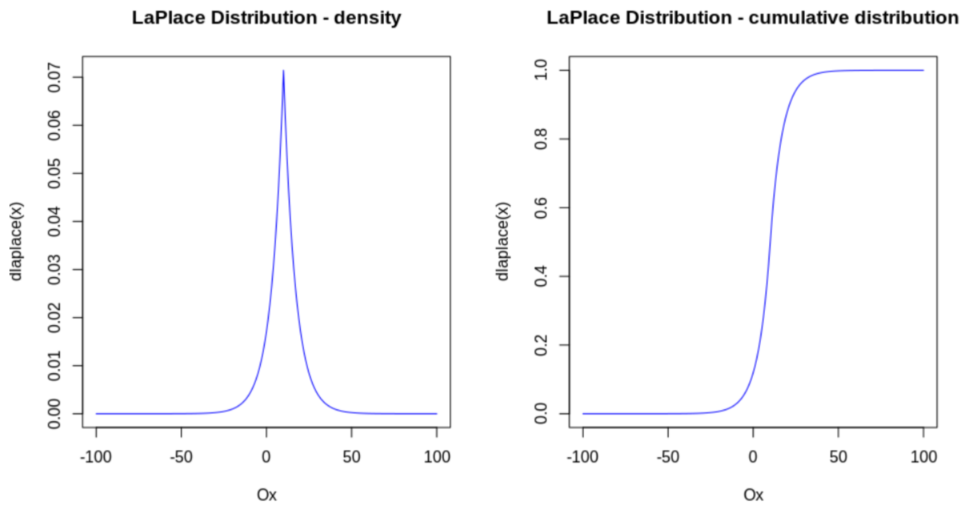

Figure 19: .
Reprezentări vizuale: funcția densitate de probabilitate, funcția de repartiție

(f) **Exemplu:**
Presupunând că reandamentul unui anumit stoc are o distribuție Laplace cu $\mu = 5$ și b = 2, calculati probabilitatea ca stocul să aibă un randament între 6 și 10.
Putem calcula acest lucru după cum urmează: [6].

$$P(6 \leq X \leq 10) = \sum_{x=6}^{10} \frac{1}{2 \times 2} \exp(-\frac{|x-5|}{2}) = 0.262223$$

[6] http://wiki.stat.ucla.edu/socr/index.php/AP_Statistics_Curriculum_2007_Laplace

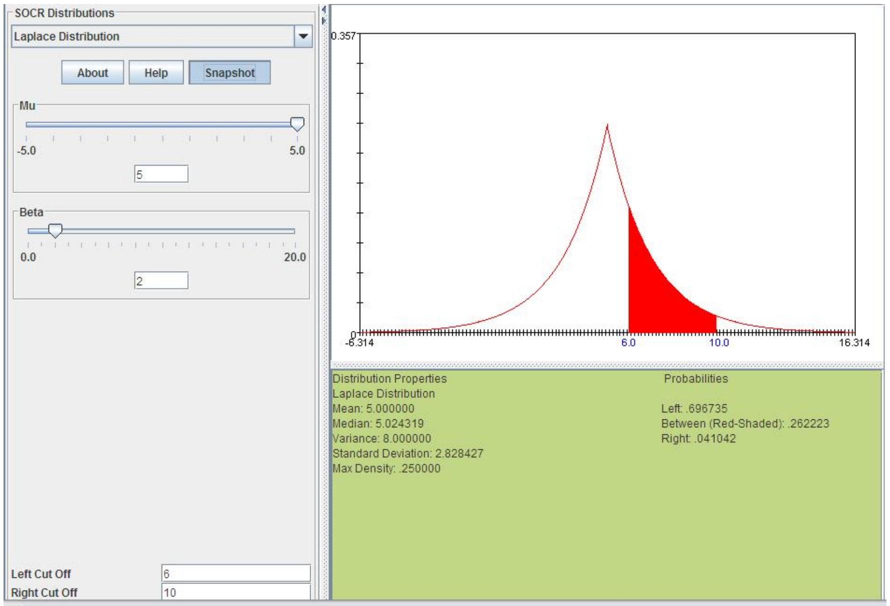

Figure 20: Repartiția Laplace

10 Referințe:

(a) http://r-statistics.co/Linear-Regression.html

(b) https://en.wikipedia.org/wiki/Laplace_distribution

(c) http://wiki.stat.ucla.edu/socr/index.php/AP_Statistics_Curriculum_2007_Laplace

(d) https://www.johndcook.com/blog/2019/02/05/normal-approximation-to-Laplace-distribu

(e) https://stats.stackexchange.com/questions/15011/generate-a-random-variable-with-a-

(f) http://rstudio-pubs-static.s3.amazonaws.com/138191_9169a18ae3d34e1492d1df67a810e5d html?fbclid=IwAR0I6erpHRt3YXESb8LvCExibCMDucibZa1iK-n0e3CjYb9lQCEUcuTyJUA

(g) https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/?fbclid=IwAR0ktY7Dcnu_8HLEildpw0U--9-oh1RweyqHedrd0QKP4XkSOQbCwYpA0Aw

(h) https://rpubs.com/Pun_/PredictiveModellingofVolumeofCheeryTrees?fbclid=IwAR1CWMFLwgWvafASA9S4KWcZSj1F_VI19E08t7NjSQ1XM2MYZs7j7w1B78I

(i) https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/treering.html?fbclid=IwAR1YiUoAzhIg8wcH0gRdLDI-ZcVMgGtRDaz54-R-Xbf6TRmY5Wlt54gCPcQ

(j) https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/?fbclid=IwAR0BSEgvcWC5yb8rVi0OK6LhpMIMBaWGlULIpm3KflfVgrF239KiIAxo9_w

(k) https://www.statmethods.net/stats/regression.html