1. Realizați două modele de regresie liniară simplă pentru a afla dacă variația numărului salariaților (Salariati) explică în mai mare măsură variația cifrei de afaceri (CA) sau variația profitului net (Profit_N).

- intai voi inspecta variabilele

Hide

```
summary(Bilant$Salariati)
```

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
  0.000    1.000    2.000    8.842    5.000 5172.000     4059
```

Hide

```
summary(Bilant$CA)
```

```
     Min.    1st Qu.     Median       Mean    3rd Qu.       Max.       NA's
  -920044      39200     158583    1910835     613840 1213005095       2385
```

Hide

```
summary(Bilant$Profit_N)
```

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
      0        0     4314   121468    50970 57041222        1
```

Observ ca variabilele contin valori lipsa si ar putea reprezenta o problema pentru analizele ulterioare daca nu sunt indepartate.Astfel, voi creea un obiect nou, “Bilant2” pe care il voi folosi in cele doua analize de regresie, intrucat nu va contine valorile lipsa.

Bilant2 <- Bilant[complete.cases(Bilant[,c(“Salariati”,“CA”, “Profit_N”)]),] –> desi i-am dat run si a mers, nu apare in nb ca celelalte coduri, i don’t know why :(

Hide

```
summary(Bilant2$Salariati)
```

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.000    1.000    2.000    9.255    5.000 5172.000
```

Hide

```
summary(Bilant2$CA)
```

```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max. 
   -920044      73434     231660    2329435     859843 1213005095 
```

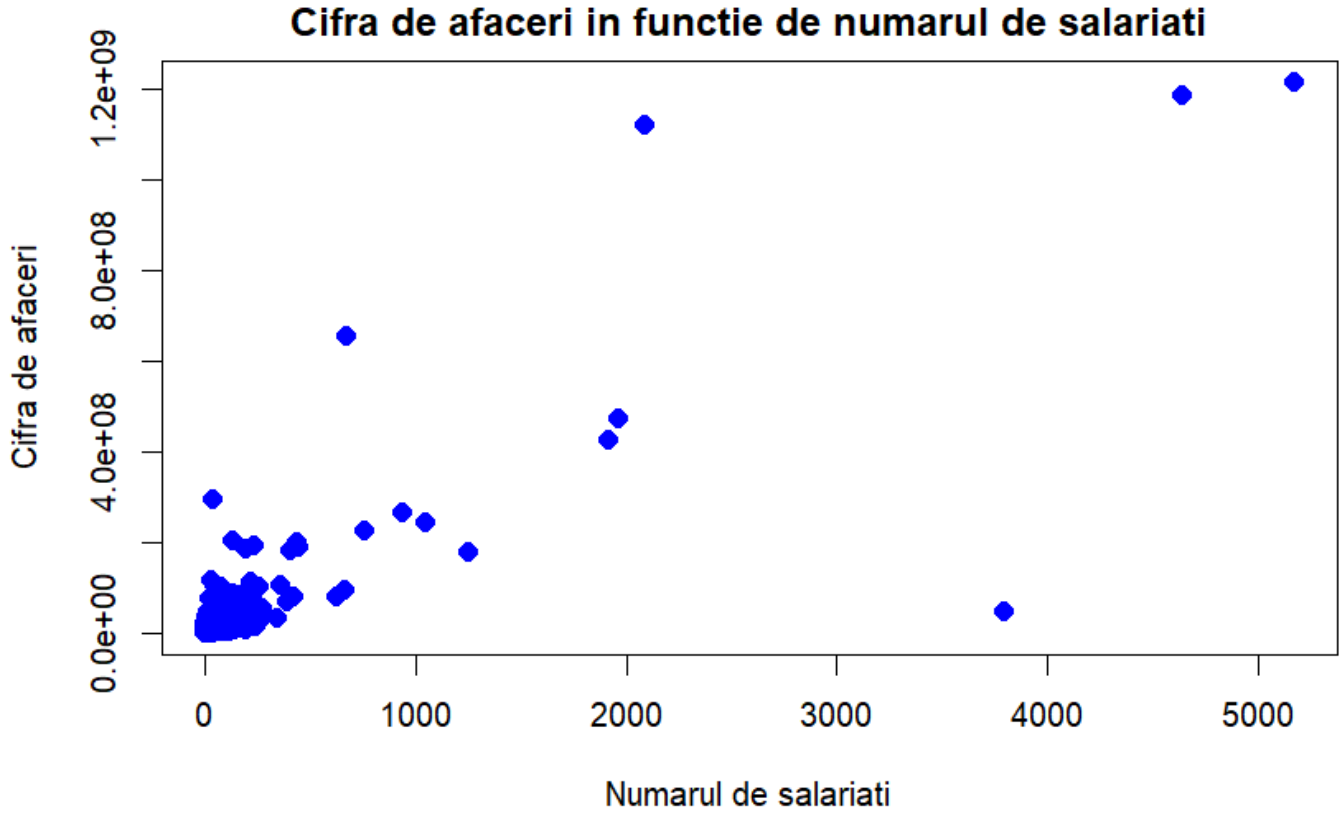
Hide

```
summary(Bilant2$Profit_N)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
       0        0    16574   172667    96563 57041222 
```

Prima analiza de regresie: Variatia numarului de salariati explica variatia cifrei de afaceri.

- voi reprezenta grafic relatia dintre cele doua variabile (unde numarul de salariati = variabila explicativa si cifra de afaceri = variabila dependenta) pentru a verifica daca exista corelatie



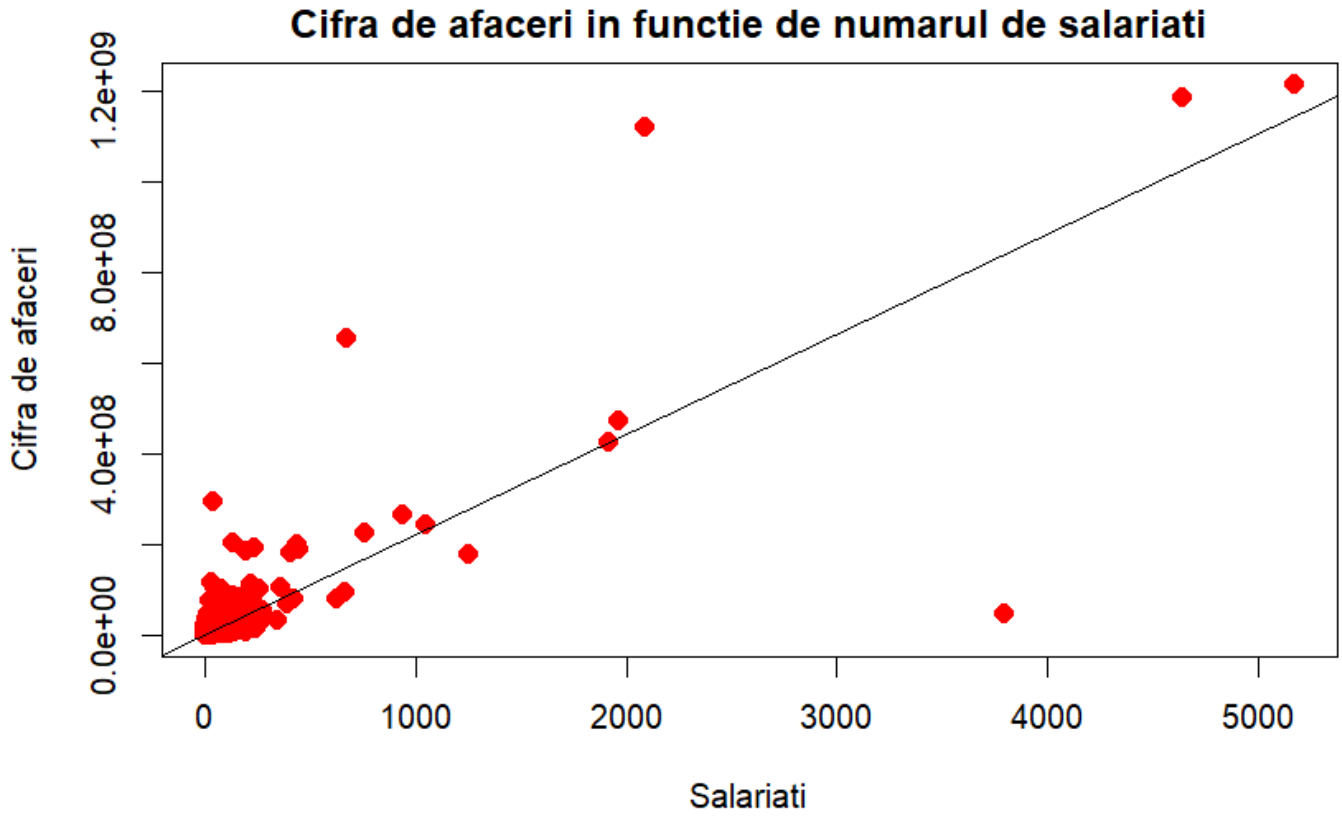
Conform graficului, intuiesc ca ar putea exista o corelatie pozitiva, desi se observa prezenta unor outliers si conglomerarea datelor in jurul unor valori foarte mici (s-ar putea sa existe probleme cu asumptiile de normalitate si de omogenitatea variantelor). Voi calcula coeficientul de corelatie Pearson.

Hide

```
cor(Bilant2$Salariati, Bilant2$CA, method="pearson")
```

```
[1] 0.8393502
```

Observ ca r=0.83, ceea ce inseamna ca exista o corelatie pozitiva foarte puternica (adica atunci cand numarul de salariati creste, creste si cifra de afaceri), lucru care imi permite sa continui analiza de regresie.



Observ ca punctele sunt destul de apropiate de linie (ceea ce ar duce la cresterea coeficientului de determinare si implicit a puterii explicative), insa aspectele mentionate anterior (outliers + conglomerarea datelor) sunt destul de ingrijoratoare cu privire la robustetea modelului.

Hide

```
arm::display(m)
```

```
lm(formula = CA ~ Salariati, data = Bilant2)
            coef.est  coef.se
(Intercept) 285448.89 142626.70
Salariati   220855.23   1492.28
---
n = 9189, k = 2
residual sd = 13607842.96, R-Squared = 0.70
```

Tabelul de mai sus ne ofera urmatoarele informatii: - Dreapta de regresie (interceptul) pleaca din punctul 285448.89, adica unei cifre de afaceri egala cu 0 i-ar corespunde aproximativ 285449 salariati. Observ, de asemenea, ca eroarea interceptului este extrem de mare, lucru care ar putea aduce scepticism in ceea ce priveste puterea explicativa a modelului. - Cand se angajeaza un salariat in plus in intreprindere, cifra de afaceri creste cu aproximativ 220855. - Valoarea coeficientul de determinare (R-Squared) imi spune ca modelul meu explica in proportie de 70% variatia cifrei de afaceri in functie de variatia numarului de salariati.

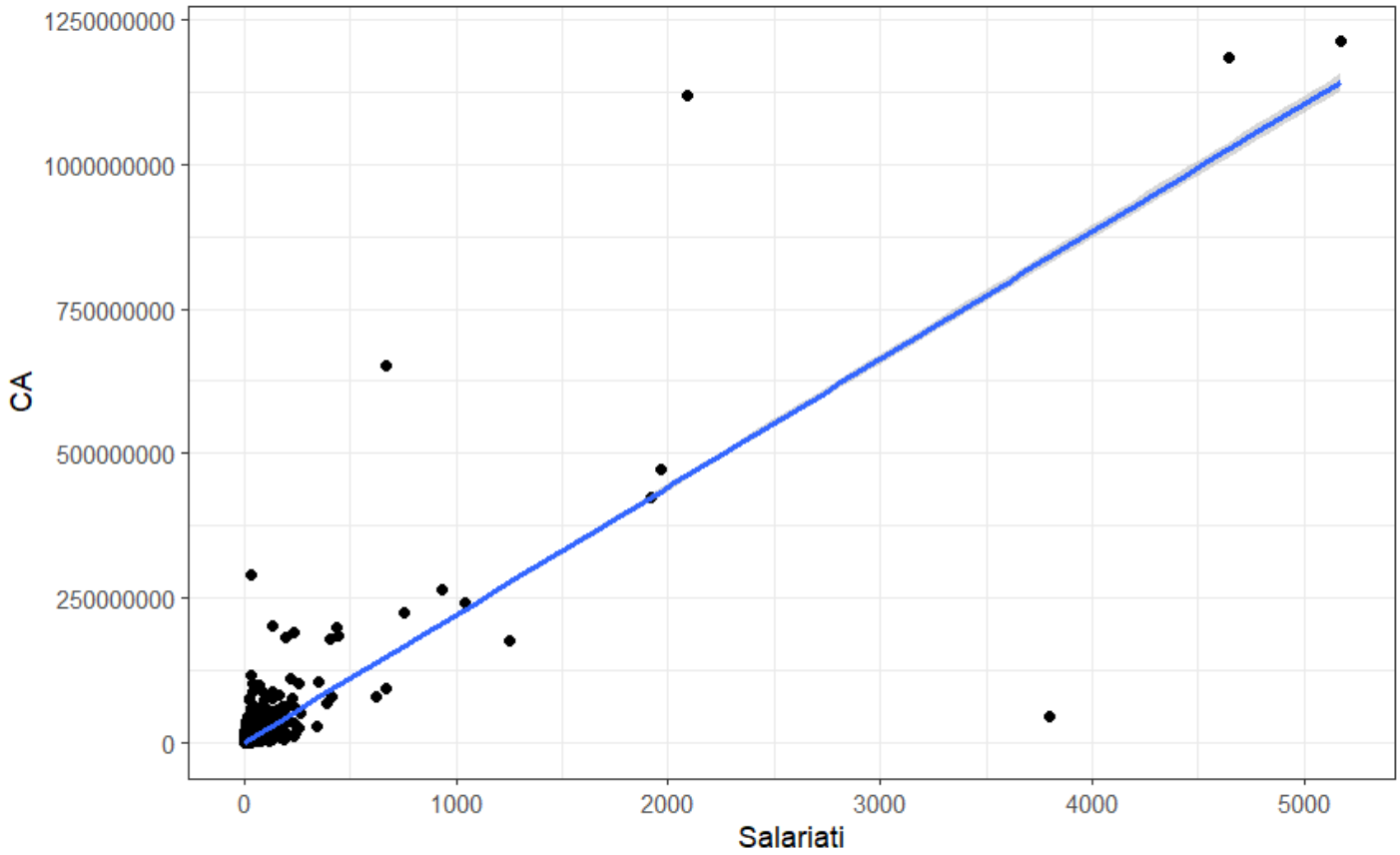
Hide

```
options(scipen=999)
summary(lm.beta::lm.beta(m))
```

```
Call:
lm(formula = CA ~ Salariati, data = Bilant2)

Residuals:
       Min         1Q     Median         3Q        Max 
-793626423    -715148    -460762    -241294  655187423 

Coefficients:
               Estimate Standardized  Std. Error t value            Pr(>|t|)    
(Intercept) 285448.8879           NA 142626.7020   2.001              0.0454 *  
Salariati   220855.2320       0.8394   1492.2784 147.999 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13610000 on 9187 degrees of freedom
Multiple R-squared:  0.7045,    Adjusted R-squared:  0.7045 
F-statistic: 2.19e+04 on 1 and 9187 DF,  p-value: < 0.00000000000000022
```

Conform testului T, variabila Salariati este semnificativa atunci cand vrem sa oferim o explicatie in ceea ce priveste variatia Cifrei de Afaceri. Practic, avem 0 dovezi in favoarea ipotezei nule intrucat p<0.01. De asemenea, testul F compara ceea ce modelul reuseste sa explice cu ceea ce nu reuseste sa explice si, in acest caz, este semnificativ (p<0.01), ceea ce inseamna ca modelul de regresie explica suficient de bine variata cifrei de afaceri in functie de numarul de Salariati. De asemenea, coeficientul de regresie standardizat imi indica faptul ca o crestere cu o abatere standard a numarului de salariati este asociata cu o crestere medie a cifrei de afaceri cu 0.83 abateri standard.

In cele ce urmeaza, voi investiga, prin intermediul intervalelor de confidenta, puterea explicativa a modelului.



Observ ca majoritatea punctelor sunt in afara intervalului marcat cu gri ceea ce inseamna ca erorile sunt prea mari pentru aceste cazuri (asa cum am si observat in tabelul de coeficienti) si o parte importanta din variatia cifrei de afaceri nu este explicata de modelul de regresie.

Pentru a intelege mai bine ecuatia de regresie, voi analiza valorile observate, valorile estimate si reziduurile.

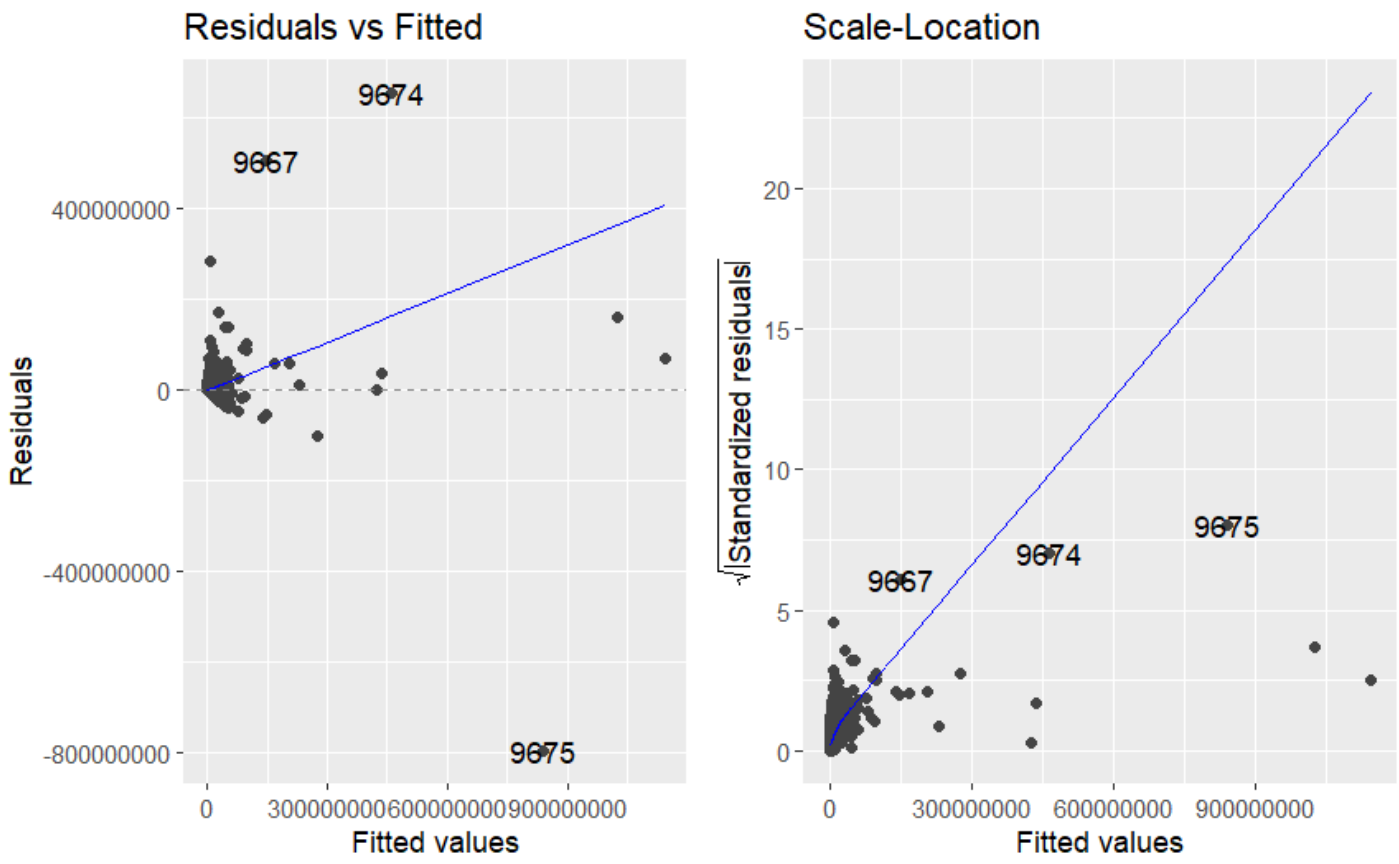
	Bilant2.Salariati <dbl>	**Bilant2.CA** <dbl>	**m.fitted.values** <dbl>	**m.residuals** <dbl>
1	0	19059501	285448.9	18774052.1
2	0	5023096	285448.9	4737647.1
3	0	4972332	285448.9	4686883.1
4	0	3065402	285448.9	2779953.1
5	0	1013003	285448.9	727554.1
6	0	806608	285448.9	521159.1

6 rows

La inspectarea tabelului, observ ca variabila “Bilant2.Salariati” contine foarte multe valori nule. Variabila “Bilant2.CA” reprezinta valorile observate in cifra de afaceri, iar m.fitted.values reprezinta valoarea prezisa prin intermediul ecuatiei de regresie. Erorile sunt extrem de mari, lucru care indica faptul ca modelul meu nu reuseste sa estimeze valori cat mai aproape de valorile observate. Cauza poate fi reprezentata de faptul ca exista si alte variabile necunoscute care pot explica mai bine variatia cifrei de afaceri.

- Testarea Asumptiilor pentru prima analiza de regresie

1. Egalitatea variantelor (Homoscedascitatea)



Observ ca asumptia este incalcata pentru ca: 1. avem outliers (9179, 9186, 9187) - problema care s-ar putea remedia prin simpla eliminare a acestora 2. ca sa existe homoscedasticitate, linia punctata ar trebui sa se suprapuna intr-o oarecare masura peste linia albastra, lucru care in graficul de mai sus nu se intampla. Observ o suprapunere in jurul valorii 0 (acolo unde exista acel conglomerat de date) si intuiesc ca ar putea fi vorba de o distributie asimetrica. 3. fluctuatiile diferentelor de varianta in functie de valorile prezise sunt destul de mari

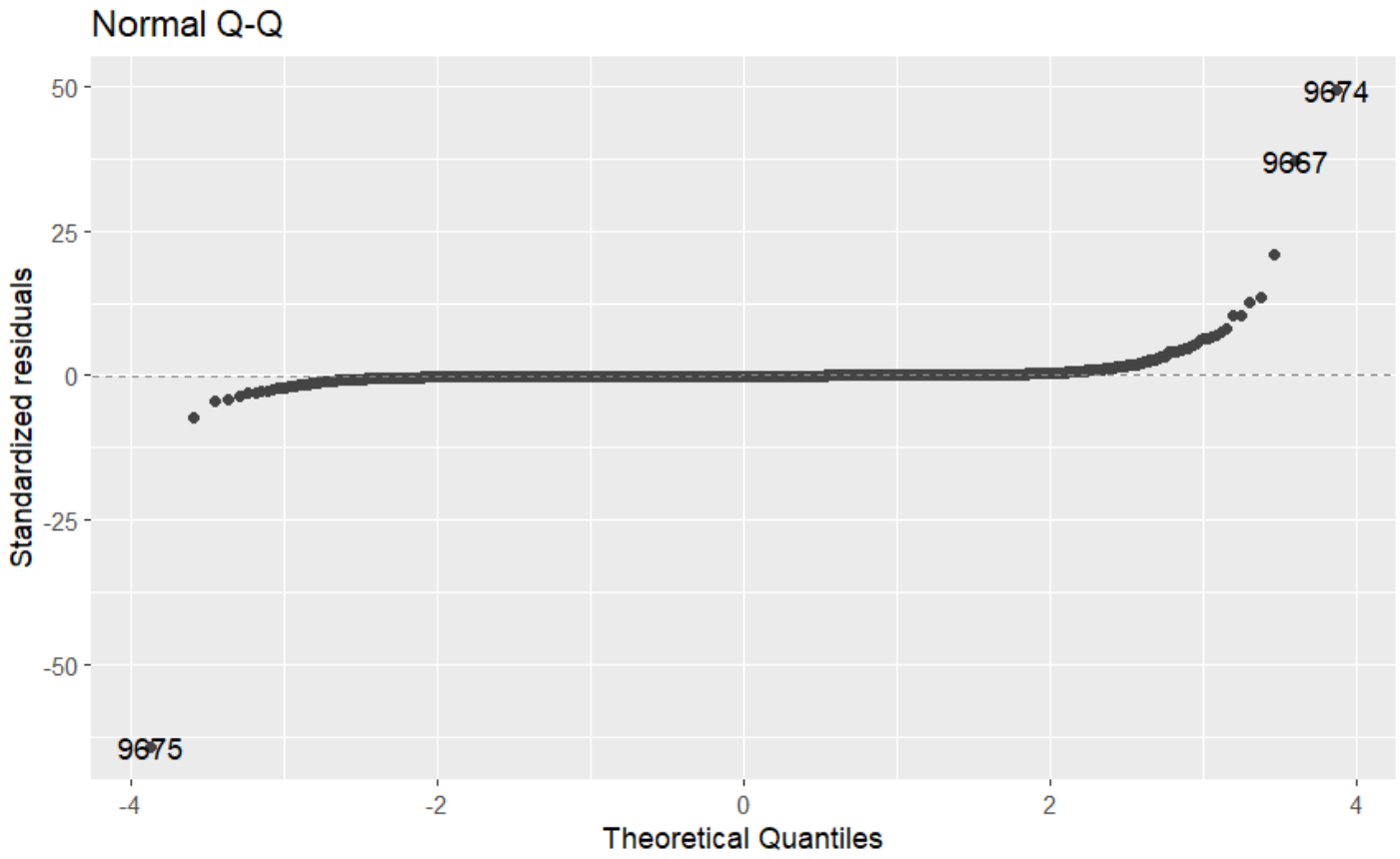
Hide

```
gqtest(m, data=Bilant2)
```

```
    Goldfeld-Quandt test

data:  m
GQ = 779.61, df1 = 4593, df2 = 4592, p-value < 0.00000000000000022
alternative hypothesis: variance increases from segment 1 to 2
```

Ipoteza nula este reprezentata de faptul ca exista homoscedasticitate. Avand in vedere faptul ca p<0.01, se respinge ipoteza nula si se confirma faptul ca asumptia omogenitatii variantelor este incalcata.

2. Normalitatea (distributia erorilor trebuie sa fie relativ normala)



Si aici se observa prezenta valorilor extreme, ceea ce face ca forma distributiei erorilor sa devieze de la normalitate. De asemenea, observ ca linia punctata nu se abate de la valoarea 0, ceea ce indica o distributie puternic asimetrica cu un conglomerat de date in jurul acestei valori (aspect observat si in reprezentarile vizuale anterioare)

Hide

```
shapiro.test(m$residuals[0:5000])
```

```
    Shapiro-Wilk normality test

data:  m$residuals[0:5000]
W = 0.32891, p-value < 0.00000000000000022
```

Deviatia de la normalitate este confirmata si statistic. Asadar, cele doua asumptii sunt incalcate. Acest lucru inseamna ca, desi am obtinut semnificativitate statistica in ceea ce priveste modelul (vezi R-squared, testul T si F), modelul meu nu este unul robust, adica variatia numarului de salariati nu reuseste sa explice variatia cifrei de afaceri pentru ca exista anormalitati (valori extreme, distributie asimetrica cu conglomerat de date in jurul valorii 0, diferente foarte mari intre valorile observate si cele estimate).

A doua analiza de regresie: Variatia numarului de salariati explica variatia profitului net.

Inspectez variabila dependenta.

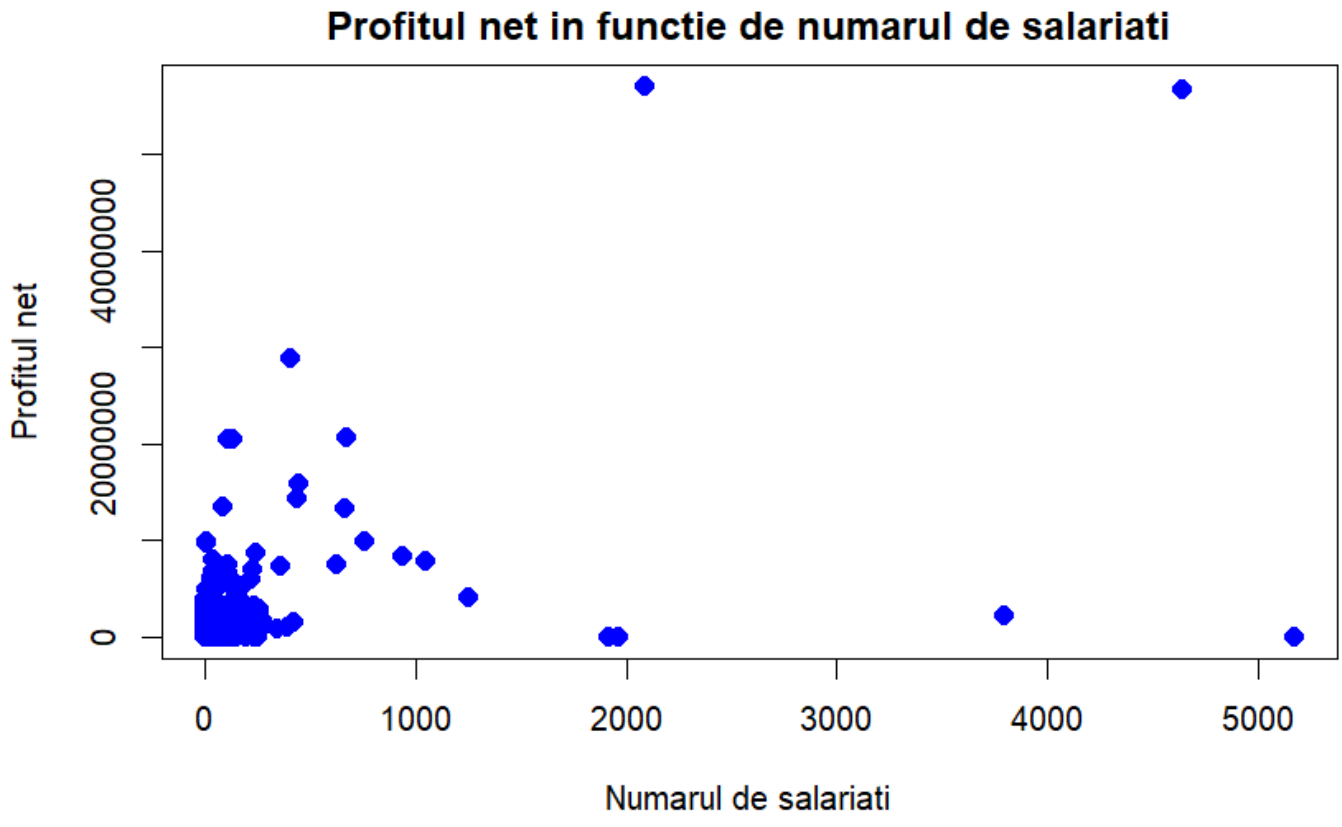
Hide

```
summary(Bilant2$Profit_N)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
       0        0    16574   172667    96563 57041222 
```

Observ si, in acest caz, o plaja larga de valori, avand in vedere min si max. De asemenea, faptul ca 25% dintre valori sunt 0 indica faptul ca distributia ar putea fi asimetrica.

Voi investiga relatia dintre variabila explicativa (numarul de salariati) si variabila dependenta (profitul net) pentru a vedea daca exista corelatie.



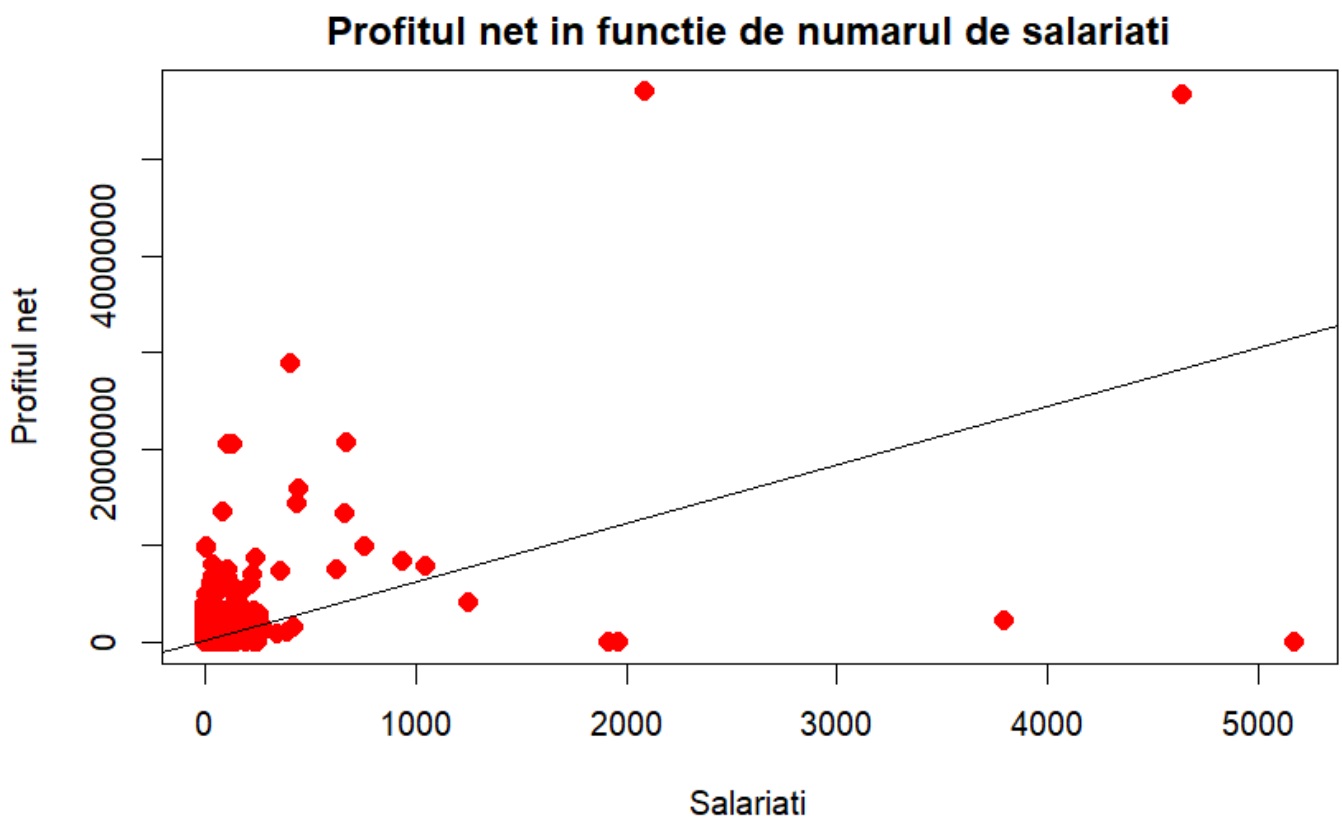
Graficul vizual indica, si de aceasta data, prezenta unei posibile corelatii pozitive, insa, la fel ca in cazul modelului trecut, se observa valori extreme si un conglomerat de date in jurul valorii 0.

Hide

```
cor(Bilant2$Salariati, Bilant2$Profit_N, method="pearson")
```

```
[1] 0.5125611
```

Coeficientul Person r=0,51, ceea ce indica o corelatie pozitiva moderata (adica si aici odata cu cresterea numarului de salariati, creste si profitul net). Prezenta corelatiei imi permite sa continui analiza de regresie.



Observ ca punctele sunt destul de apropiate de linie (ceea ce ar duce la cresterea coeficientului de determinare si implicit a puterii explicative), insa aspectele mentionate anterior (outliers + conglomerarea datelor) sunt destul de ingrijoratoare cu privire la robustetea modelului (erori foarte mari + incalcarea asumptiilor - se observa diferenta foarte mare intre variante)

Hide

```
arm::display(m1)
```

```
lm(formula = Profit_N ~ Salariati, data = Bilant2)
            coef.est  coef.se
(Intercept) 116442.96  10148.08
Salariati     6075.03    106.18
---
n = 9189, k = 2
residual sd = 968215.78, R-Squared = 0.26
```

Tabelul de mai sus ne ofera urmatoarele informatii: - Dreapta de regresie (interceptul) pleaca din punctul 116442.96, adica unui profit net egal cu 0 i-ar corespunde aproximativ 116443 salariati. Observ, de asemenea, ca eroarea interceptului este extrem de mare, lucru care ar putea aduce scepticism in ceea ce priveste puterea explicativa a modelului. - Cand se angajeaza un salariat in plus in intreprindere, profitul net creste cu aproximativ 6075.03. - Valoarea coeficientul de determinare (R-Squared) imi spune ca modelul meu explica in proportie de 26% variatia cifrei de afaceri in functie de variatia numarului de salariati. E posibil sa existe alte variabile care ar putea explica mai bine variatia profitului net decat numarul de salariati.

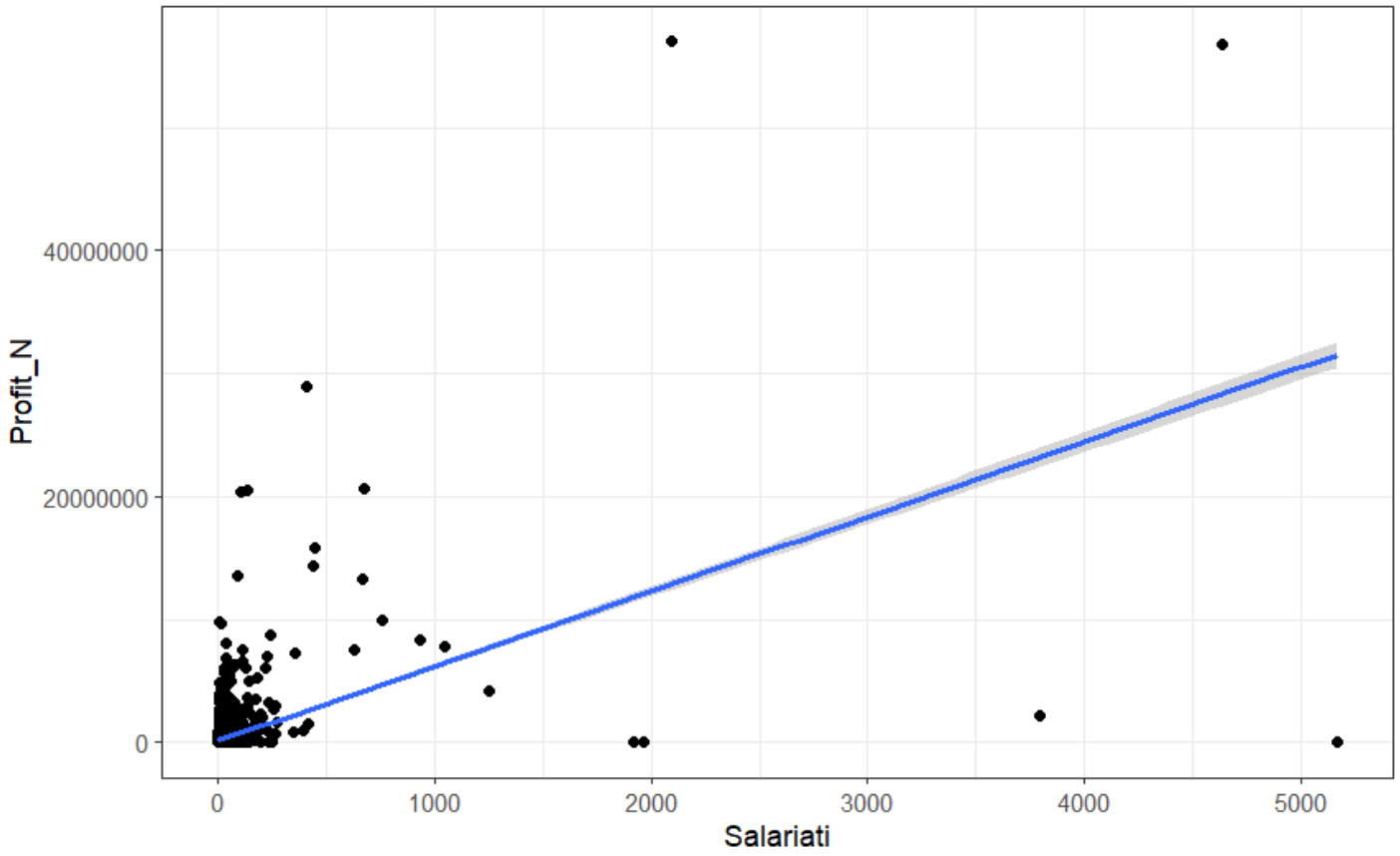
Hide

```
options(scipen=999)
summary(lm.beta::lm.beta(m1))
```

```
Call:
lm(formula = Profit_N ~ Salariati, data = Bilant2)

Residuals:
      Min        1Q    Median        3Q       Max
-31536518   -127781   -114833    -51284  44215808

Coefficients:
               Estimate Standardized  Std. Error t value            Pr(>|t|)
(Intercept) 116442.9599           NA  10148.0759   11.47 <0.0000000000000002 ***
Salariati     6075.0339       0.5126    106.1776   57.22 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 968200 on 9187 degrees of freedom
Multiple R-squared:  0.2627,    Adjusted R-squared:  0.2626
F-statistic:  3274 on 1 and 9187 DF,  p-value: < 0.00000000000000022
```

Conform testului T, variabila Salariati este semnificativa atunci cand vrem sa oferim o explicatie in ceea ce priveste variatia Profitului net. Practic, avem 0 dovezi in favoarea ipotezei nule intrucat p<0.01. De asemenea, testul F compara ceea ce modelul reuseste sa explice cu ceea ce nu reuseste sa explice si, in acest caz, este semnificativ (p<0.01), ceea ce inseamna ca modelul de regresie explica suficient de bine variata cifrei de afaceri in functie de numarul de Salariati. De asemenea, coeficientul de regresie standardizat imi indica faptul ca o crestere cu o abatere standard a numarului de salariati este asociata cu o crestere medie a profitului net cu 0.51 abateri standard.

In cele ce urmeaza, voi investiga, prin intermediul intervalelor de confidenta, puterea explicativa a modelului.



Observ si aici ca punctele sunt in afara intervalului marcat cu gri ceea ce inseamna ca erorile sunt prea mari pentru aceste cazuri (asa cum am si observat in tabelul de coeficienti) si o parte importanta din variatia profitului net nu este explicata de modelul de regresie.

Pentru a intelege mai bine ecuatia de regresie, voi analiza valorile observate, valorile estimate si reziduurile.

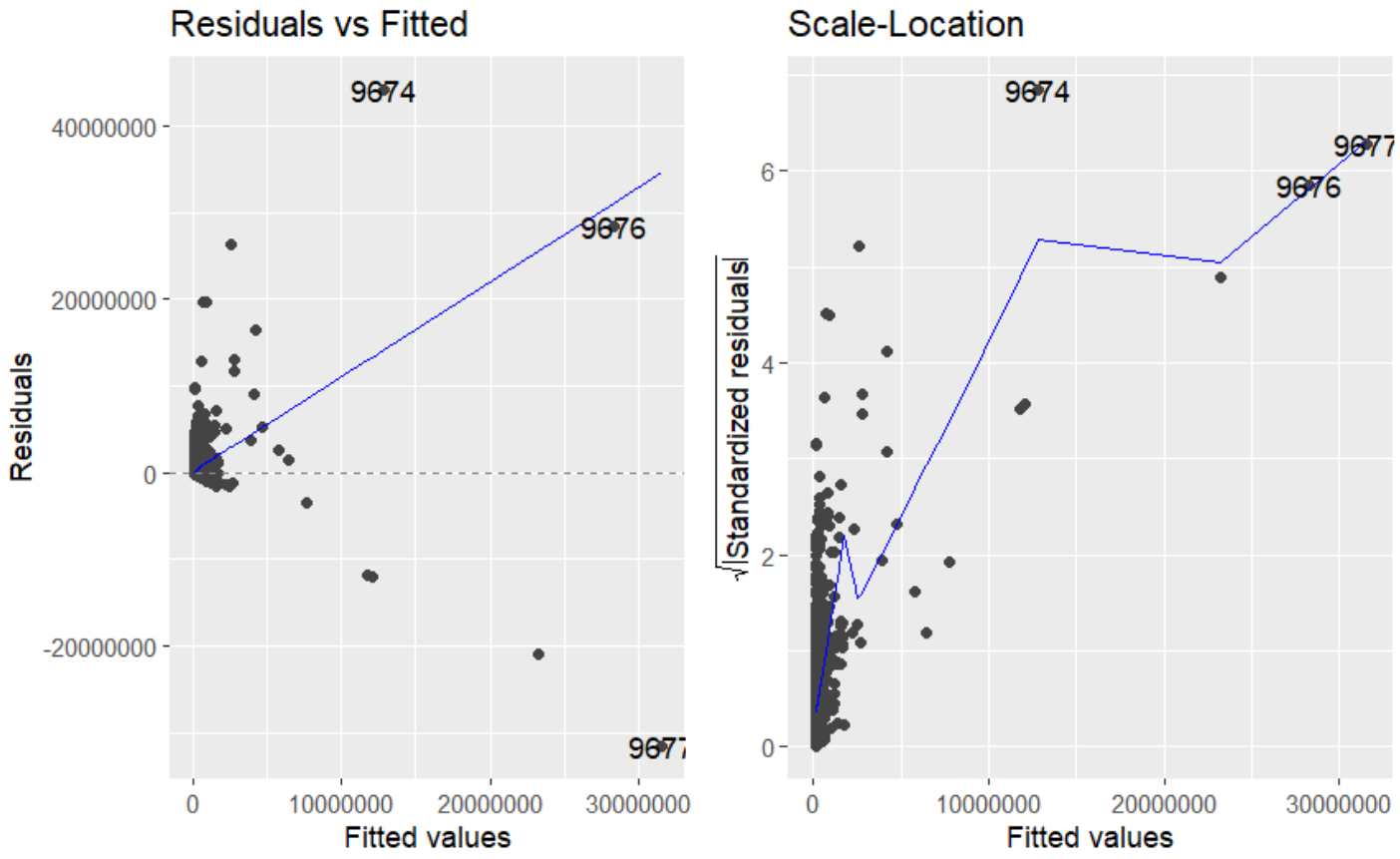
	Bilant2.Salariati <dbl>	Bilant2.Profit_N <dbl>	m.fitted.values <dbl>	m.residuals <dbl>
1	0	402102	285448.9	18774052.1
2	0	887111	285448.9	4737647.1
3	0	529171	285448.9	4686883.1
4	0	581540	285448.9	2779953.1
5	0	172050	285448.9	727554.1
6	0	99	285448.9	521159.1

6 rows

La inspectarea tabelului, observ ca variabila “Bilant2.Salariati” contine foarte multe valori nule (lucru care era de asteptat conform graficelor anterioare unde aparea acel conglomerat). Variabila “Bilant2.Profit_N” reprezinta valorile observate profitul net, iar m.fitted.values reprezinta valoarea prezisa prin intermediul ecuatiei de regresie. Erorile sunt extrem de mari, lucru care indica faptul ca modelul meu nu reuseste sa estimeze valori cat mai aproape de valorile observate. Cauza poate fi reprezentata de faptul ca exista si alte variabile necunoscute care pot explica mai bine variatia profitului net.

- Testarea Asumptiilor pentru a doua analiza de regresie

1. Egalitatea variantelor (Homoscedascitatea)



Observ ca asumptia este incalcata pentru ca: 1. avem outliers (9186, 9188, 9189) - problema care s-ar putea remedia prin simpla eliminare a acestora 2. ca sa existe homoscedasticitate, linia punctata ar trebui sa se suprapuna intr-o oarecare masura peste linia albastra, lucru care in graficul de mai sus nu se intampla. Observ o suprapunere in jurul valorii 0 si intuiesc ca ar putea fi vorba de o distributie asimetrica. 3. fluctuatiile ale diferentelor de varianta in functie de valorile prezise sunt destul de mari

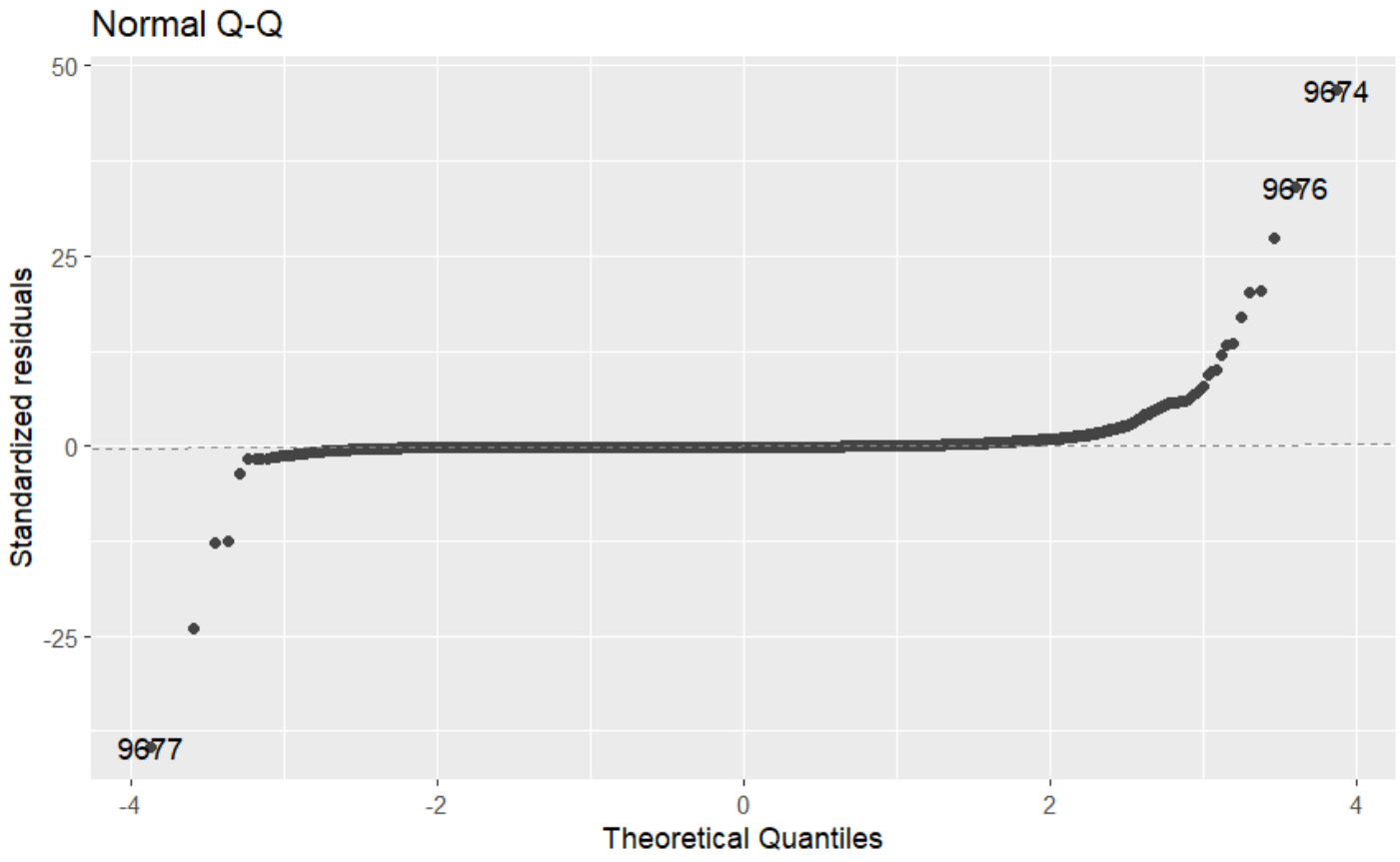
Hide

```
library(lmtest)
gqtest(m1, data=Bilant2)
```

```
    Goldfeld-Quandt test

data:  m1
GQ = 61.988, df1 = 4593, df2 = 4592, p-value < 0.00000000000000022
alternative hypothesis: variance increases from segment 1 to 2
```

Ipoteza nula spune ca exista homoscedasticitate. Avand in vedere faptul ca p<0.01, se respinge ipoteza nula si se confirma faptul ca asumptia omogenitatii variantelor este incalcata.

2. Normalitate (distributia erorilor trebuie sa fie normala)



Din nou, observ prezenta valorilor extreme care sunt ingrijoratoare, ceea ce face ca forma distributiei erorilor sa devieze de la normalitate. De asemenea, observ ca linia punctata nu se abate de la valoarea 0, ceea ce indica o distributie puternic asimetrica cu un conglomerat de date in jurul acestei valori (aspect observat si in reprezentarile vizuale anterioare).

Hide

```
shapiro.test(m1$residuals[0:5000])
```

```
    Shapiro-Wilk normality test

data:  m1$residuals[0:5000]
W = 0.30739, p-value < 0.00000000000000022
```

Deviatia de la normalitate este confirmata si statistic. Asadar, cele doua asumptii sunt incalcate. Acest lucru inseamna ca, desi am obtinut semnificativitate statistica in ceea ce priveste modelul (vezi R-squared, testul T si F), modelul meu nu este unul robust, adica variatia numarului de salariati nu reuseste sa explice variatia profitului net pentru ca exista anormalitati (valori extreme, distributie asimetrica cu conglomerat de date in jurul valorii 0, diferente foarte mari intre valorile observate si cele estimate).

Daca ar fi sa ignoram faptul ca modelele nu sunt robuste si nu pot estima cat de cat acurat valorile observate, consider ca primul model (Salariati~CA) are puterea de predictie mai mare. Coeficientii standardizati si coeficientul de determinare sunt elementele cheie prin care pot compara doua modele. Observam ca in cazul in care variabila dependenta este cifra de afaceri, avem un coeficient de regresie standardizat b=0.83 si coeficientul de determinare ajustat Adjusted R-Square = 0.70, ceea ce inseamna ca variatia numarului de salariati explica, in proportie de 70%, variatia cifrei de afaceri. Pe de alta parte, daca variabila dependenta este profitul net,

coeficientul de regresie standardizat b2=0.51, iar coeficientul de determinare ajutat Adjusted R-Square2 = 0.26, inseamna ca variatia numarului de salariati explica, in proportie de 26%, variatia profitului net. Totusi, este clar ca exista si alte variabile care pot explica mai bine variatiile atat la nivel de cifra de afaceri, cat si la nivel de profit net. Mai mult, puterea predictiva este redundanta avand in vedere faptul ca modelele nu sunt robuste si nici reliable din cauza anormalitatilor.

2. Realizați o nouă variabilă pentru a compara veniturile (Venit_T) și pierderile (Pierdere_N), după mărimea companiei în funcției de numărul de salariați (Salariați).

A. Folosiți clasificarea de mai jos pentru a defini noua variabilă (0,5p)

Microîntreprindere < 10 angajați

Întreprindere mică < 50 angajați

Întreprindere mijlocie < 250 angajați

Întreprindere mare => 250 angajați

B. Arați care sunt principalele domenii de activitate (Secțiuni) ale întreprinderilor mari și ale celor mijlocii (0.5p)

Hide

```
Bilant_$Tip_intreprindere[Bilant_$Salariati < 10] <- "Microintreprindere"
Bilant_$Tip_intreprindere[Bilant_$Salariati >10 & Bilant_$Salariati<50] <- "Intreprindere mica"
Bilant_$Tip_intreprindere[Bilant_$Salariati>50 & Bilant_$Salariati<250] <- "Intreprindere mijloc
ie"
Bilant_$Tip_intreprindere[Bilant_$Salariati >= 250] <- "Intreprindere mare"
ordered (Bilant_$Tip_intreprindere,
         levels =c("Microintreprindere", "Intreprindere mica", "Intreprindere mijlocie", "Intrep
rindere mare"))
```

Hide

```
table(Bilant_$Tip_intreprindere)
```

```
    Intreprindere mare     Intreprindere mica Intreprindere mijlocie
                    24                    901                    150
    Microintreprindere
                  8473
```

Hide

```
x <- table(Bilant_$Sectiune, Bilant_$Tip_intreprindere)
addmargins(x)
```

Intreprindere mare
  #N/A
0
  Activităţi Ale Gospodăriilor Private În Calitate De Angajator De Personal Casnic; Activităţi A
le Gospodăriilor Private De Producere De Bunuri Şi Servicii Destinate Consumului Propriu
0
  Activităţi De Servicii Administrative Şi Activităţi De Servicii Suport
0
  Activităţi De Spectacole, Culturale Şi Recreative
0
  Activităţi Profesionale, Ştiinţifice Şi Tehnice
0
  Administraţie Publică Şi Apărare; Asigurări Sociale Din Sistemul Public
0
  Agricultură, Silvicultură Şi Pescuit
0
  Alte Activităţi De Servicii
0
  Comerţ Cu Ridicata Şi Cu Amănuntul; Repararea Autovehiculelor Şi Motocicletelor
0
  Construcţii
1
  Distribuţia Apei; Salubritate, Gestionarea Deşeurilor, Activităţi De Decontaminare A Terenuri
lor
1
  Hoteluri Şi Restaurante
0
  Industria Extractivă
0
  Industria Prelucrătoare
20
  Informaţii Şi Comunicaţii
0
  Intermedieri Financiare Şi Asigurari
1
  Învăţământ
0
  Producţia Şi Furnizarea De Energie Electrică Şi Termică, Gaze, Apă Caldă Şi Aer Condiţionat
0
  Sănătate Şi Asistenţă Socială
0
  Transport Şi Depozitare
1
  Tranzacţii Imobiliare
0
  Sum
24

Intreprindere mica

  #N/A
0
  Activităţi Ale Gospodăriilor Private În Calitate De Angajator De Personal Casnic; Activităţi A
le Gospodăriilor Private De Producere De Bunuri Şi Servicii Destinate Consumului Propriu
0
  Activităţi De Servicii Administrative Şi Activităţi De Servicii Suport
33
  Activităţi De Spectacole, Culturale Şi Recreative
1
  Activităţi Profesionale, Ştiinţifice Şi Tehnice
26
  Administraţie Publică Şi Apărare; Asigurări Sociale Din Sistemul Public
0
  Agricultură, Silvicultură Şi Pescuit
29
  Alte Activităţi De Servicii
6
  Comerţ Cu Ridicata Şi Cu Amănuntul; Repararea Autovehiculelor Şi Motocicletelor
214
  Construcţii
173
  Distribuţia Apei; Salubritate, Gestionarea Deşeurilor, Activităţi De Decontaminare A Terenuri
lor
6
  Hoteluri Şi Restaurante
60
  Industria Extractivă
13
  Industria Prelucrătoare
218
  Informaţii Şi Comunicaţii
13
  Intermedieri Financiare Şi Asigurari
3
  Învăţământ
3
  Producţia Şi Furnizarea De Energie Electrică Şi Termică, Gaze, Apă Caldă Şi Aer Condiţionat
1
  Sănătate Şi Asistenţă Socială
14
  Transport Şi Depozitare
85
  Tranzacţii Imobiliare
3
  Sum
901

Intreprindere mijlocie
  #N/A
0
  Activităţi Ale Gospodăriilor Private În Calitate De Angajator De Personal Casnic; Activităţi A

le Gospodăriilor Private De Producere De Bunuri Și Servicii Destinate Consumului Propriu
0
  Activități De Servicii Administrative Și Activități De Servicii Suport
6
  Activități De Spectacole, Culturale Și Recreative
1
  Activități Profesionale, Științifice Și Tehnice
0
  Administrație Publică Și Apărare; Asigurări Sociale Din Sistemul Public
1
  Agricultură, Silvicultură Și Pescuit
0
  Alte Activități De Servicii
0
  Comerț Cu Ridicata Și Cu Amănuntul; Repararea Autovehiculelor Și Motocicletelor
17
  Construcții
20
  Distribuția Apei; Salubritate, Gestionarea Deșeurilor, Activități De Decontaminare A Terenuri
lor
4
  Hoteluri Și Restaurante
3
  Industria Extractivă
2
  Industria Prelucrătoare
80
  Informații Și Comunicații
3
  Intermedieri Financiare Și Asigurari
1
  Învățământ
0
  Producția Și Furnizarea De Energie Electrică Și Termică, Gaze, Apă Caldă Și Aer Condiționat
0
  Sănătate Și Asistență Socială
2
  Transport Și Depozitare
10
  Tranzacții Imobiliare
0
  Sum
150

Microintreprindere
  #N/A
0
  Activități Ale Gospodăriilor Private În Calitate De Angajator De Personal Casnic; Activități A
le Gospodăriilor Private De Producere De Bunuri Și Servicii Destinate Consumului Propriu
0
  Activități De Servicii Administrative Și Activități De Servicii Suport

249
  Activităţi De Spectacole, Culturale Şi Recreative
115
  Activităţi Profesionale, Ştiinţifice Şi Tehnice
719
  Administraţie Publică Şi Apărare; Asigurări Sociale Din Sistemul Public
0
  Agricultură, Silvicultură Şi Pescuit
290
  Alte Activităţi De Servicii
194
  Comerţ Cu Ridicata Şi Cu Amănuntul; Repararea Autovehiculelor Şi Motocicletelor
2242
  Construcţii
1278
  Distribuţia Apei; Salubritate, Gestionarea Deşeurilor, Activităţi De Decontaminare A Terenuri
lor
28
  Hoteluri Şi Restaurante
500
  Industria Extractivă
42
  Industria Prelucrătoare
1011
  Informaţii Şi Comunicaţii
204
  Intermedieri Financiare Şi Asigurari
82
  Învăţământ
64
  Producţia Şi Furnizarea De Energie Electrică Şi Termică, Gaze, Apă Caldă Şi Aer Condiţionat
16
  Sănătate Şi Asistenţă Socială
171
  Transport Şi Depozitare
1094
  Tranzacţii Imobiliare
174
  Sum
8473

Sum
  #N/A
0
  Activităţi Ale Gospodăriilor Private În Calitate De Angajator De Personal Casnic; Activităţi A
le Gospodăriilor Private De Producere De Bunuri Şi Servicii Destinate Consumului Propriu   0
  Activităţi De Servicii Administrative Şi Activităţi De Servicii Suport
288
  Activităţi De Spectacole, Culturale Şi Recreative
117
  Activităţi Profesionale, Ştiinţifice Şi Tehnice

```
745
  Administrație Publică Și Apărare; Asigurări Sociale Din Sistemul Public
1
  Agricultură, Silvicultură Și Pescuit
319
  Alte Activități De Servicii
200
  Comerț Cu Ridicata Și Cu Amănuntul; Repararea Autovehiculelor Și Motocicletelor
2473
  Construcții
1472
  Distribuția Apei; Salubritate, Gestionarea  Deșeurilor, Activități De Decontaminare A Terenuri
lor                                                                                          39
  Hoteluri Și Restaurante
563
  Industria Extractivă
57
  Industria Prelucrătoare
1329
  Informații Și Comunicații
220
  Intermedieri Financiare Și Asigurari
87
  Învățământ
67
  Producția Și Furnizarea De Energie Electrică Și Termică, Gaze, Apă Caldă Și Aer Condiționat
17
  Sănătate Și Asistență Socială
187
  Transport Și Depozitare
1190
  Tranzacții Imobiliare
177
  Sum
9548
```

Pentru intreprinderile mici și mijlocii, principalele domenii de activitate sunt urmatoarele comerţul cu ridicata și cu amănuntul; repararea autovehiculelor și motocicletelor, industria prelucrătoare și construcțiile.

Valori pentru fiecare in parte:

A. Intreprindere mică:

- Comerţ Cu Ridicata Şi Cu Amănuntul; Repararea Autovehiculelor Şi Motocicletelor = 214
- Industria Prelucrătoare = 218
- Construcții = 173

B. Intreprindere mijlocie:

- Industria Prelucrătoare = 80
- Constructii = 20
- Comerţ Cu Ridicata Şi Cu Amănuntul; Repararea Autovehiculelor Şi Motocicletelor = 17

Examen_Data_II - Cerinta 3

Code

3. Calculați perioada de la înființare folosind variabila „An” și momentul 2020. (0,5p)

a. Calculați media perioadei de la înființare ținând cont de mărimea firmei (după numărul de angajați - folosiți variabila creată la punctul 2.a.). (0,5p)

b. Testați ipoteza nulă conform căreia nu există diferențe semnificative ale perioadei de la înființare ținând cont de cele 4 categorii de firme. (0,5p)

c. Care este concluzia în urma aplicării metodei corespunzătoare? (0,5p)

Am creat o noua coloana in tabel prin care se calculeaza perioada de la infiintare, folosind “AN” si momentul 2020:

Hide

```
Bilant_$Durata <- (2020 - Bilant_$AN)
```

ANOVA ONE-WAY

VD = perioada de la infiintare (Durata)

VI = marimea firmei (4 categorii)

Hide

```
summary(Bilant_$Durata)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  -1.00    3.00    8.00   10.04   16.00   29.00 
```

Variabila nu prezinta valori lipsa. Voi calcula media perioadei de la infiintare in functie de marimea firmei.

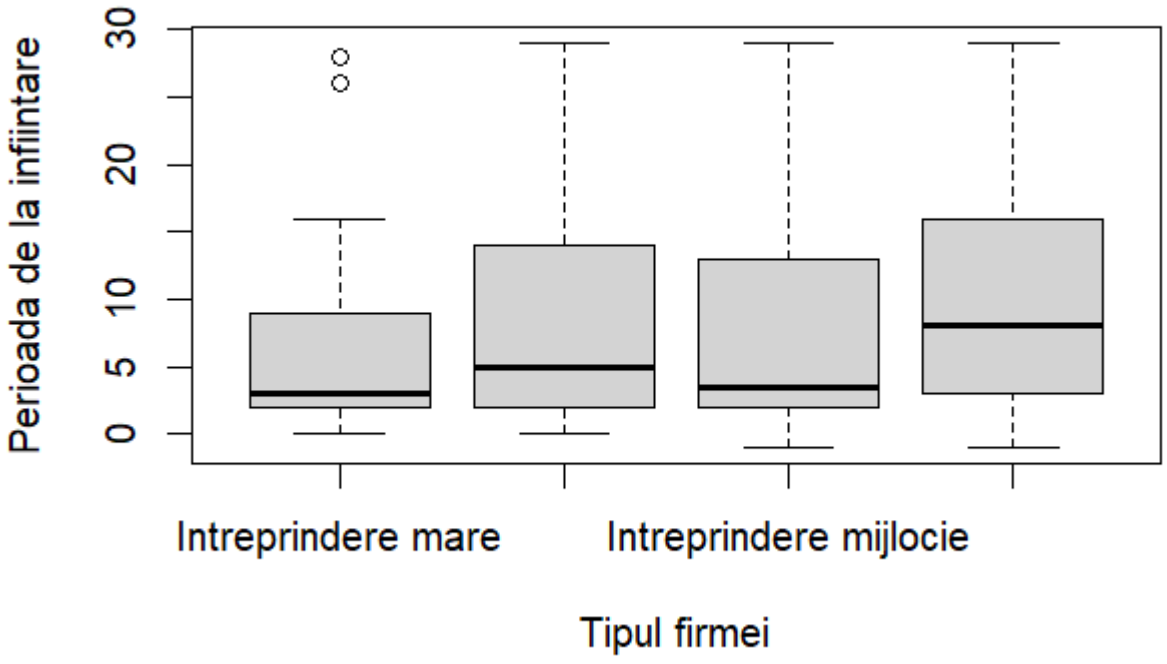
Hide

```
descriptive <- function(x) list(
                Total = length(x),
                Media = mean(x),
                Ab.std = sd(x)
                )
tapply(Bilant_$Durata, Bilant_$Tip_intreprindere, descriptive)
```

```
$`Intreprindere mare`
$`Intreprindere mare`$Total
[1] 24

$`Intreprindere mare`$Media
[1] 7.416667

$`Intreprindere mare`$Ab.std
[1] 8.423139


$`Intreprindere mica`
$`Intreprindere mica`$Total
[1] 901

$`Intreprindere mica`$Media
[1] 8.690344

$`Intreprindere mica`$Ab.std
[1] 8.420386


$`Intreprindere mijlocie`
$`Intreprindere mijlocie`$Total
[1] 150

$`Intreprindere mijlocie`$Media
[1] 7.386667

$`Intreprindere mijlocie`$Ab.std
[1] 7.770869


$Microintreprindere
$Microintreprindere$Total
[1] 8473

$Microintreprindere$Media
[1] 10.66175

$Microintreprindere$Ab.std
[1] 8.744367
```

Medie microintreprindere = 10.66 Medie intreprindere mica = 8.69 Medie intreprindere mijlocie = 7.38 Medie intreprindere mare = 7.41

La o prima vedere, observ ca exista cateva diferente care ar putea fi semnificative, mai ales in ceea ce priveste categoria “macrointreprindere” in comparatie cu celelalte. Pentru inceput, voi compara mediile categoriilor prin intermediul graficelor. Ma intereseaza sa vad cum arata distributiile pentru fiecare grup in parte.



In cazul categoriei “intreprindere mare”, observ ca avem valori extreme. Totusi, distributiile arata diferit in fiecare grup.

- Realizez ANOVA si verific asumptiile

Ipoteza nula: Mediile perioadei de la infiintare sunt egale pentru toate cele 4 categorii de firme.

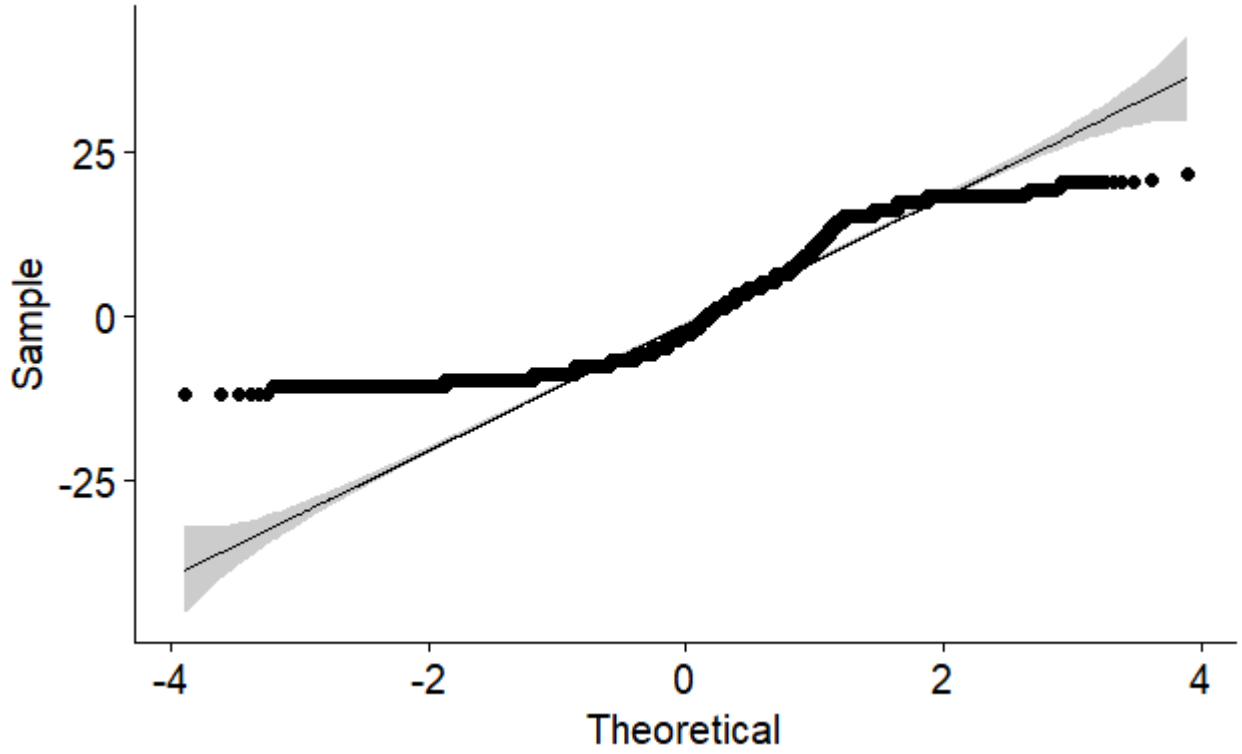
Hide

```
anova <-aov(Durata~Tip_intreprindere, data=Bilant_)
summary(anova)
```

```
                   Df Sum Sq Mean Sq F value             Pr(>F)
Tip_intreprindere    3   4787  1595.7   21.09 0.000000000000131 ***
Residuals         9544 722245    75.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
4188 observations deleted due to missingness
```

Rezultatul este semnificativ, intrucat p<0.01. Astfel, ipoteza nula este respinsa, adica exista diferente intre mediile perioadei de la infiintare pentru cele 4 categorii de firme.

1. Verificarea normalitatii (ma intereseaza ca distributia reziduurilor sa fie una normala)



Conform graficului, observ ca asumptia normalitatii este incalcata.

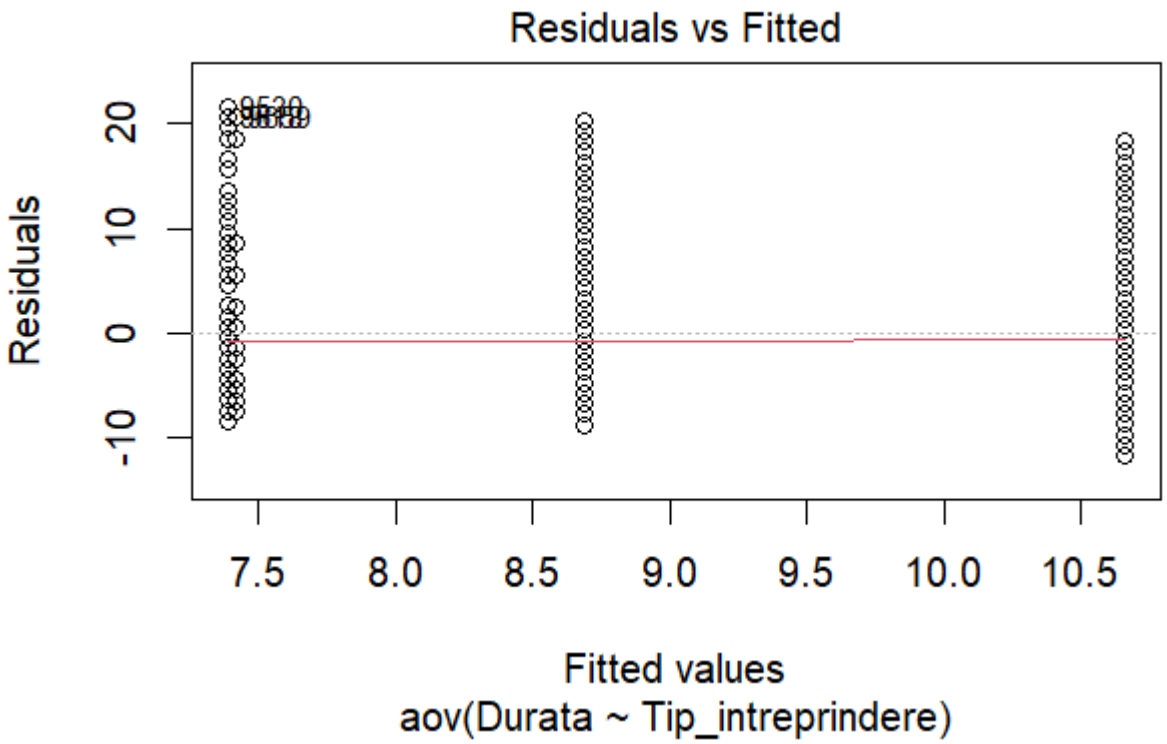
Hide

```
shapiro.test(reziduuri.anova[0:5000])
```

```
    Shapiro-Wilk normality test

data:  reziduuri.anova[0:5000]
W = 0.91876, p-value < 0.00000000000000022
```

Deviatia de la normalitate este confirmata si statistic.

2. Verificarea egalitatii variantelor (homoscedasticitate)



Conform graficului, asumptia egalitatii variantelor nu este incalcata, intrucat media reziduurilor este complet orizontala si aproape centrata de 0.

Hide

```
bartlett.test(Bilant_$Durata, Bilant_$Tip_intreprindere)
```

```
    Bartlett test of homogeneity of variances

data:  Bilant_$Durata and Bilant_$Tip_intreprindere
Bartlett's K-squared = 5.8636, df = 3, p-value = 0.1184
```

Ipoteza nula spune ca exista homoscedasticitate. Conform testului Bartlett, exista 18% dovezi in favoarea ipotezei nule (p>0.01), ceea ce indica prezenta omogenitatii variantelor.

Intrucat doar asumptia normalitatii este incalcata, voi aplica testul Kruskal-Wallis ca alternativa ce poate oferi rezultate robuste.

Ipoteza nula in cadrul testului Kruskal-Wallis: Medianele perioadei de la infiintare sunt egale pentru toate cele 4 categorii de firme.

Hide

```
kruskal.test(Bilant_$Durata, Bilant_$Tip_intreprindere)
```

```
    Kruskal-Wallis rank sum test

data:  Bilant_$Durata and Bilant_$Tip_intreprindere
Kruskal-Wallis chi-squared = 84.738, df = 3, p-value <
0.00000000000000022
```

Testul este semnificativ statistic (p<0.01), astfel se respinge ipoteza nula. Cel putin 2 valori mediane ale perioadei de la infiintare sunt diferite intre ele in functie de categoria firmei.

Pentru a compara categoriile intre ele, voi aplica teste post-hoc. In acest caz, voi aplica testul Dunn.

Hide

```
dunnTest(Durata ~ Tip_intreprindere,
         data=Bilant_,
         method="bonferroni")
```

```
Warning: Tip_intreprindere was coerced to a factor.Warning: Some rows deleted from 'x' and 'g' b
ecause missing data.Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Bonferroni method.
```

Z <dbl>	P.unadj <dbl>	P.adj <dbl>
-0.72964521	0.46560707872763162918	1.00000000000000000000
0.07294813	0.94184739126753780347	1.00000000000000000000
1.89314365	0.05833876845165902880	0.35003261070995417280
-2.03133175	0.04222135280268162622	0.25332811681608974340
-7.54313078	0.00000000000004588206	0.00000000000027529240
-5.23579333	0.00000016427769676239	0.00000098566618057430

6 rows | 2-4 of 4 columns

In urma aplicarii acestor analize pot concluziona urmatoarele: desi distributia reziduurilor deviaza de la normalitate, s-a demonstrat semnificativitate statistica in ceea ce priveste existenta unor diferente in perioada de la infiintare pentru fiecare categorie de firma. Acest lucru a fost posibil prin aplicarea testului Kruskal Wallis. Astfel, am creat un model robust rezistent la abaterile de la normalitate. De asemenea, conform testului Dunn, “intreprindere mica” si “microintreprindere” difera semnificativ una fata de cealalta (adj.p<0.01), la fel si “intreprindere mijlocie” fata de “microintreprindere” (adj.p<0.01).