Danmarks Tekniske Universitet

Anvendt medicinsk statistik
62518

Projekt 2

Troels Engsted Kiib s205492
December 27, 2023

Abstract

Your abstract.

Indholdsfortegnelse

1 Introduktion

I denne rapport vil jeg arbejde med et datasæt der beskriver lipider i blodet. Jeg vil undersøge dataen, samt beskrive sammenhænget imellem alder, køn og cholesterol. Derefter er LDL udregnet via. en formel, vi ved dog ikke hvilken ud af 2, så derfor vil jeg vurderer hvilken formel der er blevet brugt. Til slut vil jeg undersøge om mænd har en laverer HDL end kvinder med en margin.

2 Opgaver

I Dette kapitel vil jeg gennemgå opgaverne samt forklarer fremgangsmåden og resultatet. Opgaverne er udregnet vha. [Bla15]

2.1 Opgave 1: Beskriv dataene

I denne opgave undersøger jeg datasættet med passende grafer og et summary kald, for at lave en visuel undersøgelse.

2.1.1 Sex

Male	Female
3302	3482

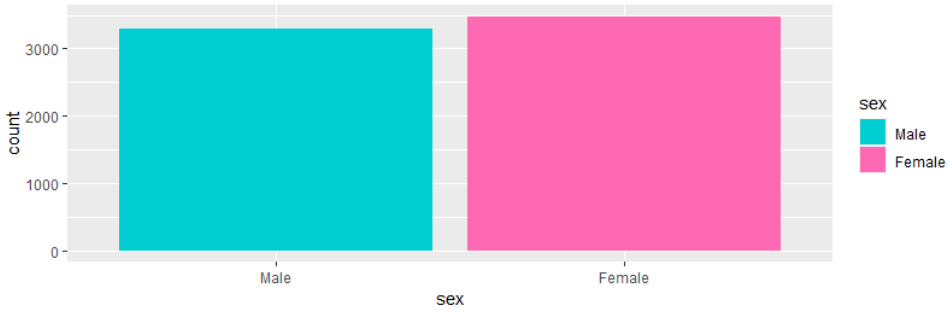


Figure 1: Male Female

2.1.2 Alder

Min.	1st Qu.	Median.	Mean.	3rd Qu.	Max.
29.69	39.98	45.09	46.05	50.26	61.34

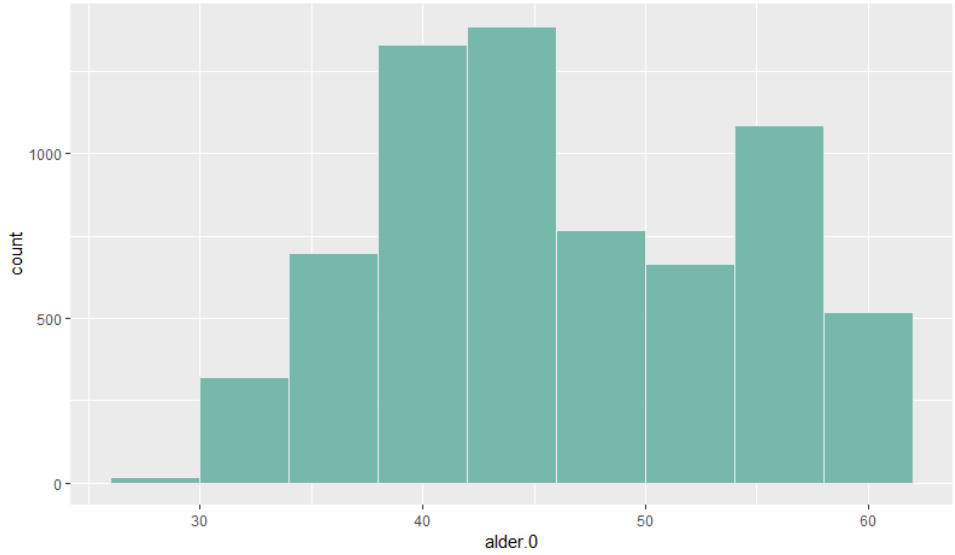


Figure 2: Alder

2.1.3 Cholesterol

Min.	1st Qu.	Median	Mean.	3rd Qu.	Max.	NA's
2.300	4.800	5.400	5.526	6.200	16.000	45

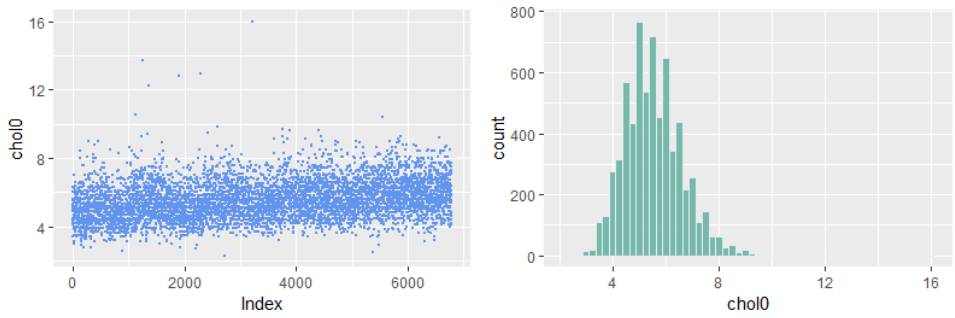


Figure 3: Cholesterol

2.1.4 HDL

Min.	1st Qu.	Median	Mean.	3rd Qu.	Max.	NA's
0.250	1.130	1.380	1.427	1.660	4.420	45

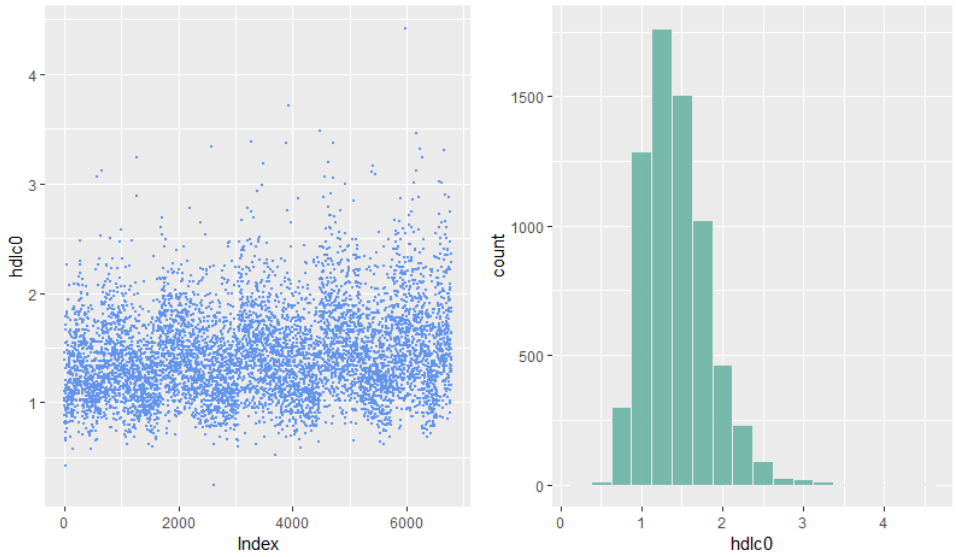


Figure 4: HDL

2.1.5 LDL

Min.	1st Qu.	Median	Mean.	3rd Qu.	Max.	NA's
0.8	2.8	3.4	3.5	4.1	11.3	128

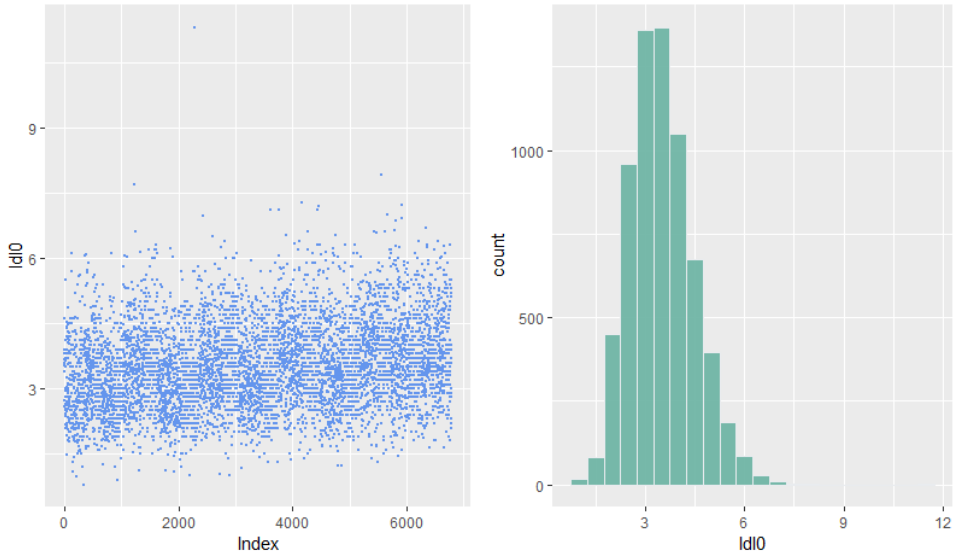


Figure 5: LDL

2.1.6 VDL

Min.	1st Qu.	Median	Mean.	3rd Qu.	Max.	NA's
0.1000	0.3636	0.5000	0.5750	0.7000	2.3000	128

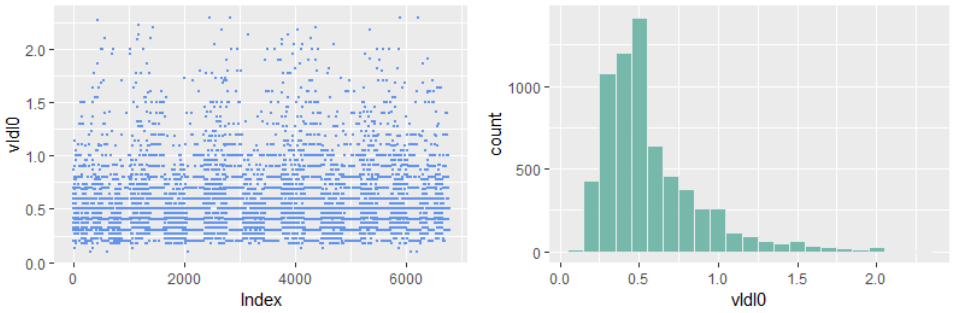


Figure 6: VDL

2.1.7 Triglycerid

Ved Triglycerid kan man se, at der er et behov for at transformere dataen på grund af de meget store svingninger. Deraf bruger vi log10.

Min.	1st Qu.	Median	Mean.	3rd Qu.	Max.	NA's
0.300	0.800	1.100	1.352	1.600	60.400	45

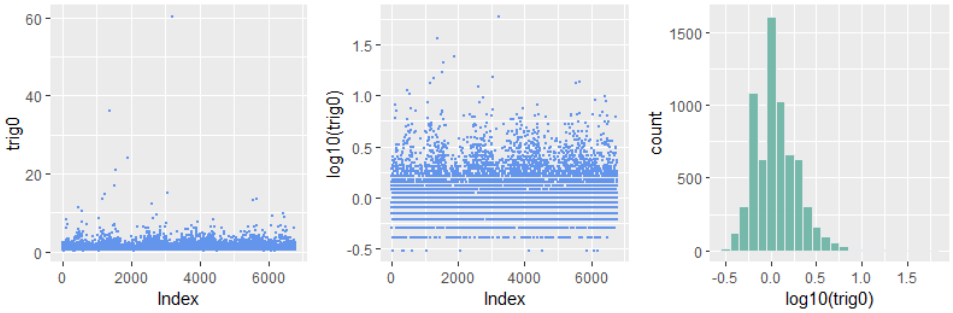


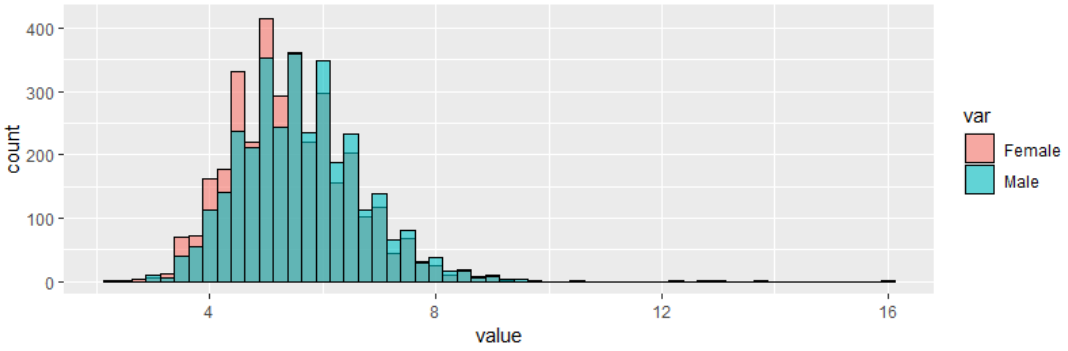
Figure 7: Triglycerid

2.2 Opgave 2: Vurder om der er en forskel i total cholesterol i forhold til kønnet

For at vurdere om der er en forskel i den totale mængde af cholesterol i datasættet, laver vi nu nogle subsets ud af vores datasæt. Dette gør vi for at adskille mænd og kvinder, som derefter plottes i samme graf. For at vurderer om der er en forskel udfører vi en T-test, hvorved vi optiller nulhypotesen:

H0: Der er ingen forskel imellem datasættene HA: Der er en forskel imellem datasættene

Og da vores test på figur: 8 viser at vi får en p-værdi under 0.05, må vi forkaste vores nulhypotese, og konkluderer at vi må vurder at der er en forskel i det totale cholesterol i forhold til kønnet.



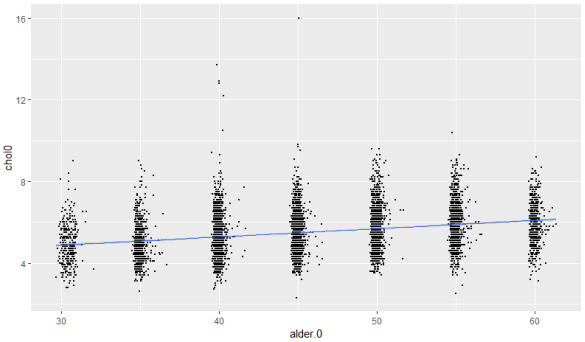

```
        Welch Two Sample t-test

data:  maleset$chol0 and femaleset$chol0
t = 6.7115, df = 6726.6, p-value = 2.083e-11
alternative hypothesis: true difference in means is not
 equal to 0
95 percent confidence interval:
 0.1253432 0.2287749
sample estimates:
mean of x mean of y
 5.616920  5.439861
```

Figure 8: Histogram og T-test

2.3 Opgave 3: Vurder afhængigheden mellem alder og total cholesterol med en lineær regression

For at undersøge sammenhænget imellem alderen og det totale cholesterol, laves der en lineær regressions model over dataen. På baggrund af vores lineær regressions, kan vi opstille den lineær model over forholdet således

$$TotaleCholesterol = alder \cdot 0.0406 + 3.6563 \tag{1}$$




```
Call:
lm(formula = proj2_data$chol0 ~ proj2_data$alder.0)

Coefficients:
       (Intercept)  proj2_data$alder.0
            3.6563              0.0406
```

Figure 9: Linær regressions model

2.4 Opgave 4: Udvid regressionsmodellen med køn og interaktion imellem alder og køn

Jeg udvider min linær regression til nu at være kønsopdelt, og placerer denne i samme plot, for at se sammenhænget. Deraf får vi to ligninger:

$$TotaleCholesterolMænd = alder \cdot 0.0293 + 4.2571 \tag{2}$$

$$TotaleCholesterolKvinder = alder \cdot 0.0503 + 3.1395 \tag{3}$$

Disse kan derefter sættes op imod hinanden og finde skæringspunktet, for at vurderer hvornår kvindernes cholesterol bliver højere end mændenes.

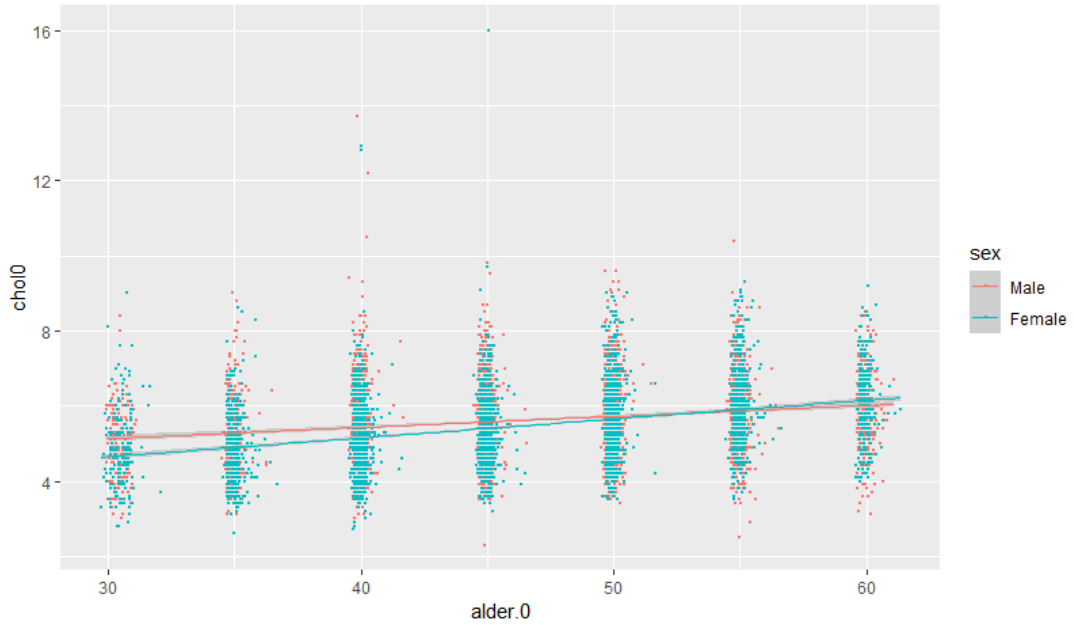


Figure 10: Totale Cholesterol vs alder

```
Call:
lm(formula = maleset$chol0 ~ maleset$alder.0)

Coefficients:
    (Intercept)  maleset$alder.0
         4.2571           0.0293
```

```
Call:
lm(formula = femaleset$chol0 ~ femaleset$alder.0)

Coefficients:
      (Intercept)  femaleset$alder.0
           3.1395             0.0503
```

Figure 11: Linær regressions model

2.5 Opgave 5: Vurder om residualerne fra overordnede model kan betagtes som følgende en normal fordeling

Herunder plotter jeg residualerne, de enkelte afstand fra punkterne til regressionslinjen. Ud fra plotsne herunder vil jeg konkluderer at residualerne er normaltfordelt.

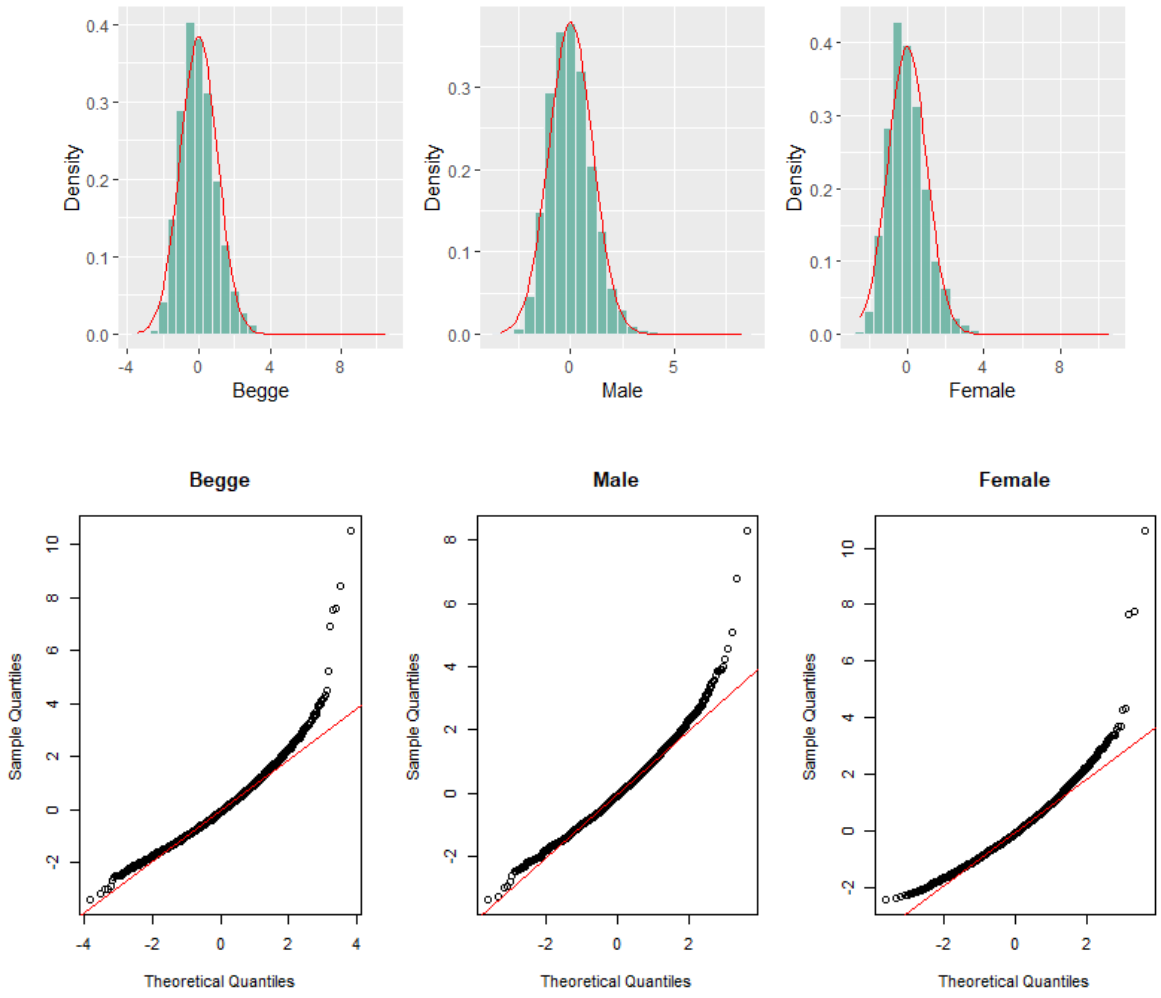


Figure 12: ResidualPlots til vurdering af normalfordelling

2.6 Opgave 6: Kør fit koden og beskriv så mange plots som muligt

2.6.1 Residusals vs Fitted

Det man ser her er residualerne plottet i forhold til de estimerede, vi kan bruge denne graf til at undersøge om der er evt. outliers og om vores datasæt opfører sig linær regulært. Her kan vi se, at residualerne gerne skal ligge nogenlunde stokastisk omkring vores 0 værdi.

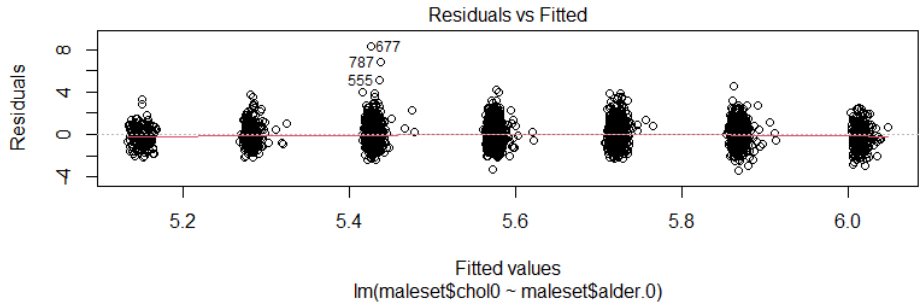


Figure 13: Residusals vs Fitted

2.6.2 Normal Q-Q

Vi kan bruge Q-Q plottet til at undersøge om dataen er normal fordelt. Dette bliver tydeligt gjort ved at dataen skal følge den estimerede linje hvis dataen er normal fordelt. Deraf kan man undersøge om dataen afviger fra linjen og deraf værende ikke normal fordelt.

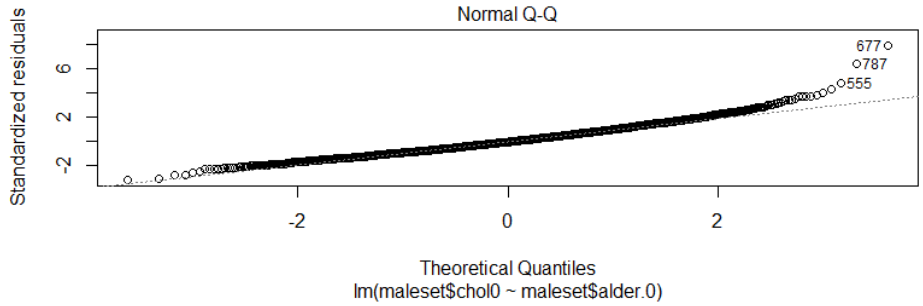


Figure 14: Normal Q-Q

2.6.3 Scale-Location

Er også kaldet spread-location plot, og undersøger om der er en homogenitet af variansen. Det vil sige, at der er en homogenitet i variansen og at den ikke vokser sig større og større. Her er en lige estimeret linje god, da denne så vil beskrive at variansen ikke vokser.

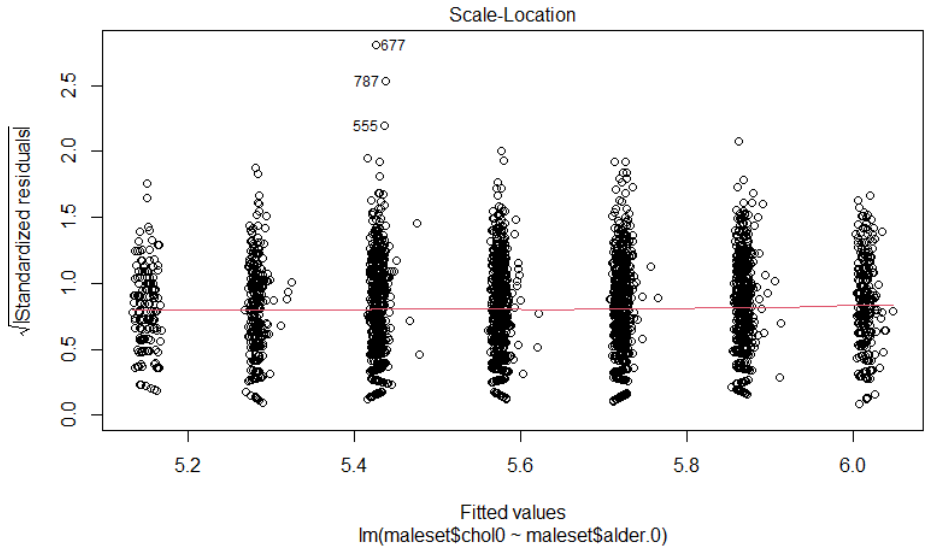


Figure 15: Scale-Location

2.6.4 Residusals vs Leverage

Dette plot undersøger om der er nogle indflydelsesrige datapunkter. Dette kunne f.eks. være en outlier der pga. lineær regression nu ændre vores resultat markant, pga. dets ekstreme værdi. Dette gør fordi, at selvom et datapunkt har en ekstrem værdi, er dette ikke ens med at den ødelægger vores regression, altså at resultatet ikke ville ændres hvis datapunktet blev eksluderet eller inkluderet.

Det vi kigger efter er om nogen af værdier går udover Cook's distance, som er en rød stiplet linje, som ikke kan ses på mit plot, og deraf kan jeg konkludere at ingen at settes evt. outliers har en markant effekt på resultatet.

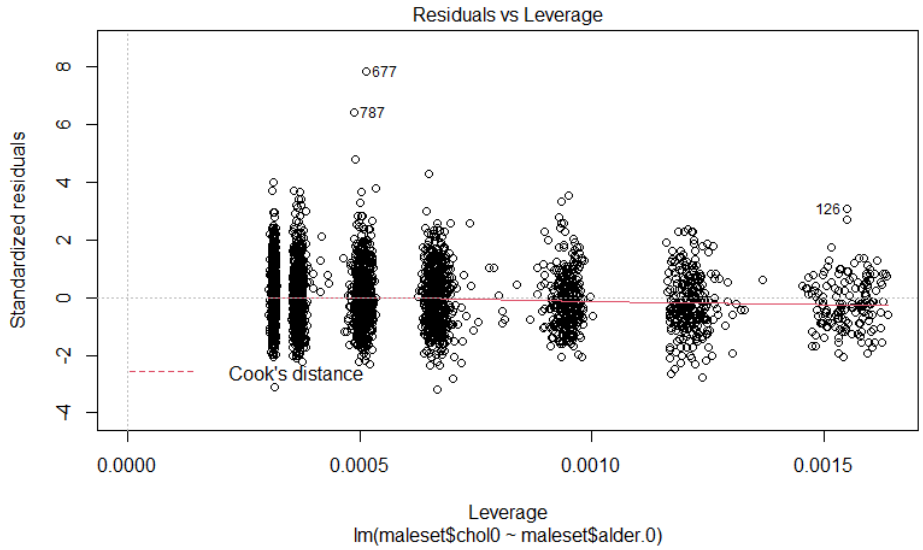


Figure 16: Residusals vs Leverage

2.7 Opgave 7: Bestem hvilken LDL formel der er blevet brugt

Her skal jeg finde ud af hvilken formel der er blevet brugt til at lave LDL i datasættet:

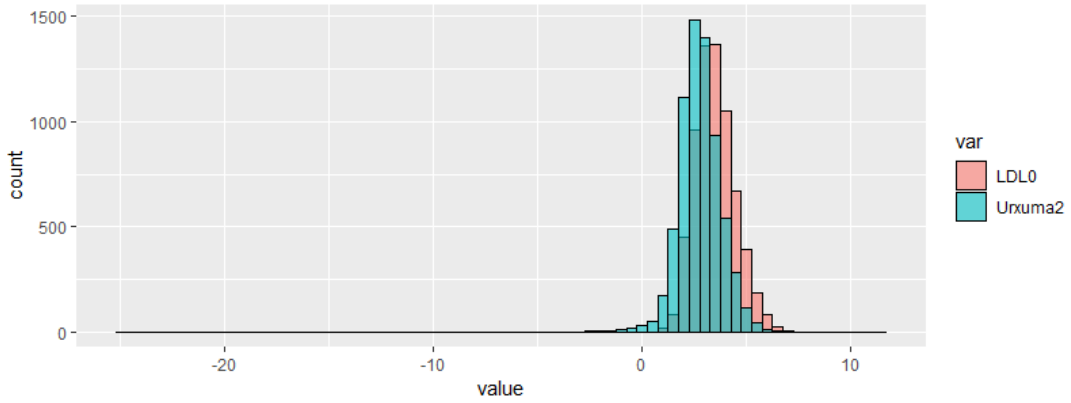
$$ldl = \alpha chol - \beta hdl - \rho triglycerid \tag{4}$$

$$ldl = \alpha chol - \beta hdl + \rho triglyceride + \phi \tag{5}$$

Deraf kan jeg genere det ene datasæt således at jeg har et datasæt der følger formel (4), hvorved jeg kan opstille hypotesen:

H0: Den udregnede LDL følger formel 1 HA: Den udregnde LDL følger ikke formel 1

På baggrund af udførte t-test herunder på figur: 17, må jeg konkluderer at nulhypotesen skal forkastes, og da LDL ikke kan være udregnet via formel (4), må den være udregnet via formel (5).




```
        Welch Two Sample t-test

data:  proj2_data$ldl0 and Formel1
t = 40.871, df = 13054, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal
 to 0
95 percent confidence interval:
 0.7092972 0.7807589
sample estimates:
mean of x mean of y
 3.500065  2.755037
```

Figure 17: Histogram og T-test

2.8 Opgave 8: Følger dataen hypotesen at mænd ikke er mere inferior end kvinder, med en 0,5 non-inferiority margin, i hdl

Hvis vi udregner middelværdier får vi disse til 1.293046 og 1.554901, for henholdvis mænd og kvinder. Deraf kan vi udfører en non-inferiority test hvor vi undersøger om mænd har en lavere HDL end kvinder med en non-inferiority margin på 0.5

$$H_0 : \mu_1 =< \mu_2 - k, H_A : \mu_1 > \mu_2 - k \tag{6}$$

Deraf kan vi udregne vores værdi, og se at mænds middelværdi ikke er mindre end kvindernes middelværdi efter at non-inferiority marginen er trukket fra.

$$1.293047 > 1.554901 - 0.5 \tag{7}$$

Deraf må vi forkaste vores nulhypotese og konkludere at mænd ikke har en hdl værdi der er mere end 0.5 lavere end kvinder.

3 Konklusion

På baggrund af de udrengede opgaver, må jeg konkluderer at jeg har undersøgt dataen, beskrevet sammenhængen mellem alder,køn og cholesterol, vurderet den korrekte formel der er brugt til udregnningen af LDL og undersøgt om mænd har en laverer HDL end kvinder med en margin på 0.5

References

[Bla15] Martin Bland. An introduction to medical statistics. (4), 23-07-2015.