

Statistik – Eksamensbesvarelse

1 Datapresentation

1.1) `d <- read.csv("http://causal.sund.ku.dk/f22/55.csv", header=TRUE, stringsAsFactors=TRUE)`

1.2) Nedenfor ses tabel 1:

Tabel 1	Crawlsvømning (Crawl = 1)	Andre svømmeformer (Crawl = 0)
Total i hver gruppe	53 (100 %)	87 (100 %)
Erosion		
1 (Ætsning), N (%)	18 (33.96 %)	6 (6.90 %)
0 (Ingen ætsning), N (%)	35 (66.04 %)	81 (93.10 %)
Soda		
Never: Aldrig, N (%)	8 (15.09 %)	6 (6.90 %)
< 5/week: Mindre end 5 gange pr. uge, N (%)	22 (41.51 %)	37 (42.53 %)
>= 5/week: Mere end eller lig med fem gange pr. uge, N (%)	16 (30.19 %)	37 (42.53 %)
NA: (Not Available) det har ikke været muligt at indsamle data, N (%)	7 (13.21 %)	7 (8.04 %)
Pooltime [timer/uge]		
Middelværdi (spredning)	9.56 (1.01)	10.53 (1.11)
Age [i år]		
Median (IQR)	14 (5)	12 (4)

Jeg har inddelt datasættet i to subgrupper:

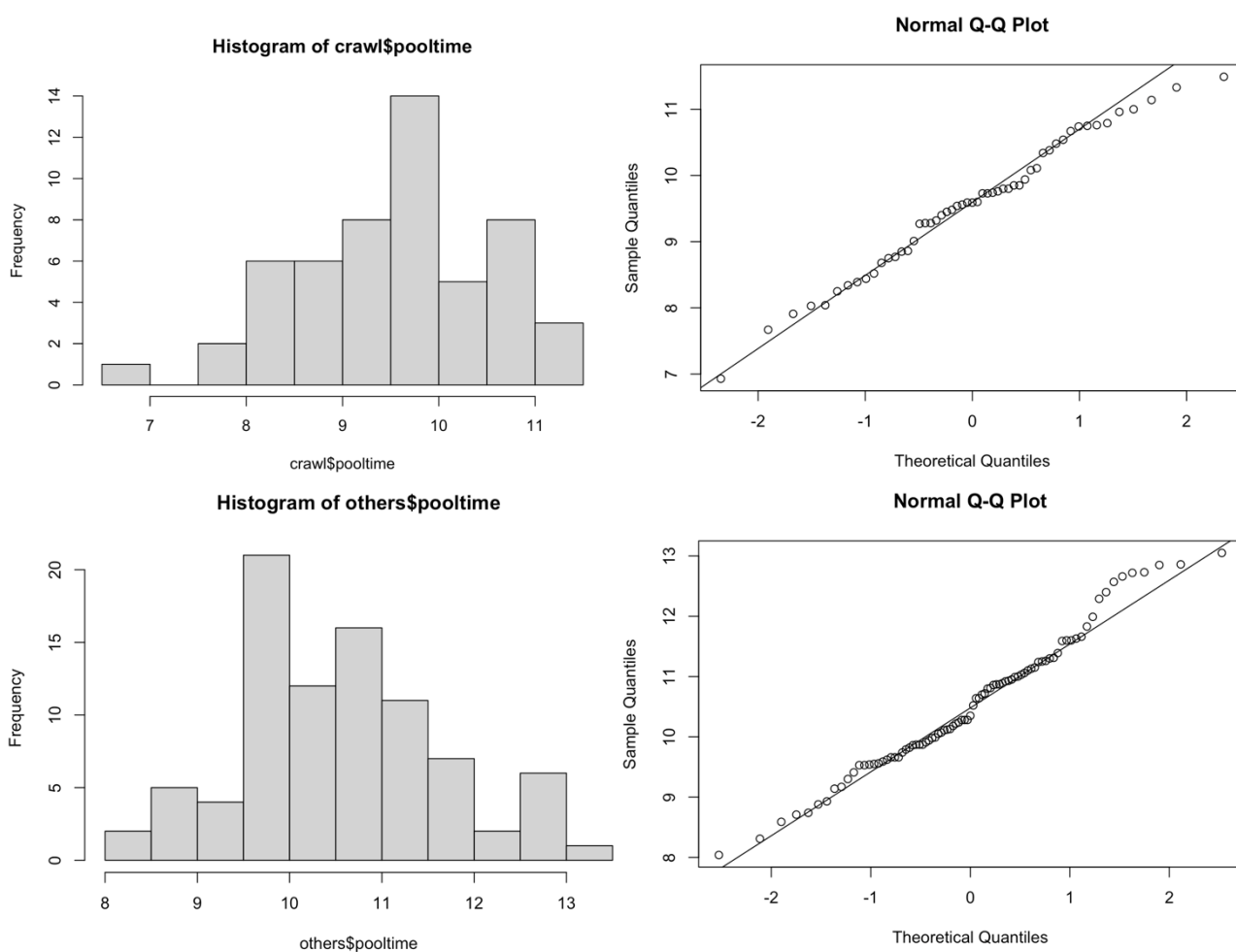
- Den ene subgruppe, som har `crawl = 1`, det vil sige de personer som benytter sig af svømmeformen: crawl.
- Den anden subgruppe, som har `crawl = 0`, det vil sige de personer som benytter sig af andre svømmeformer.

Den underliggende grund til at jeg har inddelt datasættet i to subgrupper, er fordi opgavebeskrivelsen blandt andet belyser at man ønsker at undersøge den risiko der forekommer ved crawlsvømning og andre svømmeformer, da indtagelsen af vand fra svømmebassinet under svømning kan afhænge af hvilken svømmeform som individet udøver.

Jeg har ud fra de to subgrupper opstillet en tabel, og udført deskriptiv statistik på de andre variabler, altså erosion og soda, pool time og age. Da variablen: erosion og soda er kategoriske variabler, kan jeg nemt inddеле grupperne i undergrupper, alt efter hvilken gruppe som personerne passer til i forhold til tandstatus og har af drikkevaner i forhold til sodavand. Derfor bestemmes andelen/frekvensen på disse variabler.

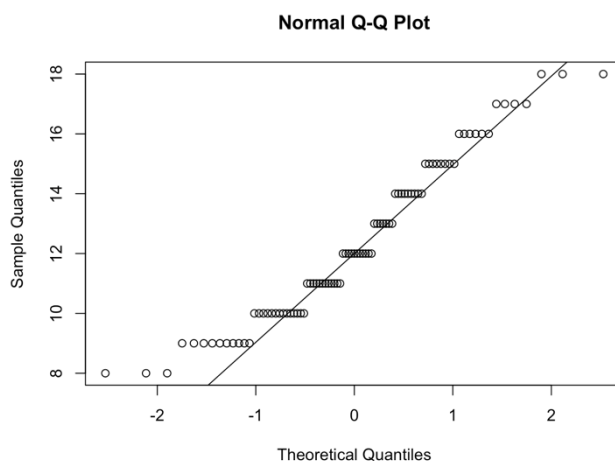
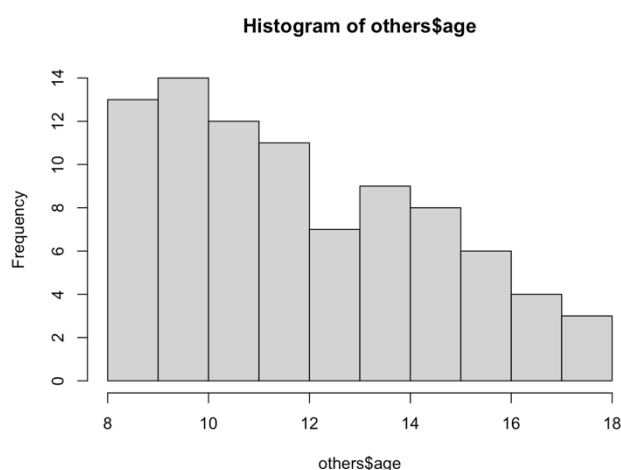
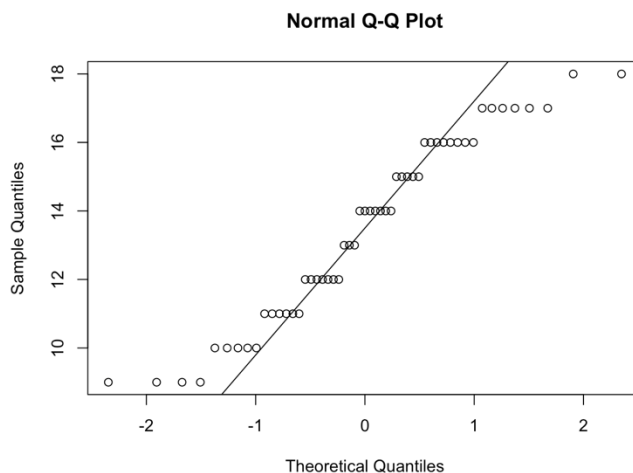
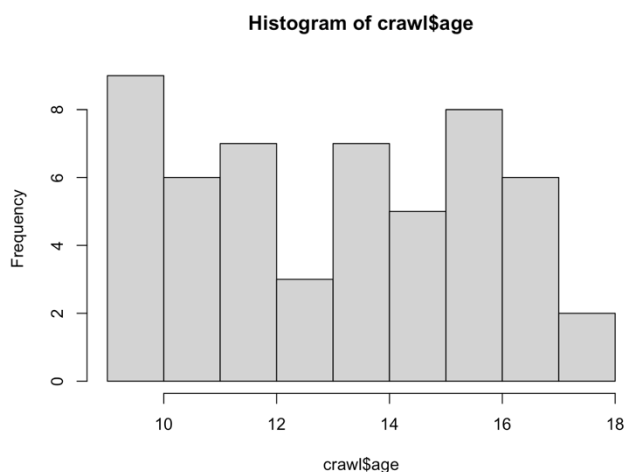
I forhold til variablerne pooltime og age som antager numeriske værdier på samlet 140 observationer, kan dette ikke være hensigtsmæssigt at indsætte i tabel 1. Derfor undersøges variablerne om hvorvidt de er normalfordelte eller ej, hvor jeg herefter har beregnet middelværdien og spredningen (hvis normalfordelte) eller medianen og IQR (hvis ikke normalfordelte) for variablerne.

I forhold til variablen pooltime hvor de to subgrupper er taget i betragtning, kan jeg ud fra vores histogram og qq-plot konkludere at variablen er normalfordelt:



Derfor bestemmes middelværdien og spredningen for denne variabel.

I forhold til variabelen age hvor de to subgrupper er taget i betragtning, kan jeg ud fra vores histogram og qq-plot konkludere at variabelen ikke er normalfordelt. Det ses også at QQ-plottet afviger fra den rette linje, og derfor kan man konkludere at variabelen ikke er normalfordelt.



Derfor bestemmes medianen og IQR for denne variabel.

2 Sammenhæng mellem crawlsvømning og alder

2.1) Her kan man benytte sig af den logistiske regression, da:

- Alderen er en kontinuert variabel, som i dette tilfælde er den uafhængige variabel
- Crawlsvømningen er en kategorisk variabel, som i dette tilfælde er den afhængige variabel.

Herved ses hvordan crawlsvømning afhænger af alder:

$$\log Odds(crawl) = -2.5264 + 0.1575 * age$$

2.2) For at undersøge den estimerende effekt for alder i modellen vil vi benytte os af oddsratioen e^b .

Oddsratioen fortæller os om den effekt der forekommer af den uafhængige variabel, som er alderen i dette tilfælde.

$$e^b = e^{0.1575} = 1.170534$$

Jeg bestemmer 95 % konfidensintervallet for oddsratioen til værende: [1.0298; 1.3375].

Den tilhørende p-værdi: 0.01765

Den tilhørende nulhypotese lyder på at effekten for alder er lig 0, som betyder at der ikke forekommer en forskel i alder, ensbetydende med at crawlsvømning ikke afhænger af alder.

Jeg har bestemt en effekt til 1.17, som befinder sig indenfor det bestemte 95%-konfidensinterval, som fortæller os at vi er 95 % sikre på at effekten ligger indenfor et interval mellem 1.0298 til 1.3375.

P-værdien er angivet til 0.01765, som er under 5 %, som gør at jeg kan tillade mig at forkaste nulhypotesen, som lyder på at crawlsvømning ikke afhænger af alder. Sagt med andre ord at crawlsvømning afhænger af alderen. Dette er muligt, da jeg ifølge min p-værdi har nok statistisk evidens til at forkaste nulhypotesen.

2.3) Den estimerede sammenhæng mellem crawlsvømning og alder på odds-skalaen er givet ved:

$$Odds(y) = e^{a+b*x}$$

⇕

$$Odds(crawl) = e^{-2.5264 + 0.1575*age}$$

2.4) Den estimerede sammenhæng mellem crawlsvømning og alder på sandsynlighedsskalaen er givet ved:

$$p(y = 1) = \frac{e^{a+b*x}}{1 + e^{a+b*x}}$$

⇕

$$p(crawl = 1) = \frac{e^{-2.5264 + 0.1575*age}}{1 + e^{-2.5264 + 0.1575*age}}$$

2.5) Sandsynligheden for crawlsvømning ved en alder på 9 år præsikter:

$$p(crawl = 1) = \frac{e^{-2.5264 + 0.1575*9}}{1 + e^{-2.5264 + 0.1575*9}} = 0.248076$$

2.6) Risiko kan udtrykkes som en kombination af en uønsket hændelse og sandsynligheden hertil. Derfor er det muligt at benytte sig af sandsynlighedsskalaen, hvor jeg bestemmer sandsynligheden for de 2 aldre, og herefter dividere dem med hinanden.

$$p(\text{crawl} = 1) = \frac{e^{-2.5264 + 0.1575 \cdot 14}}{1 + e^{-2.5264 + 0.1575 \cdot 14}} = 0.4203346$$

$$p(\text{crawl} = 1) = \frac{e^{-2.5264 + 0.1575 \cdot 12}}{1 + e^{-2.5264 + 0.1575 \cdot 12}} = 0.3460608$$

$$\text{Relativ risiko} = \frac{0.4203346}{0.3460608} = 1.214626$$

Dette kan fortolkes som at risikoen for crawlsvømning for en på 14 år er 21% højere end en på 12 år.

2.7) Oddsene bestemmes for to aldre som adskiller sig med 3 år:

$$\text{Odds}(\text{crawl}) = e^{-2.5264 + 0.1575 \cdot 3}$$

$$\text{Odds}(\text{crawl}) = e^{-2.5264 + 0.1575 \cdot 6}$$

Hvor disse to værdier bliver divideret med hinanden:

$$\text{Odds} = \frac{e^{-2.5264 + 0.1575 \cdot 6}}{e^{-2.5264 + 0.1575 \cdot 3}} = 1.603999$$

Dette betyder at oddsene for crawlsvømningen ændrer sig 1.6 gange når alderen stiger med 3 år.

2.8) Her benyttes odds-skalaen, hvor man skal isolere og løse for alderen:

$$0.25 = \frac{e^{-2.5264 + 0.1575 \cdot \text{age}}}{1 + e^{-2.5264 + 0.1575 \cdot \text{age}}} = 9.065$$

Da jeg ikke ved hvordan RStudio gør dette for en, vil jeg i øjeblikket benytte mig af Maple og få følgende værdi for alderen: 9 år. Derved kan man konkludere at sandsynligheden man crawlsvømmer når man er 9 år, er lig 25 %.

2.9) (Intercept) er skæringen med y-aksen, som beskriver odds når x er lig 0.

Sættes dette i relation til vores data, vil (Intercept) beskrive den odds for crawlsvømning der forekommer når alderen er lig 0. P-værdien for (Intercept) er givet ved 0.00418, som er under de 5 %, som gør at man kan forkaste nulhypotesen for at oddsene for crawlsvømning når alderen er lig 0, vil være 0.

3 Hvad påvirker tiden brugt i svømmebassin?

3.1) Her kan man benytte sig af den uparret t-test, da:

- Pooltime er en kontinuert variabel, som i dette tilfælde er den afhængige variabel
- Crawlsvømningen er en kategorisk variabel, som i dette tilfælde er den uafhængige variabel.

Jeg benytter mig af en uparret t-test, da personerne ikke forekommer i begge grupper, men kun en af de to grupper.

3.2) Jeg udfører en uparret t-test, som giver to middelværdier. En middelværdi for tiden brugt i svømmebassin målt i timer/uge for de personer som crawlsvømmere, og en middelværdi for de personer som udfører andre svømmeformer.

Middelværdi for crawlsvømmere (crawl = 1): 9.557736 timer/uge

Middelværdi for andre svømmeformer (crawl = 0): 10.533678 timer/uge

Den gennemsnitlige forskel i tid brugt i svømmebassin pr uge for de to grupper: 0.975942.

95 %-konfidensintervallet fås også fra outputtet fra vores t-test til værende: CI = [0.6132907; 1.3385939].

Den tilhørende p-værdi: 4.784e-07

Den tilhørende nulhypotese lyder på at der ikke forekommer en forskel i tid brugt i svømmebassin pr uge for crawlsvømmer og for personer som udøver andre svømmeformer. Altså at forskellen er lig 0.

Jeg har bestemt en forskel til 0.97, som befinder sig indenfor det bestemte 95%-konfidensinterval, som fortæller os at man er 95 % sikre på at effekten ligger indenfor et interval mellem 0.6133 til 1.3386. P-værdien er angivet til 4.784e-07, som er under 5 %, som gør at jeg kan tillade mig at forkaste nulhypotesen, som lyder på at der ikke forekommer en forskel i tiden brugt i svømmebassin pr uge for crawlsvømmere og for personer som udøver andre svømmeformer. Dette er muligt, da jeg ifølge min p-værdi har nok statistisk evidens til at forkaste nulhypotesen.

3.3) Da et 95 % referenceinterval for tid brugt i svømmebassin pr uge blandt crawlsvømmere ønskes bestemt, vil jeg benytte mig af subgruppe, crawl = 1, og dens tilhørende bestemte middelværdi og spredning som kan findes i tabel 1.

$$Referenceinterval = [middelværdi - 2 * spredning; middelværdi + 2 * spredning]$$

$$Referenceinterval = [9.56 + 2 * 1.01; 9.56 - 2 * 1.01] = [7.54; 11.58]$$

Derved fortæller 95 % referenceintervallet, at det forventes at 95 % populationen blandt crawlsvømmere vil mellem 7.54 til 11.58 timer/uge befinde sig i et svømmebassin.

3.4) Her kan man benytte sig af den lineære regression, da:

- Pooltime er en kontinuert variabel, som i dette tilfælde er den afhængige variabel
- Alder er en kontinuert variabel, som i dette tilfælde er den uafhængige variabel.

3.5) Da det er lineær regression, opstilles følgende model:

$$y = a + b * x$$

⇕

$$pooltime = a + b * age$$

⇕

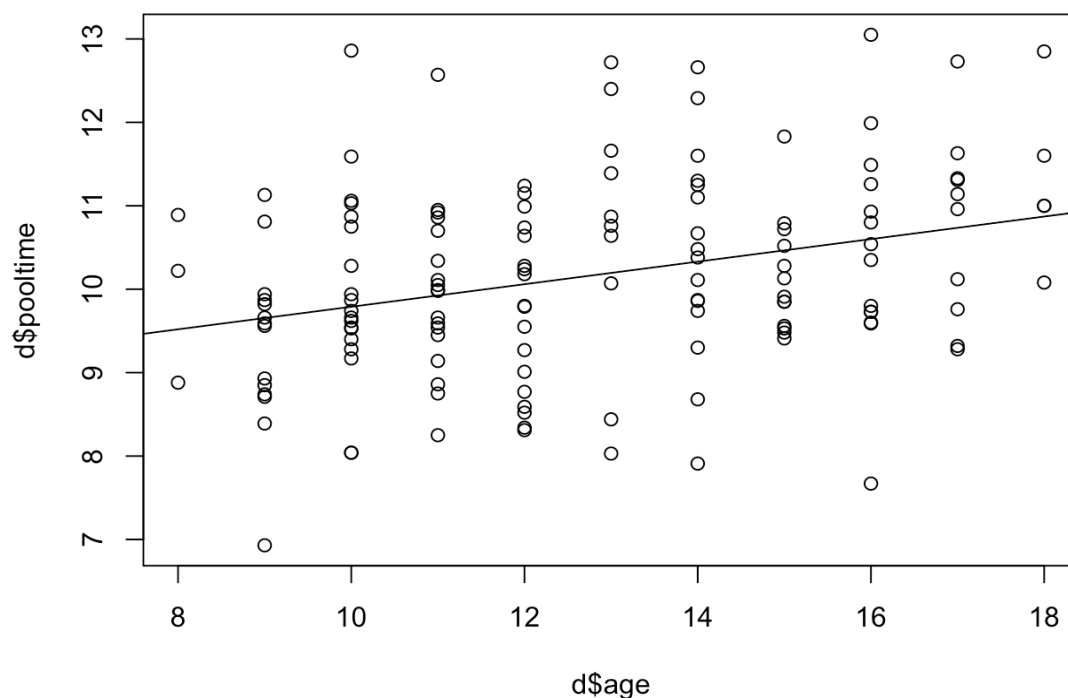
$$pooltime = 8.43640 + 0.13536 * age$$

Parameterestimatet: $a = 8.43640$, er skæringen med y-aksen. Det vil sige når x , altså age er lig 0.

Det vil sige, at ved alderen 0 forventes det at man svømmer 8.44 timer/uge i et svømmebassin.

Parameterestimatet: $b = 0.13536$, er linjens hældning, som beskriver at når x stiger med 1, altså når man bliver 1 år ældre, vil tiden/uge i et svømmebassin stige med 0.13536 timer/uge.

Der indtegnes en tilhørende graf:



3.6) RStudio benyttes til at bestemme 95%-konfidensintervallet via `confint`-kommandoen.

95%-konfidensintervallet bestemmes til: $CI = [0.06653806; 0.2041876]$, som fortæller os at man er 95 % sikre på at den sande effekt af alder på tid brugt i svømmebassin pr uge befinder sig indenfor intervallet 0.0665 til 0.2042 timer/uge. Dette kan også ses på parameterestimatet: $b = 0.13536$ timer/uge, som er indenfor intervallet.

P-værdien for ovenstående parameterestimat, b er til værende: 0.000156, som er under de 5 %, som gør at man kan forkaste nulhypotesen som lyder på at man forventer at $b = 0$, altså at tiden brugt i svømmebassin pr uge vil være lig 0 timer/uge. Med andre ord, betyder dette at tiden brugt i svømmebassin pr uge vil være højere end 0 timer/uge.

3.7) Den lineære regressionmodel benyttes:

$$pooltime = 8.43640 + 0.13536 * 13$$

$$pooltime = 10.19608 \frac{timer}{uge}$$

En 13-årig bruger gennemsnitligt 10.20 timer/uge i et svømmebassin.

95%-Referenceintervallet bestemmes for den 13-årig. Her vil jeg benytte mig af ovenstående værdi, da det er den gennemsnitlige værdi for 13-årige. Yderligere benyttes spredningen fra tabel 1:

$$Referenceinterval = [middelværdi - 2 * spredning; middelværdi + 2 * spredning]$$

$$Referenceinterval = [10.19608 + 2 * 1.01; 10.19608 - 2 * 1.01] = [8.17608; 12.21608]$$

95%-Referenceintervallet bestemmes til: $[8.18; 12.22]$, som kan fortolkes som at 95 % af populationen af 13-årige forventes at bruge mellem 8.18-12.22 timer/uge i et svømmebassin.

3.8) Da jeg skal undersøge hvordan tid brugt i svømmebassin samtidigt afhænger af både alder og crawlsvømning, og herefter undersøge om crawlsvømning påvirker sammenhængen mellem alder og pooltime, vil jeg benytte mig af multiple lineær regression. Dette er grundet at tid brugt i svømmebassin (pooltime) og alder (age) er numeriske værdier, hvor der er blevet tilføjet leddet: crawlsvømning som er en kontinuert værdi. Disse parametre opfylder kravene for en multiple lineær regression

- Pooltime er den afhængige numeriske variabel
- Age er den uafhængige numeriske variabel
- Crawl er den kategoriske variabel

I den multiple lineær regressionsmodel vil jeg give den kategoriske variabel en indikatorvariabel.
Nu opskrives den multiple lineær regressionsmodel med hovedeffekt dog uden interaktionsleddet:

$$pooltime = a + b_1 * age + b_2 + e_i \quad \text{hvis crawl} = 1$$

$$pooltime = a + b_1 * age + e_i \quad \text{hvis crawl} = 0$$

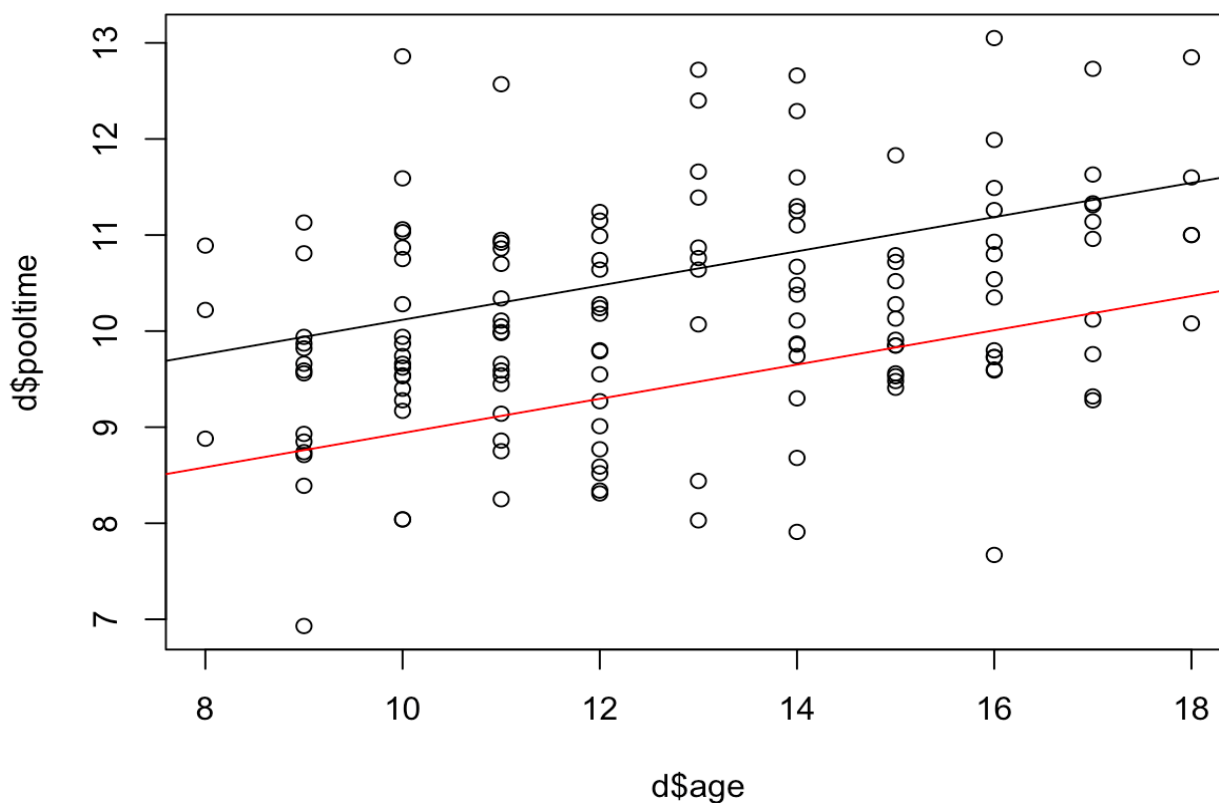
Dette giver følgende via RStudio:

$$pooltime = 8.33560 + 0.17822 * age - 1.17882 \quad \text{hvis crawl} = 1$$

$$pooltime = 7.15678 + 0.17822 * age$$

$$pooltime = 8.33560 + 0.17822 * age \quad \text{hvis crawl} = 0$$

Dette indtegnes:



Nu opskrives den multiple lineær regressionsmodel med hovedeffekt dog uden interaktionsleddet:

$$pooltime = a + b_1 * age + b_2 * 1(crawl = 1) + b_3 * 1(crawl = 1) * age + e_i$$

pooltime =

$$a + b_1 * age + b_2 + b_3 * age + e_i \quad \text{hvis crawl} = 1$$

$$b + a1 * age + e_i$$

$$hvis\ crawl = 0$$

pooltime =

$$(a + b2) + (b1 + b3) * age + e_i$$

$$hvis\ crawl = 1$$

$$a + b1 * age + e_i$$

$$hvis\ crawl = 0$$

Jeg indsætter værdierne fra RStudio-outputtet:

pooltime =

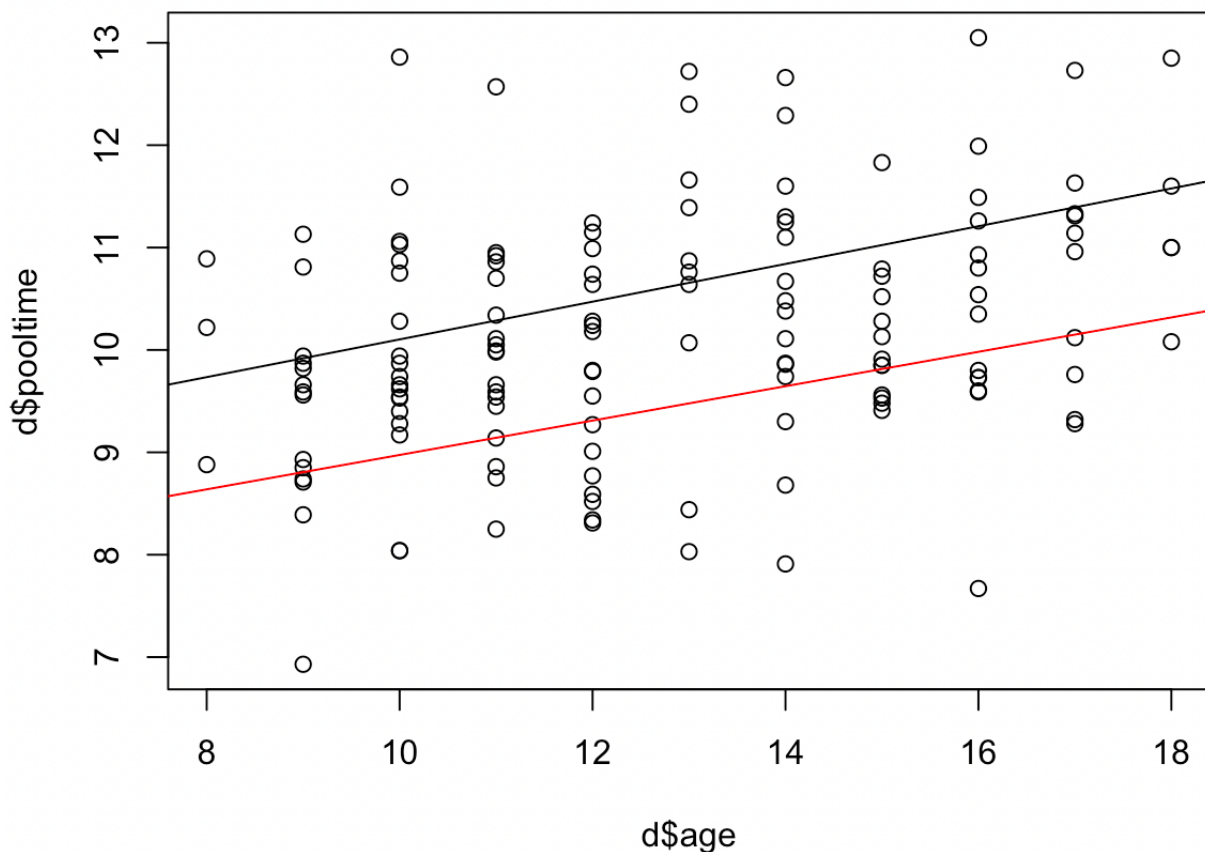
$$(8.25909 - 0.96452) + (0.18443 - 0.01643) * age$$

$$hvis\ condition = 1$$

$$8.25909 + 0.18443 * age$$

$$hvis\ condition = 0$$

Dette indtegnes:



Nu undersøges p-værdier for de to multiple lineære regressionsmodeller:

Med hovedeffekt: 2.43e-10

Med hovedeffekt + interaktionsled: 0.797

Grunden til at jeg ønsker at undersøge p-værdien for de to modeller er grundet p-værdien fortæller mig om hvorvidt jeg kan forkaste vores nulhypotese eller ej. Nulhypotesen lyder på at jeg ønsker at

undersøge om interaktionsleddet er lig 0. Da jeg har fået en p-værdi på 7.97 %, altså over 5 %, har jeg ikke nok statistisk evidens til at forkaste nulhypotesen. Dette betyder at interaktionsleddet godt kan antage denne værdi, altså værdien 0. Hvis interaktionsleddet antager værdien 0, vil dette medføre til at de to linjer bliver parallelle, og dermed kan man udføre modelselektion, som gør at man kan benytte sig af den anden multiple lineær regression med blot hovedeffekt uden interaktionsled.

3.9) Herved opskrives udtrykket for min endelige model op:

$$pooltime = 8.33560 + 0.17822 * age - 1.17882 \quad \text{hvis crawl} = 1$$

$$pooltime = 7.15678 + 0.17822 * age$$

$$pooltime = 8.33560 + 0.17822 * age \quad \text{hvis crawl} = 0$$

Parameterestimatet: $a = 8.33560$ er skæringen med y-aksen, altså når x er lig 0.

Parameterestimatet: $b_1 = 0.17822$ er linjens hældning når x stiger med 1.

Parameterestimatet: $b_2 = -1.17882$ er den ekstra effekt mellem de to grupper.

3.10) Et 95 % konfidensinterval for effekten af crawlsvømning i den endelige model bestemmes via RStudio til værende: $CI = [-1.5197080; -0.8379404]$, som fortæller at man er 95 % sikker på at den sande effekt af crawlsvømning ligger indenfor dette interval. Dette kan også ses på vores model, hvor effekten er lig -1.17882 , som er indenfor 95%-konfidensintervallet.

3.11) P-værdierne i outputtet fra den endelige model indsættes:

P-værdi for parameterestimatet: a ((Intercept)) = $< 2e-16$

P-værdi for parameterestimatet: b_1 (d\$age) = $4.71e-08$

P-værdi for parameterestimatet: b_2 (d\$crawl) = $2.43e-10$

Alle disse p-værdier har en værdi på under 5 %, som gør at jeg kan forkaste deres nulhypotese, om at de er alle sammen lig 0. Nulhypotesen for b_2 (d\$crawl) er især vigtig, da $b_2 = 0$, vil gøre at man kan modelselektere og vælge den lineær regression fremfor den multiple lineærregression. Det skal huskes at b_2 er den ekstra effekt på y-aksen, og hvis denne er lig 0, er der ingen forskel på de to ”parallelle linjer”, de vil nemlig skære samme sted på y-aksen og have samme hældning. Denne

nulhypotese forkastes grundet p-værdien på under 5 %. Sagt med andre ord, at den multiple lineær regressionsmodel benyttes i dette tilfælde.

Man kan yderligere kommentere på nulhypotesen om b_1 ($d\$age$), som blot vil medføre til at linjen ville skære i 0, ensbetydende med at en person på 0 år, vil have haft 0 timer/uge i svømmebassinet. Denne nulhypotese forkastes grundet p-værdien på under 5 %.

Til sidst kan man også kommentere på nulhypotesen om a ((Intercept)), som vil medføre til at linjen har en hældning på 0, som er en konstant horisontal linje. Dette ville betyde at når en person bliver ældre med 1 år, vil antallet af timer/uge i svømmebassinet være konstant. Ensbetydende med at personen slet ikke benytter sig af svømmebassinet gennem ugen. Denne nulhypotese forkastes grundet p-værdien på under 5 %.

3.12) Modellen benyttes til at prædiktere den gennemsnitlige tid i svømmebassin for en person på 12 år, der svømmer crawl:

$$pooltime = 8.33560 + 0.17822 * age - 1.17882 \quad \text{hvis crawl} = 1$$

$$pooltime = 7.15678 + 0.17822 * age$$

⇕

$$pooltime = 7.15678 + 0.17822 * 12$$

$$pooltime = 9.29542$$

En person på 12 år, der svømmer crawl, bruger 9.30 timer/uge i et svømmebassin

3.13) Modellen benyttes til at prædiktere den gennemsnitlige tid i svømmebassin for en person på 12 år, der ikke svømmer crawl:

$$pooltime = 8.33560 + 0.17822 * age \quad \text{hvis crawl} = 0$$

⇕

$$pooltime = 8.33560 + 0.17822 * 12$$

$$pooltime = 10.47424$$

En person på 12 år, der ikke svømmer crawl, bruger 10.47 timer/uge i et svømmebassin

3.14) Jeg benytter mig af modellen, hvor jeg blot indsætter alder som to forskellige værdier, som er forskellig fra hinanden med værdien 4:

$$pooltime = 8.33560 + 0.17822 * 12 = 10.47424 \quad crawl = 0$$

$$pooltime = 8.33560 + 0.17822 * 16 = 11.18712 \quad crawl = 0$$

Ændring i tid i svømmebassin over 4 år for en person der ikke svømmer crawl er lig: 0.71288

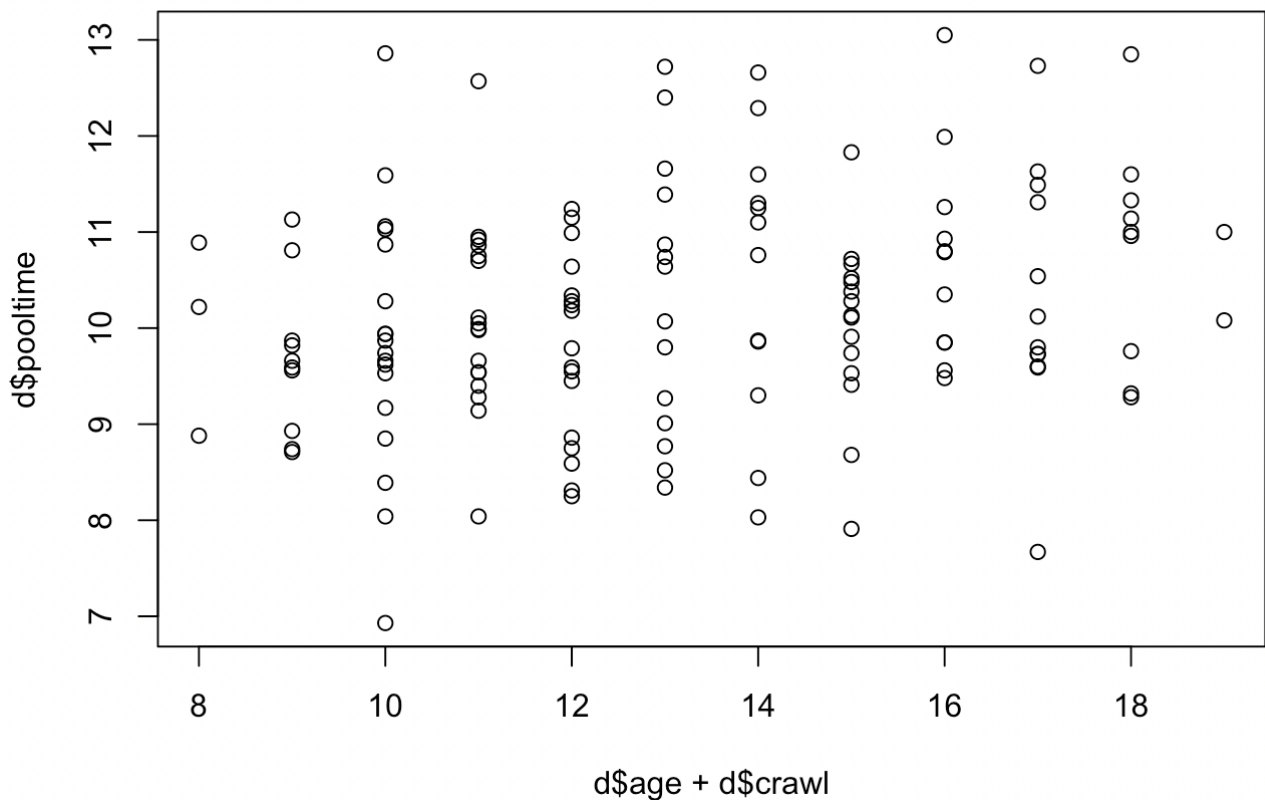
$$pooltime = 7.15678 + 0.17822 * 12 = 9.29542 \quad crawl = 1$$

$$pooltime = 7.15678 + 0.17822 * 16 = 10.0083 \quad crawl = 1$$

Ændring i tid i svømmebassin over 4 år for en person der svømmer crawl er lig: 0.71288

Den gennemsnitlige ændring i svømmebassin hen over 4 år for personer der svømmer crawl og for personer der ikke svømmer crawl, er lig hinanden: $0.71288 = 0.71288$. Dette er grundet at de personer som crawlsvømmere og de personer som ikke-crawlsvømmere regelmæssigt svømmetræner, og måden hvorpå de svømmetræner, altså deres svømmeform har ingen betydning i forhold til deres regelmæssige svømmetræning.

3.15) Nedenfor ses en graf, der illustrer hvordan tid i svømmebassin afhænger af både alder og crawlsvømning:



4 Indtagelse af kulsyreholdige læskedrikke

4.1) Sandsynligheden for at indtage kulsyreholdige læskedrikke mindre end fem gange per uge beregnes. Værdierne anskaffes fra tabel 1, og sandsynligheden bestemmes via prop.test, så vi tager højde for usikkerheder.

Sandsynligheden bestemmes til værende: 0.4214286.

Sandsynligheden for at man indtager kulsyreholdige læskedrikke mindre end fem gange per uge er lig 0.421. Her er 95 %-konfidensintervallet for sandsynligheden beregnet til: CI = [0.3394211; 0.5078135], som fortæller at man er 95 % sikre på at sandsynligheden ligger indenfor dette interval. Dette kan også ses på den bestemte værdi.

4.2) Odds for at indtage kulsyreholdige læskedrikke mindre end fem gange per uge beregnes. Værdierne anskaffes fra tabel 1, og odds bestemmes via prop.test, så vi tager højde for usikkerheder. Oddsene bestemmes til værende: 0.7283951.

Oddsene kan tolkes som at for hver 0.728 person som indtager kulsyreholdige læskedrikke mindre end fem gange per uge, har man en person som indtager et andet antal (enten intet eller mere).

4.3) Da de specifikt spørger ind til risikodifferencen og den relative risiko for crawlsvømmere ved to forskellige estimater, vil vi benytte vores subgruppe, crawl = 1, til at undersøge dette.

Risikodifferencen er lig forskellen af de to risici for de to estimater. Først bestemmes risikoen for hver af de to estimater for variablen soda. Dette gøres via prop.test, så vi tager højde for usikkerheder:
Gruppen med indtagelse af kulsyreholdige læskedrikke mindre end fem gange per uge: 0.4150943
Gruppen med indtagelse af kulsyreholdige læskedrikke mere end eller lig 5 gange per uge: 0.3018868

Risikodifferencen bestemmes:

$$\text{Risikodifference} = 0.4150943 - 0.3018868 = 0.1132075$$

Ud fra ovenstående kan vi konkludere at risikoen for at indtage kulsyreholdige læskedrikke mindre end fem gange per uge er lig 11 procentpoint.

Den relativ risiko (risikoratio) bestemmes:

$$\text{Relativ risiko} = \frac{0.4150943}{0.3018868} = 1.375$$

Dette resultat kan fortolkes som at risikoen for at crawlsvømmere indtager kulsyreholdige læskedrikke mindre end fem gange per uge, er 37.5% større end at indtage mere end eller lig 5 per uge.

4.4) Da de specifikt spørger ind til oddsratioen for crawlsvømmere ved to forskellige estimater, vil jeg benytte vores subgruppe til at undersøge dette. Jeg opstiller en matrix hvor jeg herefter benytter mig af fisher.test for at tage højde for usikkerheder.

Oddsratioen bestemmes til værende: 1.633378. Dette kan tolkes som at oddsratioen for at crawlsvømmere som indtager mindre end 5 kulsyreholdige læskedrikke per uge, er 63 % højere end de crawlsvømmere som indtager mere end eller lig 5 kulsyreholdige læskedrikke per uge. Hvis oddsratioen var lig 1 ville dette betyde at risikoen for at indtage mindre end eller mere end 5 kulsyreholdige læskedrikke per uge, være den samme.

5 Hvad påvirker risikoen for ætsninger på tænderne?

5.1) Her kan man benytte sig af chi-i-anden testen, da:

- Erosion er en kategorisk variabel, som i dette tilfælde er den afhængige variabel
- Crawlsvømningen er en kategorisk variabel, som i dette tilfælde er den uafhængige variabel.

Specifikt benyttes chi-i-anden testen, da testen sammenligner to grupper, som begge antager kategoriske variabler. Værdier for erosion og crawl fås fra tabel 1.

5.2) Chi-i-anden testen giver os en p-værdi på 0.0001, som er under 5 %, som tillader at man kan forkaste nulhypotesen, da man har nok statistisk evidens til at gøre dette. Nulhypotesen lyder på at der er ingen forskel på risikoen blandt crawlsvømmere og ikke-crawlsvømmere, altså at forskellen er lig 0. Dog grundet p-værdien på under 5 %, kan man forkaste nulhypotese, og derfor tolke det som at der er en forskel i risiko for ætsninger på tænderne blandt crawlsvømmere og ikke-crawlsvømmere.

5.3) Værdier på erosion i hver af de to grupper anskaffes fra tabel 1. Risikodifferensen bestemmes som forskellen i risici i de to grupper. Derfor bestemmes først risici for de to grupper via prop.test, som tager højde for usikkerheder:

Risiko for erosion blandt crawlsvømmere: 0.3396226

Risiko for erosion blandt ikke-crawlsvømmere: 0.07407407

Risikodifferencen bestemmes: $0.3396226 - 0.07407407 = 0.2655485$

Risikoen for erosion blandt crawlsvømmere er 26.5 procentpoint højere end hos ikke-crawlsvømmere. Hvis risikoen var den samme, ville risikodifferencen være lig 0 procentpoint.

5.4) Da risikoen for de to grupper er beregnet i ovenstående opgave, skal vi blot tage forholdet mellem de to grupper risici, for at bestemme den relative risiko:

$$Relative\ risiko = \frac{0.07407407}{0.3396226} = 0.218107$$

Den relative risiko for at ikke-crawlsvømmere får ætsninger på tænderne er 0.22 større end crawlsvømmere.

5.5) Oddsrationen er forholdet af odds mellem de to grupper. Jeg opstiller en matrix, hvor vi inddrager værdierne for erosion-status for begge grupper. Jeg kan herefter via fisher.test, bestemme en oddsratio, som også tager højde for usikkerhederne.

Dette giver en oddsratio på: $6.836359 = 6.84$. Ud fra oddsrationen kan jeg tolke det som at der er 6.84 gange større odds for at man får ætsninger på tænder hvis man er crawlsvømmere.

5.6) Her kan man benytte sig af den logistiske regression, da:

- Pooltime er en kontinuerlig variabel, som i dette tilfælde er den uafhængige variabel
- Erosion er en kategorisk variabel, som i dette tilfælde er den afhængige variabel.

$$\log Odds(erosion) = 3.4608 - 0.5062 * pooltime$$

5.7) Den estimerede model skrives ind i odds-skalaen:

$$Odds(y) = e^{a+b*x}$$

⇕

$$Odds(erosion) = e^{3.4608 - 0.5062 * pooltime}$$

Modellen fortæller om den relative ændring i odds for erosion afhængigt af tid brugt i svømmebassin. Effekten, altså oddsrationen bestemmes. Oddsrationen fortæller os om den effekt der forekommer af den uafhængige variabel, som er pooltime i dette tilfælde.

Nu beregnes effekten:

$$e^b = e^{-0.5062} = 0.6027818$$

Der fås følgende 95 %-konfidensintervallet: CI: [0.389579; 0.8991999], som fortæller os at man er 95 % sikker på at effekten af tid brugt i svømmebassinet ligger indenfor intervallet: 0.39 til 0.90. Dette kan jeg blandt andet se ved effekten, som har en værdi på 0.603 som er indenfor intervallet.

Den tilhørende p-værdi er 0.0168 og nulhypotesen lyder på at effekten af tiden brugt i svømmebassinet er lig 0, ensbetydende med at tiden brugt i svømmebassinet ingen effekt har på ætsning af tænder. Da jeg har fået en p-værdi på 0.0168, som er under de 5 %, kan man forkaste nulhypotesen, og derfor konkludere at effekten af tiden brugt i svømmebassinet ikke er lig 0, og derfor har en effekt på ætsning af tænder.

5.8) Analysen giver følgende estimerede model:

$$\text{Odds(erosion)} = e^{1.6198 - 0.2400 * \text{pooltime}}$$

Nu beregnes effekten:

$$e^b = e^{-0.2400} = 0.7866279$$

Der fås følgende 95 %-konfidensintervallet for effekten: CI: [0.4338757; 1.391811], som fortæller os at man er 95 % sikker på at effekten af tid brugt i svømmebassinet ligger indenfor intervallet: 0.43 til 1.39. Dette kan jeg blandt andet se ved effekten, som har en værdi på 0.79 som er indenfor intervallet.

Den tilhørende p-værdi er 0.412 og nulhypotesen lyder på at effekten af tiden brugt i svømmebassinet af crawlsvømmere er lig 0, ensbetydende med at tiden brugt i svømmebassinet ingen effekt har på ætsning af tænder blandt crawlsvømmere. Da jeg har fået en p-værdi på 0.412, som er over de 5 %, kan man ikke forkaste nulhypotesen, da der ikke er nok statistisk evidens for dette. Dette betyder at effekten af tiden brugt i svømmebassinet blandt crawlsvømmere godt kan antage værdien 0, og derved ingen effekt have på ætsning af tænder.

5.9) Analysen giver følgende estimerede model:

$$\text{Odds(erosion)} = e^{-0.7071 - 0.1816 * \text{pooltime}}$$

Nu beregnes effekten:

$$e^b = e^{-0.1816} = 0.8339348$$

Der fås følgende 95 %-konfidensintervallet for effekten: CI: [0.3671132; 1.760198], som fortæller os at man er 95 % sikker på at effekten af tid brugt i svømmebassinet ligger indenfor intervallet: 0.37 til 1.76. Dette kan jeg blandt andet se ved effekten, som har en værdi på 0.83 som er indenfor intervallet.

Den tilhørende p-værdi er 0.644 og nulhypotesen lyder på at effekten af tiden brugt i svømmebassinet blandt crawlsvømmere er lig 0, ensbetydende med at tiden brugt i svømmebassinet ingen effekt har på ætsning af tænder blandt ikke-crawlsvømmere. Da jeg har fået en p-værdi på 0.644, som er over de 5 %, kan man ikke forkaste nulhypotesen, da der ikke er nok statistisk evidens for dette. Dette betyder at effekten af tiden brugt i svømmebassinet blandt ikke-crawlsvømmere godt kan antage værdien 0, og derved ingen effekt have på ætsning af tænder.

5.10) Konklusion fra 5.7) fortæller at effekten af tiden brugt i svømmebassinet ikke er lig 0, og derfor har en effekt på ætsning af tænder. Denne konklusion er baseret på hele datasættet, som indeholder værdier for erosion for begge grupper (crawl = 1 og crawl = 0). Konklusionerne fra 5.8) og 5.9) fortæller at effekten af tiden brugt i svømmebassinet godt kan antage værdien 0, og derfor kan man konkludere at effekten af tiden brugt i svømmebassinet ingen påvirkning har på ætsningen af tænder. Disse konklusioner er baseret på subgrupper af datasættet, det vil sige analyse på KUN den enkelte subgruppe (enten crawl = 1 eller crawl = 0).

Konklusionerne fra disse 3 spørgsmål stemmer ikke overens. Dette kan være grundet at når man i 5.7) kigger på den samlede effekt af hvordan tiden i svømmebassinet kan medføre til en påvirkning på ætsning af tænder, består datasættet af de to grupper, hvor man derfor indirekte har taget højde for at man undersøger en effekt mellem de to subgrupper i datasættet.

Med hensyn til 5.8) og 5.9) udføres analysen KUN på den pågældende subgruppe (enten crawl = 1 eller crawl = 0), og derfor kan den pågældende subgruppe ikke sammenlignes med noget andet, altså med den anden subgruppe. Da der kun er en subgruppe at udføre analyse på, vil der ikke være nok statistisk evidens for at konkludere at tiden brugt i svømmehallen har en påvirkning på ætsning af tænder, da man netop ikke kan sammenligne det med f.eks. den anden subgruppe.

Man skal også huske at tage højde for at der udføres en analyse på: om tid i svømmebassinet påvirker ætsning af tænder. Dette er ikke den primære faktor for ætsning af tænder, da den primære faktor af ætsning af tænder skyldes svømmeformen, som tydeligt ses i andre forrige opgaver og analyser, hvor man har taget udgangspunkt i variabelen crawl.

Appendix

1.1 Datapræsentation

```
d <- read.csv("http://causal.sund.ku.dk/f22/55.csv", header=TRUE, stringsAsFactors=TRUE)
```

```
crawl <- subset(d, crawl == 1)
```

```
others <- subset(d, crawl == 0)
```

```
table(crawl$erosion)
```

18/53 * 100

35/53 * 100

```
table(others$erosion)
```

6/87 * 100

81/87 * 100

```
table(crawl$soda)
```

8 + 16 + 22

53 - 46

8/53 * 100

22/53 * 100

16/53 * 100

7/53 * 100

```
table(others$soda)
```

37 + 37 + 6

87 - 80

37/87 * 100

6/87 * 100

7/87 * 100

```
table(d$pooltime)
```

```
hist(crawl$pooltime)
```

```
qqnorm(crawl$pooltime)
```

```
qqline(crawl$pooltime)
```

```
mean(crawl$pooltime)
```

```
sd(crawl$pooltime)
```

```
hist(others$pooltime)
qqnorm(others$pooltime)
qqline(others$pooltime)
mean(others$pooltime)
sd(others$pooltime)
```

```
table(d$age)
hist(crawl$age)
qqnorm(crawl$age)
qqline(crawl$age)
median(crawl$age)
IQR(crawl$age)
```

```
hist(others$age)
qqnorm(others$age)
qqline(others$age)
median(others$age)
IQR(others$age)
```

2 Sammenhæng mellem crawlsvømning og alder

2.1 & 2.2

```
m1 <- glm(d$crawl ~ d$age, family = binomial)
m1
summary(m1)
exp(0.15746)
confint(m1)
exp(0.02944233)
exp(0.2907930)
```

2.5

```
exp(-2.5264 + 0.1575*9)/(1 + exp(-2.5264 + 0.1575*9))
```

2.6

```
exp(-2.5264 + 0.1575*14)/(1 + exp(-2.5264 + 0.1575*14))  
exp(-2.5264 + 0.1575*12)/(1 + exp(-2.5264 + 0.1575*12))  
0.4203346/0.3460608
```

2.7

```
exp(-2.5264 + 0.1575*3)  
exp(-2.5264 + 0.1575*6)  
0.2056869/0.1282338
```

2.8

Maple:

```
solve(0.25 = exp(-2.5264 + 0.1575*age)/(1 + exp(-2.5264 + 0.1575*age)), age)
```

RStudio:

```
exp(-2.5264 + 0.1575*9.065)/(1 + exp(-2.5264 + 0.1575*9.065))
```

2.9

```
summary(m1)
```

3 Hvad påvirker tiden brugt i svømmebassin?

3.2

```
t.test(d$pooltime ~ d$crawl)  
10.533678 - 9.557736
```

3.3

```
9.56 + 2*1.01  
9.56 - 2*1.01
```

3.5

```
m2 <- lm(d$pooltime ~ d$age)  
summary(m2)  
plot(d$age, d$pooltime)  
abline(8.43640, 0.13536)
```

3.6

confint(m2)

3.7

$8.43640 + 0.13536 * 13$

$10.19608 + 2 * 1.01$

$10.19608 - 2 * 1.01$

3.8

`m3 <- lm(d$pooltime ~ d$age + d$crawl)`

`summary(m3)`

8.33560-1.17882

`plot(d$age, d$pooltime)`

`abline(8.33560, 0.17822)`

`abline(8.33560 - 1.17882, 0.17822,col="red")`

`m4 <- lm(d$pooltime ~ d$age * d$crawl)`

`summary(m4)`

`plot(d$age, d$pooltime)`

`abline(8.25909, 0.18443 )`

`abline(8.25909-0.96452, 0.18443-0.01643,col="red")`

3.10

confint(m3)

3.12

$7.15678+0.17822*12$

3.13

$8.33560+0.17822*12$

3.14

$7.15678+0.17822*12$

7.15678+0.17822*16

10.0083 - 9.29542

8.33560+0.17822*12

8.33560+0.17822*16

11.18712 - 10.47424

3.15

summary(m3)

plot(d\$age + d\$crawl, d\$pooltime)

4 Indtagelse af kulsyreholdige læskedrikke

4.1

prop.test(59,140)

4.2

prop.test(59, (140-59))

4.3

prop.test(22,53)

prop.test(16,53)

0.4150943 - 0.3018868

0.4150943/0.3018868

4.4)

m2 <- matrix(c(22, 16, 31, 37), nrow = 2, ncol = 2)

fisher.test(m2)

5 Hvad påvirker risikoen for ætsninger på tænderne?

5.1 & 5.2

table(d\$erosion,d\$crawl)

m2 <- matrix(c(6, 81, 18, 35), nrow = 2, ncol = 2)

m2

chisq.test(m2)

5.3

```
prop.test(18,53)
prop.test(6,81)
0.3396226 - 0.07407407
```

5.4

```
0.07407407/0.3396226
```

5.5

```
matrix <- matrix(c(18,6,(53-18),(87-6)), nrow = 2, ncol = 2)
matrix
fisher.test(matrix)
```

5.7

```
m2 <- glm(d$erosion ~ d$pooltime, family = binomial)
summary(m2)
confint(m2)
```

5.8

```
m3 <- glm(crawl$erosion ~ crawl$pooltime, family = binomial)
summary(m3)
exp(-0.2400)
confint(m3)
exp(-0.8349972)
exp(0.3306055)
```

5.9

```
m4 <- glm(others$erosion ~ others$pooltime, family = binomial)
summary(m4)
exp(-0.1816)
confint(m4)
exp(-1.002085)
exp(0.5654263)
```