Eksamen i statistik

Opgave 1

1. Filen 63.csv tilgået via http://causal.sund.ku.dk/f22/63.csv
2. Først har jeg opdelt datasættet i en gruppe med erosioner og en gruppe uden. Vi undersøger en sammenhæng om tid brugt i svømmebassin, og evt. med forskellig svømmestil samt indtag af kulsyreholdig læskedrik, kan medføre ætsninger. Derfor er det i hypotesens interesse at vi sætter svømmere med og uden erosioner op mod hinanden og tester for, om hypotesen holder vand.

   Age. Numerisk variabel. Da stikprøven er så lille, er histogrammet for den numeriske variabel svært at tyde. Ved i tillæg at lave QQ-plots for begge grupper har jeg konkluderet, at de IKKE er normalfordelte nok til at bruge gennemsnit og spredning til at beskrive variablen. Jeg har derfor regnet med medianen og IQR i stedet for.

   Crawl. Det er en binær kategorisk variabel på trods af sit numeriske udtryk. Derfor beskriver jeg den som en normal kategorisk variabel med hyppighed og relativ frekvens.

   Soda. Dette er også en kategorisk variabel. Jeg beskriver den ved hyppigheden og den relative frekvens. I denne variabel møder jeg også mine første manglende data. Der mangler observationer på indtaget af kulsyreholdig læskedrik blandt 4 personer med erosioner og 9 personer uden erosioner. Jeg har valgt at inkludere dem i sin egen række for større gennemsigtbarhed for læseren, og da det er en betragtelig andel af stikprøven.

   Pooltime. Dette er en numerisk variabel. I histogrammer og QQ-plots ses samme tendens som for age, hvor en del af data meget fint følger en normalfordeling, mens en

anden del af data ikke gør, og derfor mener jeg IKKE det er normalfordelt data. For både pooltime og age følger QQ-plottet et mere symmetrisk fladt fordelt histogram. Jeg har i stedet brugt median og IQR til at beskrive data for pooltime i tabellen.

Jeg har struktureret tabellen så de informationer, som hypotesen lægger størst vægt på, er at finde øverst. Antal deltagere og alder er selvfølgelig først, da dette er hele grundlaget for undersøgelsen. Dernæst kommer pooltime, herunder svømmestil, og sidst indtag af kulsyreholdig læskedrik, som er en mere isoleret variabel end de andre. Dermed fremlægger de centrale elementer i hypotesen først efterfulgt at de mere perifære.

	Erosion	**No erosion**
No. of participants	30	101
Age, median (IQR)	13 (4)	13 (5)
Pooltime, median (IQR)	9,56 (1,315)	10,49 (1,77)
Crawl, N(%)		
Yes	21 (70,0)	18 (17,8)
No	9 (30,0)	83 (82,2)
Soda, N (%)		
<5/week	13 (43,3)	38 (37,6)
>=5/week	12 (40,0)	45 (44,6)
Never	1 (3,3)	9 (8,9)
NA	4 (13,3)	9 (8,9)

Figur 1 - Tabel 1 for datasæt 63

Opgave 2

1. Logistisk regression. Den afhængige variabel (crawl eller anden svømmestil) er binær, mens den uafhængige variabel (alder) er kontinuert. Derfor er det netop logistisk regression vi kan benytte til at besvare dette videnskabelige spørgsmål. I RStudio bruges glm-funktionen til at undersøge denne sammenhæng.
2. Fra modellen finder vi, at effekten af alder (odds-ratio) er $e^{0,166} = 1,18$.
   95%-konfidensinterval er [1,03; 1,36].
   P-værdi for alderens effekt på sandsynligheden for crawl er 0,02.
   Nulhypotesen er, at alder IKKE har en effekt på sandsynligheden for crawlsvømning. Det vil betyde, at regressionskoefficienten er 0, dvs. at odds-ratio er lig 1. Med en p-værdi=0,02 kan vi forkaste denne nulhypotese, så alder altså har en effekt på sandsynlighed for crawlsvømning.
   Vi kan konkludere, at alder har en statistisk signifikant effekt (p=0,02 og odds-ratio=1,18) på sandsynligheden for crawlsvømning med et 95%-konfidensinterval på 1,03 til 1,36. Effekten af alder er helt præcis, at for hver gang alder øges med et år, vil odds for crawlsvømning stige med 18%.

3. Odds-skala:
$$Odds(crawl) = e^{-3,044+0,166 \cdot age}$$

4. Sandsynlighedsskala:
$$P(crawl = 1) = \frac{e^{-3,044+0,166 \cdot age}}{1 + e^{-3,044+0,166 \cdot age}}$$

5. Jeg indsætter 9 som age i sandsynlighedsskalaen:
$$P(crawl = 1) = \frac{e^{-3,044+0,166 \cdot 9}}{1 + e^{-3,044+0,166 \cdot 9}} = 17,51\%$$
   Sandsynligheden for crawlsvømning for en person på 9 år er 17,5%.

6. Jeg beregner sandsynligheden for crawlsvømning hos en person på 14 år og dividerer med sandsynligheden for crawlsvømning hos en person på 12 år. Der er 26,5% større risiko for at en 14-årig person svømmer crawl end en 12-årig person.

7. Odds-ratio, eller effekt af alder, er 1,18(se opg. 2.2). Denne faktor er multiplikativ, og dermed estimeres odds til at ændres fra 1,18 til $1{,}18^3 = 1{,}64$. Altså estimerer vi at odds for crawlsvømning stiger 64%, når alderen stiger med 3 år.

8. Når sandsynligheden for crawlsvømning krydser 25% er odds 0,33 jf. arket "Mellemregninger" fra SAU7 (når sandsynlighed for succes skal give 0,25 må odds være 0,33 for at det går op i ligningen nederst i dokumentet).
   Dermed er $logOdds = \log\left(\frac{1}{3}\right) = -0{,}48$.
   Ved at se på Odds-skalaen (se opg. 2.3) og tage log på begge sider, kan vi nu opstille ligningen:

$$-0{,}48 = -3{,}044 + 0{,}166 \cdot age$$

$$\Updownarrow$$

Ligningen løses for age vha. CAS-værktøjet WordMat.

$$age = 15{,}5$$

   Altså vil sandsynligheden for crawl først krydse 25% ved alderen 15,5 år.

9. Ud for intercept ser vi vores a-værdi i logOdds-modellen. Den fortæller os om linjens skæring med y-aksen, dvs. y-værdien når x=0. Her er den -3,044. Det er altså logaritmen til odds for crawlsvømning, når alderen er 0 år. Den kan udregnes til at være 4,7%. Dette viser en af faldgruberne ved at prædiktere data for ekstrapolerede x-værdier. Ingen nyfødte svømmer crawl.
   RStudio tester altid nulhypotesen der lyder, at hver af regressionsparametrene er lig 0. I dette tilfælde er p < 0,001 for a, så derfor kan vi **forkaste** nulhypotesen. Ved ligeledes at se på p-værdien for parameteren b(p=0,02), kan vi nu komme med en konklusion for modellen.

Vi kan konkludere, at der er evidens for, at alder har en statistisk signifikant effekt på sandsynligheden for crawlsvømning, og at denne har en anden værdi end 1*, når x=0.*

Opgave 3

1. Uparret t-test. Vores afhængige variabel er kontinuert numerisk mens vores uafhængige variabel er kategorisk binær. Havde der også været et kontinuert led i vores uafhængige variabel, ville jeg i stedet bruge multipel lineær regression. Det er der ikke, og derfor bruger jeg T-test til at undersøge sammenhængen. Vi bruger den uparrede test da det er to separate grupper, hvor ingen personer går igen.

2. Ikke-crawlsvømmere bruger i gennemsnit 1,07 timer mere i svømmebassin pr uge end crawlsvømmere.
   95% konfidensintervallet er [0,64; 1,50]
   $p < 0{,}001$ for nulhypotesen der lyder, forskellen i middelværdi mellem de to grupper er nul, og dermed kan vi **forkaste** nulhypotesen.

   En uparret t-test viste en signifikant forskel i middelværdi (p<0,001) for gennemsnitlig tid brugt i svømmebassin pr uge for ikke-crawlsvømmere og crawlsvømmere. Forskellen mellem ikke-crawlsvømmere og crawlsvømmere var i gennemsnit på 1,07 timer/uge med et 95% konfidensinterval på [0,64;1,50], og vi kan på baggrund af dette konkludere, at nulhypotesen forkastes.

3. Først sikrer jeg mig, at data for crawlsvømmere er normalfordelt. På baggrund af et histogram og QQ-plot vil mene, at data er normalfordelt med tanke på den lille stikprøve.

   95%-referenceintervallet udregnes som middelværdien ± 2 gange spredningen. Intervallet beregnes til at være 7,4 timer i ugen til 11,8 timer i ugen. Det betyder, at for

95% af crawlsvømmerne vil tiden bruge i svømmebassinet pr uge ligge mellem 7,4 timer og 11,8 timer.

4. Simpel lineær regression. Den afhængige variabel er kontinuert og den uafhængige variabel er også kontinuert. Hvis den uafhængige variabel også omfattede svømmestil, ville jeg i stedet bruge multipel lineær regression hvor svømmestil ville skabe interaktionsleddet, men det er ikke hvad vi bliver bedt om her.

5. Analysen giver følgende model:

$$pooltime = 7{,}934 + 0{,}187 \cdot age$$

Vores a-værdi er skæring med y-aksen, eller y når $x = 0$. Hvor b-værdi fortæller hvor meget y-værdien stiger med, når x stiger med 1. Vi ser ud fra modellen at der ved alderen 0 år bliver brugt knap 8 timer i svømmebassinet i ugen. Igen viser dette nødvendigheden af at vurdere hvor meget en model kan bruges til at prædiktere data. For hvert år der stiger, vil en person gennemsnitligt bruge knap 0,2 timer mere i svømmebassinet i ugen. Her skal man også overveje hvor langt modellen kan prædiktere data, da svømmetiden pr uge for fx en 90-årig vil være knap 25 timer ifølge modellen.

En vigtig pointe ved denne regressionsmodel er, at der er en stor spredning, eller residual standard error, på 1,15 timer i svømmebasin pr uge. Ved at lave et scatterplot og en grafisk fremstilling af regressionen ser vi tydeligt, at mange datapunkter ligger langt fra linjen. Vi kan også se på den lave r^2-værdi, at den lineære model stemmer dårligt overens med datapunkterne. Sagt med andre ord - linjen er den bedst mulige for vores data, men den er stadig relativt ukorrekt.

6. 95%-konfidensinterval for a-værdien er [7,00; 8,87]
   95%-konfidensinterval for b-værdien er [0,12; 0,26]

Vi er altså 95% sikre på, at den sande værdi af skæringspunktet med y-aksen ligger i intervallet [7,00; 8,87] *og tilsvarende 95% sikre på, at den sande værdi af hældningskoefficienten ligger i intervallet* [0,12; 0,26].

Hvis funktionen er opskrevet som $pooltime = a + b \cdot age$ svarer de to p-værdier i outputtet til nulhypoteserne $H_0: a = 0$ og $H_0: b = 0$
Begge parametre har en p-værdi der er langt under 0,05 (p<0,001 for begge), og dermed kan vi **forkaste** nulhypoteserne.

Sammenhængen mellem tid brugt i svømmebassinet og alder blev analyseret ved simpel lineær regression. Sammenhængen var signifikant (p < 0,001), og for hver ændring i alder på ét år øgedes den gennemsnitlige tid brugt i svømmebassin pr uge med 0,2 timer (95%-konfidensinterval = [0,12; 0,26]*). Ved alderen 0 var den gennemsnitlige tid brugt i svømmebassinet 7,93 timer pr uge (95%-konfidensinterval* [7,00; 8,87])

7. Jeg indsætter 13 som age i modellen og prædikterer en gennemsnitlig tid brugt i svømmebassin pr uge på 10,37 timer.
   Ved at bruge spredningen, som er omtalt i opg. 3.5, kan jeg udregne et 95% referenceinterval på [8,06; 12,67].

8. Multipel lineær regression. Nu får vi den binære variabel ind som et ekstra led i den uafhængige variabel, og har derfor en uafhængig variabel der indeholder både et kontinuert og et binær led. Den afhængige variabel er udelukkende kontinuert. Derfor skal vi bruge multipel lineær regression.

   Jeg vil først estimere regressionen med crawl som hovedeffekt.
   Først bruger jeg lm-kommandoen med crawl som hovedeffekt.
   Jeg får følgende regressionsmodeller (p < 0,001 for alle tre parametre):

$$pooltime = \begin{cases} 7{,}75 + 0{,}23 \cdot age, & hvis\ crawl = 0 \\ 6{,}38 + 0{,}23 \cdot age, & hvis\ crawl = 1 \end{cases}$$

(For at være sikker på, at RStudio bruger crawl=1 som referenceværdi, brugte jeg gsub-funktionen til at ændre 1 til Yes og 0 til No i d$crawl. Så kan jeg se i summary at den som antaget bruger crawl=1 til at beregne hovedeffekten.)

Nu estimerer jeg modellen med både hovedeffekt og interaktionsled for crawl for at se om der er statistisk evidens for, at hældningskoefficienten for crawlsvømmere er signifikant anderledes end for ikke-crawlsvømmere.
Nu bliver p-værdien for nulhypotesen svarende til, at interaktionsleddet er lig 0 større end 0,05 (0,80) og p-værdien for nulhypotesen at hovedeffekten er lig 0 stiger også til 0,262. Derfor kan vi IKKE forkaste nulhypoteserne, og vi foretrækker derfor modellen med crawl som hovedeffekt og ingen interaktionseffekt med alderen.

9. Som opskrevet i opg. 3.8 estimerer jeg udtrykket for sammenhængende i modellen med crawl som hovedeffekt til at være:

$$pooltime = \begin{cases} 7{,}75 + 0{,}23 \cdot age, & hvis\ crawl = 0 \\ 6{,}38 + 0{,}23 \cdot age, & hvis\ crawl = 1 \end{cases}$$

a-værdien, eller skæring med y-aksen, er højere for ikke-crawlsvømmere end for crawlsvømmere med 1,37 timers gennemsnitlig tid i bassinet pr uge.
b-værdien, eller hældningskoefficienten, er 0,23 timers gennemsnitlig tid i bassinet pr uge, og derfor stiger y-værdien med 0,23 timer hver gang alderen stiger med én.

10. 95%-konfidensintervallet er $[-1{,}75; -0{,}99]$
Jeg er 95% sikker på, at den sande effekt af crawlsvømning ift. andre svømmestile i min endelige model ligger i intervallet $[-1{,}75; -0{,}99]$.

11. P-værdierne for både min a- og b-værdi i den endelige model med crawl som hovedeffekt er alle tre langt under 0,05. Dermed kan jeg konkludere, at jeg kan forkaste nulhypotesen om, at nogen af disse parametre er lig 0. Dette betyder, at modellen er korrekt valgt, da kompleksitetstrinnet op til en model med interaktionsled medførte nulhypoteser, der ikke kunne forkastes. I statistik ønsker vi altid at bruge den mest simple model, der forklarer data bedst muligt. Dette har jeg gjort.
Vi kan på baggrund af de forkastede nulhypoteser konkludere følgende:

Sammenhængen mellem tid brugt i svømmebassinet og alder under hensyntagen til crawlsvømning (p<0,001) blev analyseret ved multipel lineær regression. For både crawlsvømmere og ikke-crawlsvømmere steg tiden gennemsnitligt brugt i svømmebassin pr uge hvert år med 0,23 timer (95%-konfidensinterval = [0,17;0,29], p<0,001), og i gennemsnit var tid gennemsnitligt i svømmebassin pr uge 1,37 timer (95%-konfidensinterval = [-1,75;-0,99], p<0,001) lavere for crawlsvømmere end for ikke-crawlsvømmere.
Der blev ikke fundet nogen signifikant interaktion mellem alder og crawlsvømning (p=0,80), og derfor er der ikke evidens for, at tidsforbruget ændrer sig forskelligt over tid afhængigt af svømmestil.

12. Jeg indsætter 12 som age i modellen fra opg. 3.9 for crawl=1:

$$pooltime = 6{,}38 + 0{,}23 \cdot 12 = 9{,}14\ gennemsnitlige\ timer\ i\ bassin\ pr\ uge$$

13. Samme procedure, nu for formlen med crawl=0:

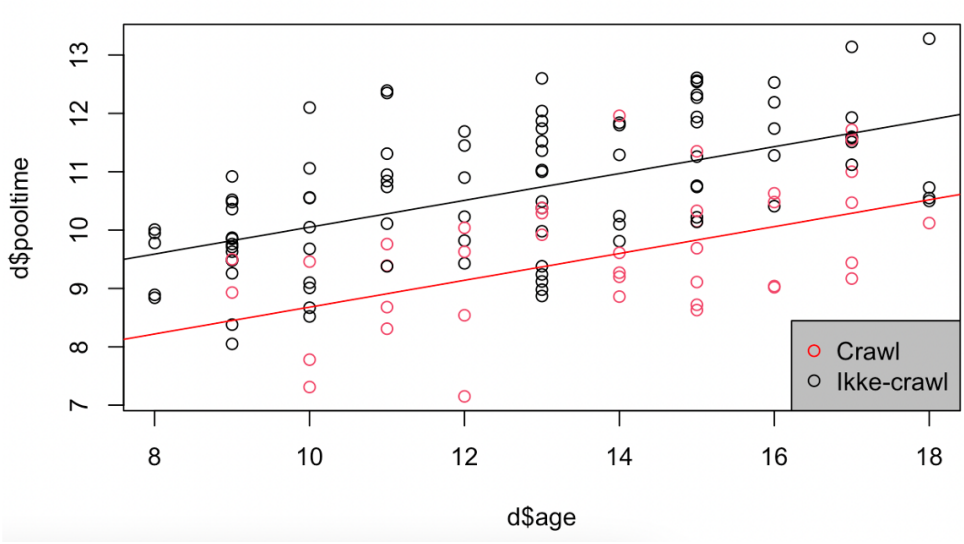
$$pooltime = 7{,}75 + 0{,}23 \cdot 12 = 10{,}51\ gennemsnitlige\ timer\ i\ bassin\ pr\ uge$$

14. Samme procedure:

$$pooltime = \begin{cases} 7{,}75 + 0{,}23 \cdot 4 = \ 8{,}67\ timer \\ \ 6{,}38 + 0{,}23 \cdot 4 = \ 7{,}3\ timer \end{cases}$$

Tallene er ikke de samme. Da vi har forskellige a-værdier for de to forskellige svømmestilsgrupper, vil to svømmere fra hver sin gruppe aldrig få det samme tidsforbrug, da funktionen er lineær, og forskellen derfor vil persistere uafhængigt af x-værdi.

15. Nedenfor er en figur af modellen, der illustrerer de ovenstående sammenhænge:



Opgave 4

Jeg syntes opgaveformuleringen har været lidt uklar, og har derfor gjort mig nogle forbehold i delopgave 4.1-4.3.
Med indtag af kulsyreholdig læskedrik mindre end 5 gange i ugen er jeg i tvivl om der udelukkende menes kategorien <5/week eller kategorierne <5/week og never. Mine beregninger i 4.1-4.3 bygger på, at mindre end 5 gange i ugen inkluderer begge kategorierne <5/week og never, da det ud fra konteksten giver mest mening. I 4.4 har jeg adskilt dem, da det her står skrevet eksplicit, og konteksten i denne delopgave undersøger en anden sammenhæng.

1. Sandsynligheden for at indtage kulsyreholdig læskedrik mindre end 5 gange om ugen estimeres til at være 57%. Her har jeg udregnet den relative frekvens på baggrund af de besvarelser vi har, og har dermed ikke medregnet den del af besvarelserne, hvor der ikke er data for indtag af kulsyreholdig læskedrik.
   95%-konfidensintervallet udregnes til at være [0,47; 0,66] ved hjælp af prop.test-kommandoen.
   I vores stikprøve er der gennemsnitligt 57% chance for at drikke kulsyreholdig læskedrik mindre end 5 gange om ugen. Rstudio tester her nulhypotesen, at samme

sandsynlighed er 50% i populationen. P-værdien for dette er 0,17, og derfor kan vi IKKE forkaste nulhypotesen. Det kan også ses i 95% konfidensintervallet, at 50% sandsynlighed ligger inde i intervallet.
Vi kan på baggrund af dette konkludere, at der **IKKE** er statistisk evidens for at kunne forkaste nulhypotesen, at sandsynligheden for at indtage kulsyreholdig læskedrik mindre en 5 gange i ugen er 50%.

2. Odds er antallet af personer, som er omfattet af en hændelse, divideret med den del af populationen, som ikke er omfattet af samme hændelse.
   Igen beregner jeg kun på baggrund af de personer, som vi har data for indtag af kulsyreholdig læskedrik på:

$$\frac{57 + 10}{51} = 1{,}31$$

   Det kan fortolkes som følger: For hvert barn, der drikker kulsyreholdig læskedrik 5 eller flere gange i ugen, forventer vi 1,31 barn, der drikker det færre end 5 gange i ugen.

3. Risikodifferens:
   Jeg bruger table-kommandoen til at opstille følgende matrix:

Kulsyreholdig læskedrik	$Crawl: Ja$	$Crawl: Nej$
$< 5/ugen$	18	43
$\geq 5\ ugen$	16	41

   Estimeret risiko for at svømme crawl og drikke kulsyreholdig læskedrik <5 gange i ugen: 29,5%
   Estimeret risiko for at svømme crawl og drikke kulsyreholdig læskedrik ≥5 gange i ugen: 28,1%

   Risikodifferensen er lig 1,4 procentpoint. Det er forskellen i risiko.

Den tilsvarende relative risiko er lig $\frac{29,5}{28,1} = 1,05$. Altså er der 5% større risiko for at svømme crawl, hvis du drikker kulsyreholdig læskedrik mindre end 5 gange i ugen ift. gruppen der drikker kulsyreholdig læskedrik 5 eller flere gange i ugen.

Altså er der en estimeret større risiko for at svømme crawl, hvis du drikker mindre kulsyreholdig læskedrik.

4. Odds for crawlsvømning blandt dem, der aldrig drikker kulsyreholdig læskedrik, er $\frac{3}{7} = 0,43$
   Odds for crawlsvømning blandt dem, der drikker kulsyreholdig læskedrik <5 gange i ugen, er $\frac{15}{36} = 0,42$
   Oddsratioen er $\frac{\frac{3}{7}}{\frac{15}{36}} = 1,03$
   Det vi ser er, at der er næsten lige store odds for at svømme crawl blandt dem, der ikke drikker kulsyreholdig læskedrik, og dem der drikker kulsyreholdig læskedrik <5 gange i ugen. Der er 3% højere chance for at svømme crawl, hvis du ikke drikker kulsyreholdig læskedrik, end hvis du drikker kulsyreholdig læskedrik <5 gange i ugen. For dem, der ikke drikker kulsyreholdig læskedrik, er der 0,43 der svømmer crawl, for hver person der ikke gør. For dem, der drikker kulsyreholdig læskedrik <5 gange i ugen, er der 0,42 personer, der svømmer crawl, for hver person der ikke gør.

Opgave 5

1. χ^2-test Den afhængige variabel er binær. Den uafhængige variabel er binær. Med analysen kan vi se om der er signifikant evidens for om crawlsvømning påvirker risikoen for ætsninger på tænder.

2. Ved analyse får jeg p<0,001, og jeg kan dermed forkaste nulhypotesen om, at der ikke er sammenhæng mellem den uafhængige og den afhængige variabel. Dermed er der signifikant sammenhæng mellem crawlsvømning og risikoen for ætsninger på tænder.

3. Jeg har i RStudio opstillet følgende matrix:

$Crawl$	$Ætsninger$	$Ingen\ ætsninger$
Ja	21	18
Nej	9	83

Estimeret risiko for ætsninger på tænder for crawlsvømmere: 53,8%
Estimeret risiko for ætsninger på tænder for ikke-crawlsvømmere: 9,8%

Risikodifferens

$$53{,}8\% - 9{,}8\% = 44\%$$

Altså er risikodifferencen 44 procentpoint, og der er således en meget stor risikodifferens. Vi kan allerede nu se, at der er langt større risiko for ætsninger som crawlsvømmer end for ikke-crawlsvømmere.

4. Den relative risiko er:

$$\frac{53{,}8}{9{,}8} = 549\%$$

   Altså er den relative risiko for at få ætsninger, hvis man svømmer crawl, 449% større end hvis man har en anden svømmestil. Det er meget.

5. Odds for ætsninger på tænder blandt crawlsvømmere: 1,17
   Odds for ætsninger på tænder blandt ikke-crawlsvømmere: 0,11

   Oddsratioen:

$$\frac{1{,}1666}{0{,}1084} = 10{,}76$$

Der er altså 10,76 gange større (eller 976% højere) odds for at få ætsninger på tænderne ved crawl end ved anden svømmestil. Altså er der næsten 10 gange så mange personer med ætsninger på tænderne i forhold til personer uden ætsninger på tænderne blandt crawlsvømmere i forhold til ikke-crawlsvømmere.

6. Logistisk regression. Den uafhængige variabel er kontinuert numerisk mens den afhængige variabel er binær. Derfor bruger jeg logistisk regression til at analysere sammenhængen.

7. Odds-skala:

$$Odds(erosion) = e^{3{,}305 - 0{,}44 \cdot pooltime}$$

Den estimerede effekt af tid, eller odds-ratio, er $e^{-0{,}4442} = 0{,}64$. Dette betyder, at for hver gang den gennemsnitlige mængde tid i svømmebassinet pr uge stiger med en time, bliver odds for at få ætsninger 0,64 gange mindre, svarende til et fald på ca. 36%.

95%-konfidensintervallet for effekten af gennemsnitlig mængde tid i svømmebassinet pr uge på ætsninger udregnes til [0,44; 0,90]. Dette stemmer overens med vores nedenstående konklusion om signifikant effekt af tid brugt i svømmebassinet, da intervallet ikke indeholde 1.

Den testede nulhypotese er, at gennemsnitlig mængde tid i svømmebassinet pr uge IKKE har en effekt på ætsninger på tænderne, dvs. at odds-ratioen er 1 og regressionskoefficienten er 0.
Denne nulhypotese kan vi forkaste, da p=0,01 for denne effekt. En lille detalje her er, at nulhypotesen om, at a-værdien er 0, ikke kan forkastes, da p=0,068. En ændring i a-værdi forskyder dog blot sammenhængen på y-aksen, og her er vi mest af alt interesserede i b-værdien.

Vi kan konkludere, at gennemsnitlig mængde tid i svømmebassinet pr uge har en statistisk signifikant effekt (p=0,01 og odds-ratio på 0,64) på risikoen for ætsninger i tænderne med et 95%-konfidensinterval på [0,44; 0,90] *for effekten*.

8. Kun for crawlsvømmere:
   Odds-skala:

$$Odds(erosion) = e^{2,14-0,21 \cdot pooltime}$$

Den estimerede effekt af tid, eller odds-ratio, er $e^{-0,4442} = 0,81$. Dette betyder, at for hver gang den gennemsnitlige mængde tid i svømmebassinet pr uge stiger med en time, bliver odds for at få ætsninger 0,81 gange mindre, svarende til et fald på ca. 19%.

95%-konfidensintervallet for effekten af gennemsnitlig mængde tid i svømmebassinet pr uge på ætsninger udregnes til [0,44; 1,45]. Her indgår 1 i odds-ratioen. Det betyder, at vi er 95% sikre på at odds-ratio ligger i et interval, hvor det kan have værdien 1, og dermed betyde, at der ikke er nogen effekt. Allerede her kan vi ikke forkaste nulhypotesen.

P-værdien for nulhypotesen, at gennemsnitlig mængde tid i svømmebassinet pr uge IKKE har en effekt på ætsninger på tænderne hos crawlsvømmere, er 0,49, og derfor kan vi IKKE forkaste nulhypotesen. P-værdien for nulhypotesen at a-værdien er lig 0, er også høj nok til, at vi ikke kan forkaste nulhypotesen for denne.

Vi kan konkludere, at gennemsnitlig mængde tid i svømmebassinet pr uge for crawlsvømmere ***IKKE*** *har statistisk signifikant effekt (p=0,49 og odds-ratio på 0,81) på risikoen for ætsninger i tænderne med et 95%-konfidensinterval* på [0,44; 1,45].

9. Kun for ikke-crawlsvømmere:

   Odds-skala:

$$Odds(erosion) = e^{-2,12-0,01 \cdot pooltime}$$

Den estimerede effekt af tid, eller odds-ratio, er $e^{-0,009982} = 0,99$. Dette betyder, at for hver gang den gennemsnitlige mængde tid i svømmebassinet pr uge stiger med en time, bliver odds for at få ætsninger 0,99 gange mindre, svarende til et fald på ca. 1%.

95%-konfidensintervallet for effekten af gennemsnitlig mængde tid i svømmebassinet pr uge på ætsninger udregnes til [0,55; 1,78]. Her indgår 1 i 95%-konfidensintervallet. Det betyder, at vi er 95% sikre på, at odds-ratio ligger i et interval, hvor det kan have værdien 1, og dermed betyde, at der ikke er nogen effekt. Her kan vi igen allerede afvise at nulhypotesen kan forkastes.

P-værdien for nulhypotesen, at gennemsnitlig mængde tid i svømmebassinet pr uge IKKE har en effekt på ætsninger på tænderne hos ikke-crawlsvømmere, er 0,97, og derfor kan vi IKKE forkaste nulhypotesen. P-værdien for nulhypotesen om at a-værdien er lig 0, er også høj nok til, at vi ikke kan forkaste nulhypotesen.

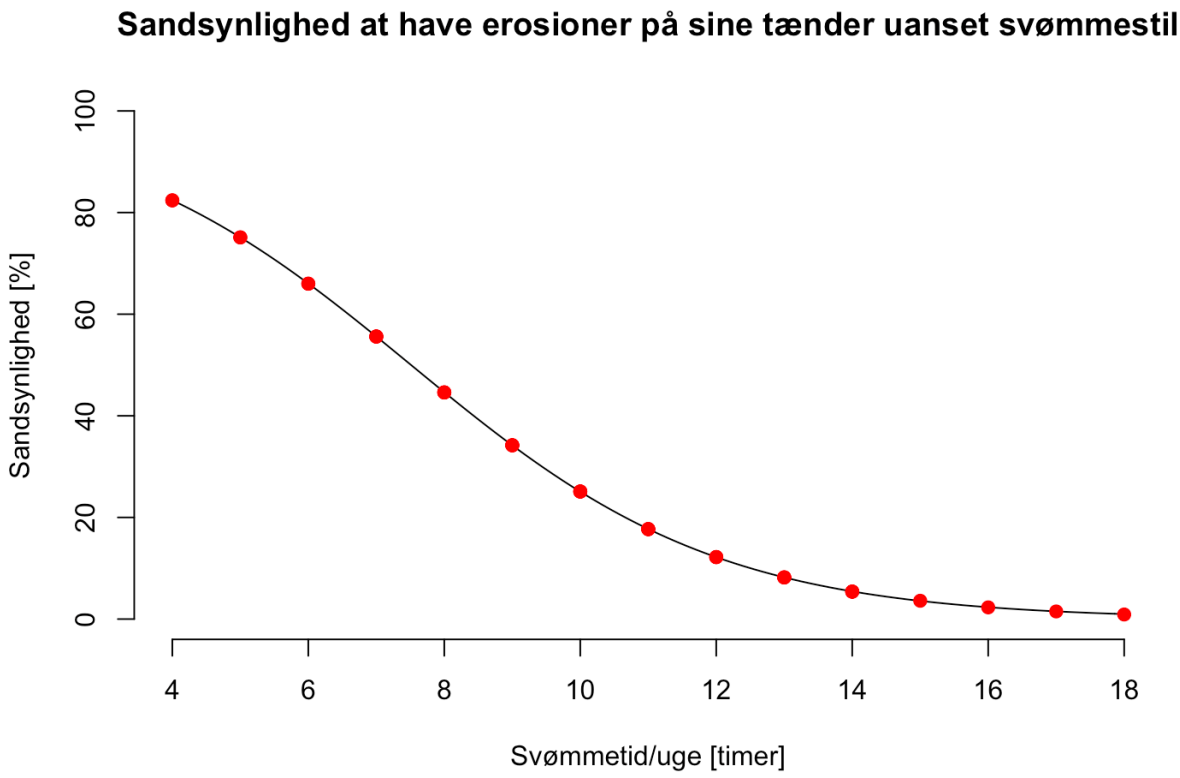
Vi kan konkludere, at gennemsnitlig mængde tid i svømmebassinet pr uge for ikke-crawlsvømmere ***IKKE*** *har statistisk signifikant effekt (p=0,97 og odds-ratio på 0,99) på risikoen for ætsninger i tænderne med et 95%-konfidensinterval* på [0,55; 1,78].

10. Konklusionen fra opgave 5.7 er, at tid i svømmebassinet har en statistisk signifikant effekt på risikoen for ætsninger i tænderne. Denne effekt er ikke negativ, som man skulle tro baseret på tidligere udregninger i opg. 5, men derimod positiv. Det skal forstås i den forstand, at et øget antal timer vil medføre nedsat odds for at få ætsninger på tænderne.

    Konklusionerne i opg. 5.8 og 5.9 støtter reelt set op om dette, men viser ingen statistisk signifikant effekt. Dette kan der være flere årsager til, men en af de primære faktorer kan være det nedsatte antal personer i stikprøverne.
    Crawlsvømmere har markant (10x) forøget øget odds for at få ætsninger på tænderne ift. sine andre medsvømmere. Dog har vi ikke kunne finde en statistisk signifikant

sammenhæng, der forklarede dette - hverken når det kom til tid i bassinet eller indtag af kulsyreholdige læskedrikke. Derfor ender vi i en situation, hvor vi har bevist en sammenhæng i en population, men ikke kan finde årsagen til denne sammenhæng, med de data vi har. Suk.

Slutteligt har jeg lavet en graf for sammenhængen fundet i 5.7 der viser, at det bedste du kan gøre for dine tænder, er at bruge en masse timer i svømmebassinets klorvand. Det er selvfølgelig en joke, men om ikke illustrerer det fint konklusionen fra opg. 5.7 grafisk.



Appendix

```r
#Opgave 1
d <- read.csv("http://causal.sund.ku.dk/f22/63.csv", header=TRUE, stringsAsFactors=TRUE)
derosionJA <- subset(d, erosion == "1")
derosionNEJ <- subset(d, erosion == "0")
hist(derosionJA$age)
hist(derosionNEJ$age)
table(derosionJA$soda)
table(derosionNEJ$soda)
qqnorm(derosionJA$age)
qqline(derosionJA$age)
qqnorm(derosionNEJ$age)
qqline(derosionJA$age)
mean(derosionJA$age)
median(derosionJA$age)
mean(derosionNEJ$age)
median(derosionNEJ$age)
sd(derosionJA$age)
sd(derosionNEJ$age)
hist(derosionJA$pooltime)
hist(derosionNEJ$pooltime)
IQR(derosionJA$age)
median(derosionJA$age)
IQR(derosionNEJ$age)
median(derosionNEJ$age)
table(derosionJA$crawl)
table(derosionNEJ$crawl)
table(derosionJA$soda)
table(derosionNEJ$soda)
qqnorm(derosionNEJ$pooltime)
qqline(derosionNEJ$pooltime)
qqnorm(derosionJA$pooltime)
qqline(derosionJA$pooltime)
median(derosionJA$pooltime)
IQR(derosionJA$pooltime)
median(derosionNEJ$pooltime)
IQR(derosionNEJ$pooltime)

```

```r
#Opgave 2
logm <- glm(d$crawl ~ d$age, family = binomial)
summary(logm)
confint(logm)
exp(-0.1395248)
exp(-0.06411011)
exp(0.16605)
exp(0.02915454)
exp(0.3109659)
exp(-3.044+0.166*9)/(1+exp(-3.044+0.166*9))
exp(6.44256 - 0.09757 * 20) / (1 + exp(6.44256 - 0.09757 * 20))
(exp(-3.044+0.166*14)/(1+exp(-3.044+0.166*14)))-(exp(-3.044+0.166*11)/(1+exp(-3.044+0.166*11)))
exp(-3.044)
exp(0)

#Opgave 3
crawlNEJ <- subset(d, crawl == "0")
crawlJA <- subset(d, crawl == "1")
t.test(d$pooltime ~ d$crawl)
hist(crawlJA$pooltime)
qqnorm(crawlJA$pooltime)
qqline(crawlJA$pooltime)
sd(crawlJA$pooltime)
mean(crawlJA$pooltime)-2*sd(crawlJA$pooltime)
mean(crawlJA$pooltime)+2*sd(crawlJA$pooltime)
model <- lm(d$pooltime ~ d$age)
summary(model)
plot(d$age, d$pooltime)
abline(7.934, 0.187)
confint(model)
7.934+0.187*13
10.365-2*1.154
10.365+2*1.154
m1 <- lm(d$pooltime ~ d$age + d$crawl)
d$crawl<-gsub(0,"No",d$crawl)
d$crawl<-gsub(1,"Yes",d$crawl)
summary(m1)
m2 <- lm(d$pooltime ~ d$age * d$crawl)
summary(m2)
confint(m1)
plot(d$age, d$pooltime, col=d$crawl)
legend("bottomright",legend=c("Crawl", "Ikke-crawl"),
       pch=1, col=c("red","black"), bg="grey")
abline(7.75,0.23)
abline(6.38,0.23, col='red')
d <- read.csv("http://causal.sund.ku.dk/f22/63.csv", header=TRUE, stringsAsFactors=TRUE)
```

```r
#Opgave 4
table(d$soda)
(57+10)/(51+57+10)
(57+10)/51
prop.test(67, 67+51)
table(d$soda, d$crawl)
18/(18+43)
16/(16+41)
ikkesoda <- subset(d, soda == "never")
lidtsoda <- subset(d, soda == "<5/week")
table(ikkesoda$crawl)
table(lidtsoda$crawl)
3/7
15/36
(3/7)/(15/36)
table(crawlJA$erosion)
table(crawlNEJ$erosion)


```

```r
#Opgave 5
m5 <-matrix(c(21, 9, 18, 83), nrow = 2, ncol = 2)
chisq.test(m5)
21/(21+18)
9/(9+83)
21/18
9/83
(21/18)/(9/83)

m52 <- glm(d$erosion ~ d$pooltime, family = binomial)
summary(m52)
exp(-0.4442)
confint (m52)
exp(-0.81511578)
exp(-0.1035912)

m53 <- glm(crawlJA$erosion ~ crawlJA$pooltime, family = binomial)
summary(m53)
exp(-0.2067)
confint (m53)
exp(-0.8286335)
exp(0.3755577)

m54 <- glm(crawlNEJ$erosion ~ crawlNEJ$pooltime, family = binomial)
summary(m54)
exp(-0.009982)
confint (m54)
exp(-0.5999418)
exp(0.575402)
```

```r
exp(3.305 - 0.44 * 4) / (1 + exp(3.305 - 0.44 * 4))
exp(3.305 - 0.44 * 5) / (1 + exp(3.305 - 0.44 * 5))
exp(3.305 - 0.44 * 6) / (1 + exp(3.305 - 0.44 * 6))
exp(3.305 - 0.44 * 7) / (1 + exp(3.305 - 0.44 * 7))
exp(3.305 - 0.44 * 8) / (1 + exp(3.305 - 0.44 * 8))
exp(3.305 - 0.44 * 9) / (1 + exp(3.305 - 0.44 * 9))
exp(3.305 - 0.44 * 10) / (1 + exp(3.305 - 0.44 * 10))
exp(3.305 - 0.44 * 11) / (1 + exp(3.305 - 0.44 * 11))
exp(3.305 - 0.44 * 12) / (1 + exp(3.305 - 0.44 * 12))
exp(3.305 - 0.44 * 13) / (1 + exp(3.305 - 0.44 * 13))
exp(3.305 - 0.44 * 14) / (1 + exp(3.305 - 0.44 * 14))
exp(3.305 - 0.44 * 15) / (1 + exp(3.305 - 0.44 * 15))
exp(3.305 - 0.44 * 16) / (1 + exp(3.305 - 0.44 * 16))
exp(3.305 - 0.44 * 17) / (1 + exp(3.305 - 0.44 * 17))
exp(3.305 - 0.44 * 18) / (1 + exp(3.305 - 0.44 * 18))


timer <- c(4,5,6,7,8,9,10,11,12,13,14, 15, 16, 17, 18)
sandsynlighed <- c(82.4, 75.11, 66, 55.6, 44.6, 34.2, 25.1, 17.7, 12.2, 8.2, 5.4, 3.6, 2.3, 1.5,0.9)
curve(100 * exp(3.305 - 0.44 * x) / (1 + exp(3.305 - 0.44 * x)),
      from=4, to=18, ylim=c(0,100), xlab="Svømmetid/uge [timer]", ylab="Sandsynlighed [%]",
      main="Sandsynlighed at have erosioner på sine tænder uanset svømmestil", bty="n")
points(timer, sandsynlighed, col='red', pch=19)
```