Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 1, April 2018

Vejledende besvarelse udarbejdet af Steffen Lauritzen og Helle Sørensen

Opgave 1

Ved opsendelsen af rumfærgen Challenger den 28. januar 1986 omkom syv astronauter i forbindelse med en eksplosion. En undersøgelseskommission blev nedsat og fastslog at årsagen til ulykken var en såkaldt O-ring, som gik i stykker i forbindelse med opsendelsen, sandsynligvis forårsaget af ekstremt koldt vejr. Filen `challenger.txt` indeholder data fra 23 tidligere opsendelser af rumfærgen. Kolonnen `temp` angiver temperaturen i Fahrenheit ved opsendelsen og kolonnen `critical` angiver om man i forbindelse med opsendelsen har observeret en kritisk tilstand for en af rumfærgens seks O-ringe, idet 1 angiver at man har observeret et problem og 0 at man ikke har.

1. Opstil en passende generaliseret lineær model til beskrivelse af afhængigheden mellem opsendelsestemperaturen og en kritisk begivenhed og angiv maximum likelihood estimatet for de indgående parametre.

   Lad $Y_1, \ldots, Y_{23}$ være indbyrdes uafhængige og Bernoullifordelte med parametre μ_i, hvor

   $$\eta_i = \text{logit}(\mu_i) = \log \frac{\mu_i}{1-\mu_i} = \alpha + \beta t_i$$

   hvor t_i angiver temperaturen ved den i-te opsendelse og Y_i indikerer om der har været en kritisk hændelse i forbindelse med samme opsendelse.

   Dette er en generaliseret lineær model med binomial fejlfordeling, kanonisk link, og opsendelsestemperaturen som kovariat. Den specificeres som følger:

   ```
   > logreg <- glm(critical ~ temp, family="binomial", data=challenger)
   ```

   og estimater for parametrene (α, β) fås via kommandoen

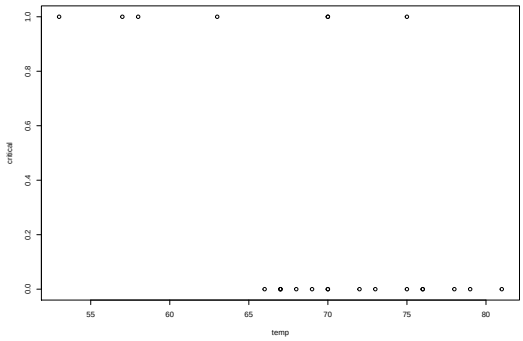
   ```
   > summary(logreg)
   ...
   Coefficients:
               Estimate Std. Error z value Pr(>|z|)
   (Intercept)  15.0429     7.3786   2.039   0.0415 *
   temp         -0.2322     0.1082  -2.145   0.0320 *
   ...
   ```

   hvilket giver estimaterne $(\hat{\alpha}, \hat{\beta}) = (15.0439, -0.2322)$.

2. Brug modellen og de angivne data til at afgøre om opsendelsestemperaturen har betydning for sandsynligheden for en kritisk hændelse.

   Et asymptotisk test for temperaturafhængigheden giver iflg. ovenstående output en p-værdi på 0.032 hvilket er signifikant på 5% niveau; så man må konkludere at at opsendelsestemperaturen har en betydning for forekomst af kritiske hændelser.

   Et plot af data giver samme konklusion idet der aldrig er set kritiske hændelser ved temperaturer over 75 grader og altid kritiske hændelser ved temperaturer under 65 grader.



3. Da rumfærgen blev sendt op var temperaturen 31 grader Fahrenheit. Estimer sandsynligheden for en kritisk hændelse med en O-ring ved denne temperatur.

   Her fås

$$\hat{\mu}(31) = \frac{e^{\hat{\alpha}+31\hat{\beta}}}{1+e^{\hat{\alpha}+31\hat{\beta}}} = 0.9996$$

   så hvis denne voldsomme ekstrapolation ellers står til troende, ville man med stor sikkerhed forvente en kritisk hændelse.

4. Find et approksimativt 95% konfidensinterval for den samme sandsynlighed som estimeret under punkt 3. *Vink:* Brug for eksempel deltametoden på funktionen $f(\alpha,\beta) = \alpha + 31\beta$.

   Den approximative kovariansmatrix for estimaterne kan fås ved kommandoen

```
> vcov(logreg)
             (Intercept)        temp
(Intercept)  54.4441826 -0.79638547
temp         -0.7963855  0.01171512
```

   Ved deltametoden fås dernæst

$$\mathbf{V}\{f(\hat{\alpha},\hat{\beta})\} \approx \mathbf{V}(\hat{\alpha}) + 31^2\mathbf{V}(\hat{\beta}) + 62\mathbf{V}(\hat{\alpha},\hat{\beta}) = 16.3265$$

   og videre fås nu et approximativt 95% konfidensinterval for den lineære prediktor $f(\alpha,\beta)$

$$\hat{\alpha} + 31\hat{\beta} \pm 1.96\sqrt{16.3265} = (-0.07, 15.8)$$

og dermed et approximativt 95% konfidensinterval for den ønskede sandsynlighed til

$$\left(\frac{e^{-0.07}}{1+e^{-0.07}}, \frac{e^{15.8}}{1+e^{15.8}}\right) = (0.482, 1).$$

Bemærk, at usikkerheden på denne interpolation er ganske stor, så selvom man estimerer sandsynligheden til at være tæt på 100%, kunne den også være så lille som 48%. Men selv den nedre grænse er stor nok til, at man klart vil fraråde opsendelse.

Opgave 2

Paretofordelingen anvendes for eksempel til at beskrive fordelingen af formuer over en givet tærskelværdi c og den har tæthedsfunktion

$$f_\theta^c(x) = \frac{\theta c^\theta}{x^{\theta+1}}, \quad \text{for } x > c,$$

hvor $c > 0$ er fast og kendt mens $\theta > 0$ er en ukendt parameter som kaldes fordelingens *index*.

1. Gør rede for, at familien af Paretofordelinger med fast tærskel c udgør en eksponentiel familie; angiv familiens grundmål.

   Vi omskriver tæthedsfunktionen således

   $$f_\theta(x) = \frac{\theta c^\theta}{x^{\theta+1}} = \theta c^\theta e^{-\theta \log x} \frac{1}{x}, \quad x > c$$

   hvorefter vi genkender den eksponentielle form med grundmål $\mu = \mathbf{1}_{(c,\infty)}(x)/x \cdot \lambda$, hvor λ er Lebesguemålet.

2. Angiv familiens dimension, den kanoniske parameter, den kanoniske stikprøvefunktion, og kumulantfunktionen.

   Familien har dimension 1, den kanoniske parameter er θ og den kanoniske stikprøvefunktion er $t(x) = -\log x$; kumulantfunktionen er

   $$\psi(\theta) = -\log(\theta c^\theta) = -\log\theta - \theta \log c.$$

   Alternativt kan man skrive

   $$f_\theta(x) = \theta e^{-\theta \log(x/c)} \frac{1}{x}, \quad x > c$$

   hvor nu den kanoniske stikprøvefunktion er $-\log(x/c)$ og kumulantfuntionen bliver $\psi^*(\theta) = -\log\theta$.

   Endnu et alternativ er at skrive

   $$f_\theta(x) = \frac{\theta c^\theta}{x^{\theta+1}} = \theta c^\theta e^{-(\theta+1)\log x}, \quad x > c$$

   med kanonisk parameter $\theta + 1$ og grundmål Lebesguemålet på den positive akse.

3. Lad nu $X_1, \ldots, X_n$ være uafhængige og Paretofordelte med kendt tærskel c og ukendt index θ. Find maximum likelihood estimatoren for θ og angiv dens asymptotiske fordeling.

   Vi finder først middelværdifunktionen

   $$\tau(\theta) = \mathbf{E}_\theta(-\log X) = \psi'(\theta) = -1/\theta - \log c$$

   hvilket fører til likelihood ligningen

   $$\sum_{i=1}^{n} \log x_i = \frac{n}{\theta} + n \log c$$

   som har den entydige løsning

   $$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \{\log(x_i) - \log c\}}.$$

   $\hat{\theta}$ er asymptotisk normalfordelt med den reciprokke Fisher information som asymptotisk varians. Vi får

   $$\kappa(\theta) = i(\theta) = \psi''(\theta) = \tau'(\theta) = \theta^{-2}$$

   og dermed er

   $$\hat{\theta} \overset{\text{as}}{\sim} N(\theta, \theta^2/n).$$

Tidsskriftet *Forbes magazine* angiver hvert år en liste over verdens største personlige formuer. Nedenfor ses for 2018 størrelsen af alle personlige formuer over 50 milliarder US dollars som angivet af Forbes magazine.

År	Formue i milliarder US dollars								
2018	112	90	84	72	71	70	67	60	58.5

4. Under antagelse af at disse observationer følger en Paretofordeling med tærskel $c = 50$ og ukendt indexparameter, ønskes værdien af maximum likelihood estimatoren $\hat{\theta}$ samt et approksimativt 95% konfidensinterval for indexparameteren θ.

   Maximum likelihood estimatet beregnes her til $\hat{\theta} = 2.502$ og den tilsvarende approximative standardafvigelse til $\sqrt{\hat{\theta}^2/9} = \hat{\theta}/3 = 0.834$.

   Baseret på den asymptotiske fordeling får vi så konfidensintervallet

   $$\hat{\theta} \pm 1.96 \cdot \hat{\theta}/3 = (0.867, 4.137).$$

Opgave 3

Lad $X_1, \ldots, X_n$ være uafhængige og identisk gammafordelte med med samme skala- og formparameter, d.v.s. deres fordeling har tæthed

$$f_\theta(x) = \frac{x^{\theta-1} e^{-x/\theta}}{\Gamma(\theta)\theta^\theta}, \quad x > 0$$

hvor $\theta > 0$ er ukendt.

I det følgende indgår *digamma*funktionen ψ og *trigamma*funktionen ψ', hvor

$$\psi(y) = D\log\Gamma(y) = \frac{\Gamma'(y)}{\Gamma(y)}, \quad \psi'(y) = D^2\log\Gamma(y).$$

Begge funktioner er implementeret som standard i `R` og kaldes som `digamma()` og `trigamma()`.

Antag nu, at der foreligger en observation $x = (x_1, \ldots, x_n)$.

1. Bestem log-likelihoodfunktionen og scorefunktionen.

   Vi får, idet vi ignorerer led som kun afhænger af observationerne

   $$\ell_x(\theta) = -\log L_x(\theta) = n\log\Gamma(\theta) + n\theta\log\theta - \theta\sum_i \log x_i + \sum_i x_i/\theta$$

   og for scorefunktionen ved differentiation

   $$S(x,\theta) = n\psi(\theta) + n\log\theta + n - \sum_i \log x_i - \sum_i x_i/\theta^2.$$

2. Bestem informationsfunktionen og Fisherinformationen.

   Ved yderligere differentiation fås

   $$I(x,\theta) = n\psi'(\theta) + n/\theta + 2\sum_i x_i/\theta^3. \tag{1}$$

   I en gammafordeling $\Gamma(\alpha,\beta)$ hvor β er skalaparameter, er middelværdien $\mathbf{E}(X) = \alpha\beta$ så i vores tilfælde er middelværdien $\mathbf{E}_\theta(X) = \theta^2$. Heraf følger at informationsfunktionen i tilfældet $n = 1$ er givet som

   $$i(\theta) = \psi'(\theta) + 1/\theta + 2/\theta = \psi'(\theta) + 3/\theta.$$

3. Scoreligningen kan ikke løses explicit. Vis, at scoreligningen har en entydig løsning og at denne løsning $\hat\theta$ er maximum likelihood estimator for θ; det kan uden bevis benyttes, at

   $$\psi'(y) = \sum_{k=0}^{\infty} \frac{1}{(y+k)^2}.$$

   Idet $\psi'(\theta) > 0$ og $x_i > 0$ giver (**??**) at scorefunktionen er strengt voksende; da vi for alle x har at

   $$\lim_{\theta\to 0} S(x,\theta) = -\infty, \quad \lim_{\theta\to\infty} S(x,\theta) = \infty$$

   har scoreligningen præcis en løsning.

4. Angiv maximum likelihood estimatorens asymptotiske fordeling.

   Maksimum likelihood estimatoren er asymptotisk normalfordelt med den reciprokke information som varians, altså er

   $$\hat\theta \overset{\text{as}}{\sim} N\left(\theta, \frac{1}{n(\psi'(\theta) + 3/\theta)}\right).$$

5. En alternativ estimator for θ er givet som

$$\tilde{\theta} = \sqrt{\frac{\sum_i x_i}{n}} = \sqrt{\bar{x}}.$$

Gør rede for, at denne estimator er en momentestimator; er estimatoren central?

Estimatoren er en momentestimator i og med at

$$\mathbf{E}_\theta\left(\frac{\sum_i X_i}{n}\right) = \theta^2$$

så estimatoren $\tilde{\theta}$ ovenfor er bestemt ved at løse denne ligning m.h.t. θ; estimatoren er ikke central idet

$$\mathbf{E}_\theta(\tilde{\theta}) = \mathbf{E}_\theta(\sqrt{\bar{X}}) < \sqrt{\mathbf{E}_\theta(\bar{X})} = \theta.$$

ifølge Jensens ulighed.

Nedenfor er angivet et eksempel på en stikprøve som anført ovenfor med $n = 10$:

```
> x
 [1] 0.69 3.64 0.96 1.28 1.79 0.12 4.82 0.68 0.55 0.52
```

Værdien af maksimum likelihood estimatoren for θ baseret på den angivne stikprøve kan beregnes til $\hat{\theta} = 1.2278$.

6. Betragt nu hypotesen $H_0 : \theta = 1$, svarende til, at observationerne stammer fra en eksponentialfordeling; beregn en approksimativ p-værdi for likelihood ratio testet for den pågældende hypotese. Kunne observationerne være eksponentialfordelte?

Log-likelihoodfunktion i minimumspunktet beregnes til 13.96 og i punktet 1 fås værdien 15.05, så kvotienttesstørrelsen bliver

$$LR = 2(15.05 - 13.96) = 2.187$$

og baseret på χ^2 fordelingen med 1 frihedsgrad fås en p-værdi på 0.14; man kan altså ikke afvise at tallene stammer fra en eksponentialfordeling.

Opgave 4

Lad (X,Y) være normalfordelt på $\mathbb{R}^2$ med middelværdi 0 og varians Σ, altså $(X,Y) \sim N(0,\Sigma)$, hvor

$$\Sigma = \begin{pmatrix} \alpha & \frac{\alpha}{2} \\ \frac{\alpha}{2} & \alpha \end{pmatrix}.$$

Her er α en konstant der opfylder visse betingelser, se spørgsmål 1.

1. For hvilke værdier af α er Σ en lovlig variansmatrix? For hvilke værdier af α er fordelingen af X en regulær hhv. singulær normalfordeling?

   Σ er en lovlig variansmatrix hvis og kun hvis den er positiv semidefinit, dvs. hvis og kun hvis

   - $\Sigma_{11} \geq 0$, dvs. $\alpha \geq 0$
   - $\det(\Sigma) \geq 0$ dvs. $\frac{3}{4}\alpha^2 \geq 0$.

   Tilsammen får vi altså at Σ en lovlig variansmatrix hvis og kun hvis $\alpha \geq 0$.

   Fordelingen er regulær hvis og kun hvis Σ er invertibel. Da $\det(\Sigma) = \frac{3}{4}\alpha^2$, har vi altså at fordelingen er regulær for $\alpha > 0$ og singulær for $\alpha = 0$.

2. Bestem fordelingen af $\begin{pmatrix} X+Y \\ X-Y \end{pmatrix}$, og vis derved at $X+Y$ og $X-Y$ er uafhængige, $X+Y \sim N(0,3\alpha)$ og $X-Y \sim N(0,\alpha)$.

   Hvis vi definerer

   $$C = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

   og bruger transformationssætningen for normalfordelingen, får vi at

   $$\begin{pmatrix} X+Y \\ X-Y \end{pmatrix} = C \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0, C\Sigma C^T).$$

   Variansmatricen viser sig at være

   $$C\Sigma C^T = \begin{pmatrix} 3\alpha & 0 \\ 0 & \alpha \end{pmatrix}.$$

   Det følger umiddelbart at $X+Y$ og $X-Y$ er uafhængige da elementet udenfor diagonalen er nul, og marginalfordelingerne aflæses også direkte: $X+Y \sim N(0,3\alpha)$ henholdsvis $X-Y \sim N(0,\alpha)$.

I det følgende kan du uden bevis benytte omskrivningen

$$XY = \frac{1}{4}(X+Y)^2 - \frac{1}{4}(X-Y)^2$$

samt at $\mathbf{E}(Z^4) = 3\sigma^4$ hvis $Z \sim N(0,\sigma^2)$.

3. Definer estimatoren

   $$\hat{\alpha} = 2XY$$

   Vis at $\hat{\alpha}$ er en central estimator for α og bestem variansen $\mathbf{V}(\hat{\alpha})$. Gør desuden rede for at $P(\hat{\alpha} \geq 0) < 1$ hvis $\alpha > 0$.

Vi definerer estimatoren $\hat{\alpha} = 2XY$, der er central fordi

$$\mathbf{E}\hat{\alpha} = 2\mathbf{E}XY = 2\mathrm{Cov}(X,Y) = 2\frac{\alpha}{2} = \alpha$$

Ved hjælp af det første vink of uafhængigheden af $X+Y$ og $X-Y$, får vi variansen:

$$\mathbf{V}(XY) = \frac{1}{16}\mathbf{V}\Big((X+Y)^2\Big) + \frac{1}{16}\mathbf{V}\Big((X-Y)^2\Big)$$

Det andet vink giver at $\mathbf{V}(Z) = 3\sigma^4 - \sigma^4 = 2\sigma^4$ hvis $Z \sim N(0,\sigma^2)$, så

$$\mathbf{V}(XY) = \frac{1}{16}2(3\alpha)^2 + \frac{1}{16}2\alpha^2 = \frac{5}{4}\alpha^2$$

og dermed er

$$\mathbf{V}(\hat{\alpha}) = 4\mathbf{V}(XY) = 5\alpha^2.$$

Hvis $\alpha > 0$, så er fordelingen af (X,Y) regulær på $\mathbb{R}^2$, således at (X,Y) „lever på“ hele $\mathbb{R}^2$, og der er derfor positiv sandsynlighed for at havne i anden eller fjerde kvadrant hvor $XY < 0$. Altså er $P(\hat{\alpha} > 0) < 1$, og vi kan altså risikere at få et estimat udenfor parametermængden.

4. Definer estimatoren

$$\tilde{\alpha} = \frac{1}{2}\big(X^2 + Y^2\big)$$

Vis at $\tilde{\alpha}$ er en central estimator for α og at variansen er $\mathbf{V}(\tilde{\alpha}) = 5\alpha^2/4$. Diskuter kortfattet om du foretrækker $\tilde{\alpha}$ eller $\hat{\alpha}$ som estimator for α.

Vi definerer en ny estimator, $\tilde{\alpha} = \frac{1}{2}(X^2+Y^2)$. Estimatoren er central da

$$\mathbf{E}\tilde{\alpha} = \frac{1}{2}(\alpha+\alpha) = \alpha$$

Andetmomentet er

$$\begin{aligned}\mathbf{E}\tilde{\alpha}^2 &= \frac{1}{4}\mathbf{E}\Big(X^4 + Y^4 + 2X^2Y^2\Big)\\ &= \frac{1}{4}\Big(3\alpha^2 + 3\alpha^2 + 2\mathbf{E}(X^2Y^2)\Big)\\ &= \frac{9}{4}\alpha^2\end{aligned}$$

hvor det til sidst er benyttet at

$$\mathbf{E}(X^2Y^2) = \mathbf{V}(XY) + (\mathbf{E}XY)^2 = \frac{5}{4}\alpha^2 + \frac{1}{4}\alpha^2 = \frac{3}{2}\alpha^2$$

Således er, som ønsket,

$$\mathbf{V}\tilde{\alpha} = \frac{9}{4}\alpha^2 - \alpha^2 = \frac{5}{4}\alpha^2.$$

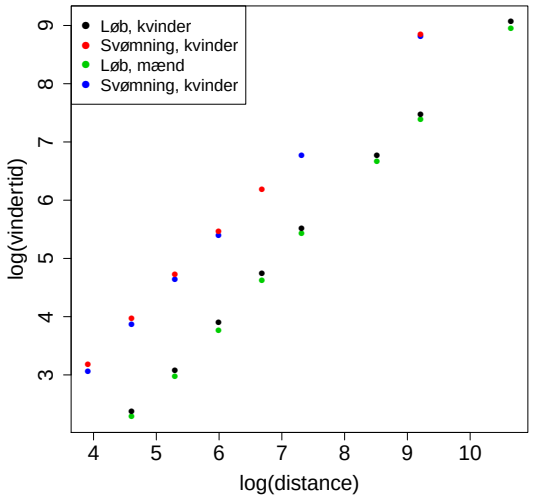
Estimatorerne $\hat{\alpha}$ og $\tilde{\alpha}$ er begge centrale, og $\tilde{\alpha}$ har mindre varians end $\hat{\alpha}$ (pånær i det uinteressante tilfælde hvor $\alpha = 0$). Desuden er $P(\tilde{\alpha} \geq 0) = 1$, så $\tilde{\alpha}$ rammer altid parameterområdet hvilket ikke gælder for $\hat{\alpha}$. Estimatoren $\tilde{\alpha}$ er af disse grunde klart at foretrække.

Opgave 5

Data til denne opgave består af vindertiderne i løbe- og svømmedisciplinerne ved OL i Rio de Janeiro, 2016. Data er tilgængelige i filen `rio2016.txt` på den vedlagte USB-nøgle. Der er følgende variable:

- `koen`: Køn, med værdien 0 for mænd og 1 for kvinder
- `type`: Typen af disciplin, med værdien 0 for løb og 1 for svømning
- `distance`: Distancen for disciplinen
- `vindertid`: Vindertiden i den pågældende disciplin, målt i sekunder

Figuren nedenfor viser data, hvor både distance og vindertid er log-transformeret og punkterne er farvet efter kombinationen af køn og typen af disciplin.



Du skal først betragte den multiple regressionsmodel hvor `log(vindertid)` benyttes som responsvariabel og `log(distance)`, `type` og `koen` benyttes som forklarende variable. Hvis data er indlæst i R som `rio2016`, så kan modellen fittes med kommandoen

```
reg <- lm(log(vindertid) ~ type + koen + log(distance), data=rio2016)
```

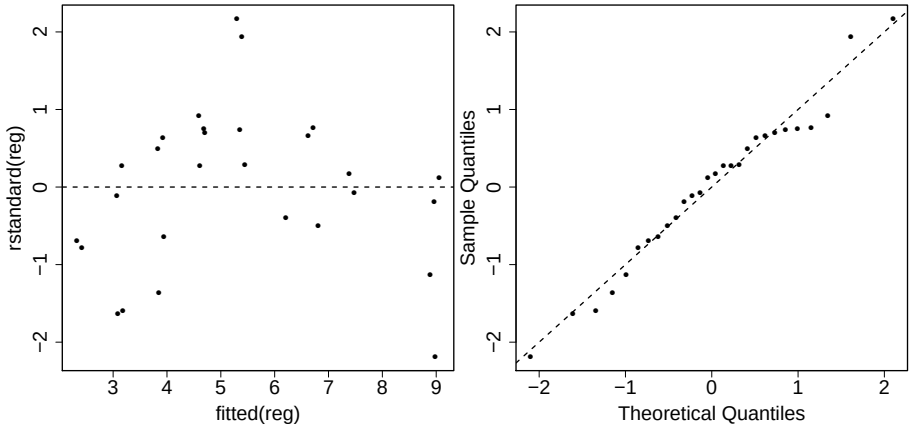
1. Opskriv den statistiske model svarende til `reg` (med papir og blyant). Fit modellen i R, og udfør modelkontrol. Svaret vedrørende modelkontrol skal indeholde skitser af relevante figurer og kommentarer til figurerne.

   Den statistiske model antager at observationen $x = (x_1, \ldots, x_{28})$ af log-vindertiderne er udfald af en stokastisk variabel $X \sim N(\xi, \sigma^2 I)$ hvor $\xi \in L$ og $\sigma^2 > 0$ er ukendte parametre ξ har formen

   $$\xi_i = \beta_0 + \beta_1 \cdot \mathbf{1}_{\text{type=svømning}} + \beta_2 \cdot \mathbf{1}_{\text{koen=kvinde}} + \beta_3 \cdot \log(\text{distance})$$

   hvor $\mathbf{1}$ er indikatorfunktioner.

   Residualplot og QQ-plot for de standardiserede residualer er vist nedenfor:



QQ-plottet ser fint ud eftersom punkterne ligger omkring 0/1-linien. Residualplottet ser derimod ikke helt godt ud: Der er en tendens til at punkterne udgør en „sur“ parabel svarende til at modellen overvurderer tiden på langsomme og hurtige discipliner og undervurderer tiden mellemhurtige discipliner.

Hvis man kigger nøjere efter, viser det sig at de to datapunkter der ligger lidt for sig selv, med store residualer, viser sig at svare til 1500 m løb for mænd og kvinder. Her er de observerede vindertider altså væsentligt større end forventet udfra modellen, hvilket indikerer at disse distancer er hårdere end de øvrige.

Uanset hvad du konkluderede vedrørende modelkontrol i spørgsmål 1, så skal du fortsætte med modellen i spørgsmål 2–4.

Husk desuden at både `distance` og `vindertid` indgår log-transformeret i modellen.

2. Kvinderne svømmer 800 m og mændene svømmer 1500 m, men ikke omvendt. Brug modellen til at bestemme et estimat for vindertiden for kvinder på 1500 m og et estimat for vindertiden for mænd på 800 m (hvis disse discipliner fandtes).

   Middelværdien af log-vindertiden for 1500 m svømning for kvinder er

   $$\xi' = \beta_0 + \beta_1 + \beta_2 + \beta_3 \cdot \log(1500).$$

   Hvis vi indsætter parameterestimaterne, så fås estimatet for ξ':

   $$\hat{\xi}' = -2.732 + 1.506 + 0.094 + 1.098 \cdot \log(1500) = 6.899$$

   Dette er på log-skala, så et estimat for vindertiden er $\exp(6.899) = 991.2$ sekunder. Dette svarer til 16 min, 31 sek.

   Middelværdien af log-vindertiden for 800 m svømning for mænd er

   $$\xi'' = \beta_0 + \beta_1 + \beta_3 \cdot \log(800).$$

   der estimeres til $\hat{\xi}'' = 6.114756$. Dette giver estimatet 452.5 sekunder, eller 7 min, 32 sek.

3. Betragt to distancer hvor den ene er dobbelt så lang som den anden. Gør rede for at modellen antager at den forventede forskel i `log(vindertid)` er den samme for alle fire kombinationer af køn og disciplintype, og bestem et estimat for den fælles forventede

forskel. Bestem derefter et estimat for den faktor, som vindertiden forøges med, når distancen fordobles (uanset køn og disciplintype).

Betragter distancerne d og $2d$. Uanset køn og disciplintype er forskellen i forventet log-vindertid mellem de to distancer lig

$$\delta = \beta_3 \log(2d) - \beta_3 \log(d) = \beta_3 \log(2).$$

Dette skyldes at de øvrige led i modellen enten optræder eller ikke optræder for begge distancer, og således udgår når der vi betragter forskellen.

Forskellen i forventet log-vindertid estimeres derfor til

$$\hat{\delta} = \hat{\beta}_3 \log(2) = 1.098 \cdot \log(2) = 0.761$$

Dette svarer til at vindertiden (ikke log-transformeret) øges med en faktor $2^{\hat{\beta}_2} = \exp(0.761) = 2.14$ når distancen fordobles.

4. Gør rede for at modellen antager at den forventede forskel i `log(vindertid)` mellem svømning og løb er den samme for alle distancer og begge køn, og bestem et estimat for den fælles forventede forskel. Bestem derefter et estimat for den faktor, som vindertiden er længere ved svømning end ved løb (uanset køn og distance).

   Forskellen i forventet log-vindertid mellem svømning og løb er, for alle distancer og begge køn, lig parameteren β_1. Dette skyldes at de øvrige led indgår i begge forventede værdier og dermed går ud når der tages differens.

   Forskellen i forventet log-vindertid estimeres derfor til $\hat{\beta}_1 = 1.506$, hvilket svarer til at vindertiden (ikke log-transformeret) er en faktor $\exp(\hat{\beta}_1) = \exp(1.506) = 4.51$ større for svømning end for løb.