UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1000 — Innføring i anvendt statistikk
Eksamensdag:	Mandag 05. desember 2022
Tid for eksamen:	09:00 – 13:00
Oppgavesettet er på 6 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Alle

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Oppgave 1 Fjortiser

1a

Høyden til 14 år gamle jenter målt i cm på fødselsdagen er normalfordelt $X \sim N(\mu = 163{,}40, \sigma = 6{,}25)$.

Høyden til 14 år gamle gutter målt i cm på fødselsdagen er normalfordelt $Y \sim N(\mu = 166{,}90, \sigma = 7{,}75)$.

$$P(X > 165) = P(Z > \frac{165 - 163{,}40}{6{,}25}) = 1 - P(Z < 0{,}256) = 0{,}399, \quad (1)$$

der Z er en standard normalfordelt variabel.

i) Sannsynligheten for at en tilfeldig valgt 14 år gammel jente er høyere enn 165cm er 0,399.

$$P(Y > 165) = P(Z > \frac{165 - 166{,}90}{7{,}75}) = 1 - P(Z < -0{,}245) = 0{,}597, \quad (2)$$

der Z er en standard normalfordelt variabel.

ii) Sannsynligheten for at en tilfeldig valgt 14 år gammel gutt er høyere enn 165cm er 0,597.

1b

$$\mu_{X-Y} = \mu_X - \mu_Y = 163{,}40 - 166{,}90 = -3{,}50 \quad (3)$$

Høyden til en en tilfeldig valgt 14 år gammel jente og en tilfeldig valgt 14 år gammel gutt er uavhengige.

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y = 6{,}25^2 + 7{,}75^2 = 99{,}125 = 9{,}956154^2 \quad (4)$$

(Fortsettes på side 2.)

i) Høydeforskjellen $X - Y$ mellom en tilfeldig valgt 14 år gammel jente og en tilfeldig valgt 14 år gammel gutt (høyde målt i cm på 14-årsdagen) er normalfordelt $X - Y \sim N(\mu = -3{,}50, \sigma = 9{,}956154)$

$$P(X-Y>0) = P(Z > \frac{0-(-3{,}50)}{9{,}956154}) = 1 - P(Z < 0{,}3515414) = 0{,}363 \quad (5)$$

ii) Sannsynligheten for at en tilfeldig valgt 14 år gammel jente er høyere enn en tilfeldig valgt 14 år gammel gutt er 0.363.

Oppgave 2 Ukjent forventningsverdi

2a

Vi er gitt n uavhengige og identisk fordelte observasjoner, $X_1, X_2, \cdots, X_n$, som er normalfordelte med kjent standardavvik σ. Tetthetsfunksjonen til normalfordelinga $N(\mu, \sigma)$ er

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right). \quad (6)$$

Likelihooden til observasjonene er

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(X_i-\mu)^2}{2\sigma^2}\right). \quad (7)$$

Log-likelihooden til observasjonene er

$$\log(L) = -n \cdot \log\left(\sqrt{2\pi\sigma^2}\right) - \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{2\sigma^2}. \quad (8)$$

Den deriverte av log-likelihooden med hensyn på μ, er

$$\frac{d(\log L)}{d\mu} = 0 - \sum_{i=1}^{n} 2 \cdot \frac{(X_i-\mu)}{2\sigma^2} \cdot (-1) = n \cdot \frac{\bar{X}-\mu}{\sigma^2}, \quad (9)$$

der gjennomsnittet til observasjonene $\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$.

Den deriverte av log-likelihooden til observasjonene har verdi 0 nettopp når $\mu = \bar{X}$. Vi har derfor at Maximum Likelihood-estimatoren $\hat{\mu}$ for verdien av parameteren μ (forventningsverdien i normalfordelinga) er gjennomsnittsverdien av observasjonene, $\hat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$.

2b

Hvert måleresultat er normalfordelt $X_i \sim N(\mu, \sigma = 0{,}69)$. Gjennomsnittet av tre måleresultat er normalfordelt $\bar{X} \sim N(\mu, \sigma_{\bar{X}} = \frac{0{,}69)}{\sqrt{3}}$.

Observert gjennomsnitt av tre måleresultat er $\bar{X} = (119{,}30 + 118{,}63 + 119{,}01)/3 = 118{,}98$.

(Fortsettes på side 3.)

i) Et 95% konfidensintervall for den sanne lengden til maskindelen μ er

$$\bar{X} \pm z_{0,975} * \sigma_{\bar{X}} = 118{,}98 \pm 1{,}96 * \frac{0{,}69)}{\sqrt{3}} = [118{,}20,\ 119{,}76]. \qquad (10)$$

ii) Konfidensintervallet presenterer et intervall av mulige verdier for μ i samsvar med dataene, og er konstruert med en metode som i 95% av tilfellene den blir brukt vil gi et intervall som inneholder den sanne parameteren μ.

Oppgave 3 Bjørn

3a

En enkel lineær regresjonsmodell for responsvariabel log-vekt (y_i) og forklaringsvariabelen log-lengde (x_i), er

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i, \quad i = 1, 2, \cdots, 223 \qquad (11)$$

der individuell variasjon $\epsilon_i \sim N(0, \sigma)$ er uavhengig og normalfordelt for hvert av individene.

Modellantagelsene er dermed:

1) Linearitetsantagelsen: Forventa log-vekt $\mu_{y_i} = \beta_0 + \beta_1 \cdot x_i$ er **lineært** avhengig av verdien på forklaringsvariabelen log-lengde x_i.

2) Antagelse om konstant varians: Spredninga i log-vekt y_i er gitt ved standardavvik σ i underpopulasjonen der log-lengde har verdi x_i, for enhver verdi av x_i; med andre ord, **spredninga** om forventningsverdien **varierer ikke med forklaringsvariabelen** x_i.

3) Uavhengighetsantagelsen: Leddene for individuell variasjon, ϵ_i, $i = 1, 2, \cdots, 223$ er **uavhengige**. Ekvivalent: gitt verdiene til log-lengde-målene x_i for hvert individ, er verdiene for log-vekt y_i uavhengige, og

4) Normalantagelsen: Leddene for individuell variasjon, ϵ_i, $i = 1, 2, \cdots, 223$ er **normalfordelte**. Ekvivalent: Gitt verdien til log-lengde-målet x_i for individ i, er log-vekta y_i normalfordelt.

3b

i) Residualene er differansen mellom observert og predikert verdi for dataene vi bruker til å estimere lineærmodellen; for individ i er residualet $e_i = y_i - \widehat{y}_i = y_i - (b_0 + b_1 \cdot x_i)$.

   – For å evaluere normalantagelsen, undersøker vi om residualene ser ut til å kunne komme fra en normalfordeling. QQ-plottet viser residualene tilnærmet på en rett linje, som angir at de empiriske kvantilene i residualene har en tilnærma lineær sammenheng med de tilsvarende kvantilene i en standard normalfordeling. Det ser ut til at residualene kan regnes som normalfordelte Ingen bekymringer

(Fortsettes på side 4.)

- For å evaluere linearitetsantagelsen. ser vi etter mønstre i residualene mot forklaringsvariabelen. De ser fint spredt ut om 0, uavhengig av verdien på forklaringsvariabelen.Man kan ane en tendens til at residualene for de minste verdiene av forklaringsvariabelen er positive, men det kan enkelt tilskrives tilfeldigheter da det er tynt med data i dette området. Ingen bekymringer
- For å evaluere antagelsen om konstant varians, ser vi etter tendenser til at spredninga til residualene er ulik for ulike verdier på forklaringsvariabelen. Ingen bekymringer.
- Vi har ikke grunnlag til å vurdere uavhengighetsantagelsen ut fra plottene.

ii) $R^2 = 0.8751$. Empirisk korrelasjon r mellom målingene av log-vekt og log-lengde er $r = \sqrt{R^2} = \sqrt{(0.8751)} = 0.94$.

3c

i)
- Parameteren β_0 er konstantleddet i lineærmodellen, og beskriver forventa log-vekt for et individ med log-lengde $x_i = 0$.
- Parameteren β_1 er stigningstallet i lineærmodellen, og beskriver forventa antall enheter økning i log-vekt y_i når log-lengde øker med én enhet.
- Parameteren σ beskriver spredninga i log-vekt y_i om forventningsverdien μ_{y_i}, der μ_{yi} bestemmes av forklaringsvariabelen x_i.

ii) Estimatet til parameteren β_0 er $b_0 = -13{,}37$. Estimatet til parameteren β_1 er $b_1 = 3{,}51$. Estimatet til parameteren σ er $s = 0{,}21$.

ii) Nullhypotesen i den statistiske signifikanstesten presentert for β_1 i R-utskriften er $H_0 : \beta_1 = 0$, og alternativhypotesen er $H_a : \beta_1 \neq 0$.

iv) Et 95% konfidensintervall for β_1 er gitt ved $b_1 \pm t^* SE_{b_1}$, der t^* er 97,5%-persentilen til t-fordelinga med 141 frihetsgrader.

Fra tabell D finner vi 97,5%-persentilen til t-fordelinga med 100 frihetsgrader: $t^* = 1{,}984$ (konservativt valg, da neste alternativ er for 1000 frihetsgrader).

Nedre grense $b_1 - t^* SE_{b_1} = 3{,}5084 - 1{,}984 * 0{,}1116 = 3{,}286986$, og øvre grense $b_1 + t^* SE_{b_1} = 3{,}5084 + 1{,}984 * 0{,}1116 = 3{,}729814$.

Et 95% konfidensintervall for β_1 er dermed [3,29, 3,73].

3d

i) Et 95% konfidensintervall for forventa log-vekt for en bjørn med lengde 155cm er [4,29, 4,36]

(Fortsettes på side 5.)

ii) Et 95% prediksjonsintervall for log-vekt for en 155cm lang bjørn er [3,90, 4,75]

iii) Senter/midtpunktet i 95% konfidensintervall for forventa log-vekt er 4.33. Senter/midtpunktet i 95% prediksjonsintervall for log-vekt er.også 4.33. Prediksjonsintervallet inkluderer også variabiliteten i en fremtidig observasjon om forventningsverdien i underpopulasjonen av bjørner med lengde 155cm, og prediksjonsintervallet for log-vekt blir derfor bredere enn konfidensintervallet for forventningsverdien..

iv) exp(4.29263) = 73.15862, og exp(4.363616) = 78.54062. Intervallet av verdier for vekt (målt i kg) som tilsvarer verdiene for log-vekt i 95% konfidensintervallet for forventa log-vekt for en bjørn med lengde 155cm er [73,16, 78,54].

v) exp(3.902754) = 49.53869, og exp(4.753493) = 115.9887. Intervallet av verdier for vekt (målt i kg) som tilsvarer verdiene for log-vekt i 95% prediksjonsintervallet for log-vekt for en 155cm lang bjørn er [49,54, 115,99]

Oppgave 4 Suksess!

4a

En logistisk regresjonsmodell for sammenhengen mellom responsvariabelen 'kobberfunn' (y) og forklaringsvariablene $x_{i1}, x_{i2}, x_{i3}, x_{i4}$, er

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3} + \beta_4 \cdot x_{i4}, \qquad (12)$$

der $p_i = P(y_i = 1|x_{i1}, x_{i2}, x_{i3}, x_{i4})$ er sannsynligheten for å finne kobber i område i med kjente verdier $x_{i1}, x_{i2}, x_{i3}, x_{i4}$ av forklaringsvariablene.

Videre, for hver j = 1,2,3,4, er β_j er den naturlige logaritmen av odds-ratioen for å finne kobber assosiert med en enhets økning i x_j.

4b

i) Odds er forholdet mellom suksess-sannsynlighet p og sannsynligheten for ikke-suksess/feil, altså $\frac{p}{1-p}$.

Videre er odds ratio forholdet mellom to ulike odds.

For to ulike individer med suksess-sannsynlighet henholdsvis p_1 for individ 1 og p_2 for individ 1, er odds ratio $\frac{p_1/1-(p_1)}{p_2/1-(p_2)}$ for individ 1 sammenligna med individ 2.

ii) exp(b_1) = exp(0,9032) = 2.467486..Et estimat for odds-ratioen (OR) for kobberfunn assosiert med en enhets økning i magnetisk intensitet x_{i1}, når de andre forklaringsvariablene holdes fast, er 2.47.

(Fortsettes på side 6.)

iii) Et 95% kondensintervall for parameteren β_1 er gitt ved formelen

$$b_1 \pm z_{0,975} \cdot SE_{b_1} \tag{13}$$

der $z_{0,975} = 1{,}96$ er 97,5% persentilen til en standard normalfordelt variabel. Et 95% kondensintervall for parameteren β_1 er derfor

$$[0{,}9032 - 1{,}96 \cdot 0{,}1701,\ 0{,}9032 + 1{,}96 \cdot 0{,}1701] = [0{,}569804,\ 1{,}236596]. \tag{14}$$

$\exp(0{,}569804) = 1.767921$, og $\exp(1{,}236596) = 3.443871$.

Et 95% kondensintervall for odds-ratioen (OR) for kobberfunn assosiert med en enhets økning i magnetisk intensitet x_{i1}, når de andre forklaringsvariablene holdes fast er [1,77, 3.44].