Løsningsforslag

Oppgavesettet og løsningsforslaget er utarbeidet av Jarle Møen. Det anbefales at sensorene gir inntil 10 poeng per delspørsmål.

Oppgave 1

a) Når fars høyde øker med én tomme, øker forventet høyde på barnet med 0,398 tommer. Når mors høyde øker med én tomme, øker forventet høyde på barnet med 0,321 tommer. Sønner forventes å bli 5,21 tommer høyere enn døtre, og forventet høyde på barna faller med 0,0438 tommer for hvert ekstra barn det er i søskenflokken. De tre første effektene er signifikante på 5 %-nivå, mens den siste effekten ikke er signifikant forskjellig fra null. Modellen har god forklaringskraft med en justert R^2 på 63,9%.

b) Den teoretiske motivasjonen for å inkludere antall barn som en forklaringsvariabel tilsier at vi kan utelukke at den sanne koeffisienten er positiv. Det kan forsvare å bruke ensidig test. Kritisk grense for ensidig T-test med 893 frihetsgrader er 1,645. Testobservatoren er T=-0,0438/0,0272=1,61. Vi kan ikke forkaste en nullhypotese om ingen effekt av antall barn med ensidig test og fem prosent signifikansnivå, men vi er ganske nær forkastning. (En god begrunnelse for tosidig test vil også gi poeng.)

c) I modell 1, 3 og 4 er den systematiske forskjellen i høyde mellom gutter og jenter tatt ut. I modell 2, 5 og 6 ligger kjønnsvariasjonene i høyde inne som en del av den variasjonen som skal forklares. Det er derfor mer variasjon som skal forklares i disse modellene, og siden denne ekstra variasjonen kan forklares med den inkluderte kjønnsdummien blir det en større andel av totalvariasjonen som kan forklares.

d) Modell 1: 18,8+0,729*(78,5+1,08*67)/2=73,79
Modell 2: 15,3+0,406*78,5+0,322*67+5,23=73,98
(Vi ser fra figur 1 at sønnens virkelige høyde var 73,2 så begge modellene predikerte ganske godt i dette tilfellet. Merk også at modellene egentlig er svært like til tross for ulik forklaringskraft. Det er først og fremst en teknikalitet knyttet til hvordan høydeforskjellen mellom kvinner og menn er behandlet som skiller modellene.)

e) T=(0,406-0,365)/0.0292=1,40. Kritisk verdi for tosidig T-test med 893 frihetsgrader er 1,96. Følgelig kan vi ikke forkaste en nullhypotese om at koeffisient er 0,365.

f) $Height=\alpha+\beta_1*Male+\beta_2*Father+\beta_3*Male*Father+\beta_4*Mother+\beta_5*Male*Mother+\varepsilon$.

g) Kjønnsdummien er den eneste variabelen som varierer innenfor familien og som dermed ikke absorberes av den familiespesifikke effekten. (Å estimere modellen med en familiespesifikk fast effekt tilsvarer å inkludere en dummyvariabel for hver familie.)

Spesielt interesserte kan lese en interessant biografi om Francis Galton på den engelske wikipedia og finne mer stoff om Galtons rolle i utviklingen av korrelasjons og regresjonsanalyse i Jeffrey M. Stanton : Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors i *Journal of Statistics Education*, Volume 9, Number 3, 2001. Artikkelen er tilgjengelig her: http://www.amstat.org/publications/jse/v9n3/stanton.html. Se også http://galton.org/.

Oppgave 2

a) Vi bruker T-test for to uavhengige utvalg med lik varians. Testobservatoren er $T=\frac{64{,}110-63{,}984}{0{,}0203}=0{,}620$ der nevneren er beregnet som $\sqrt{\frac{432\cdot 5{,}618+196\cdot 5{,}549}{433+197-2}\cdot\left(\frac{1}{433}+\frac{1}{197}\right)}=\sqrt{0{,}0413}=0{,}203$. Kritisk grense for tosidig T-test med 628 frihetsgrader er 1,96. Følgelig er vi ikke i nærheten av å kunne forkaste nullhypotesen om lik forventning for mødre og døtre. Vi kan heller ikke forkaste nullhypotesen om vi velger å bruke ensidig t-test (ut fra en tanke om at sann differanse ikke kan være negativ).

b) Vi antar uavhengighet mellom mors og fars høyde, jfr. spørsmål d. (Det burde vært presisert i oppgaven.
Var(*Midparent*)= Var(0,5**Father*+0,5*1,08**Mother*)=0,25*Var(*Father*)+ 0,2916*Var(*Mother*)=0,5416σ^2. Da har vi at 0,5416σ^2=3,30 og dermed at σ^2=3,30/0,5416= 6,093.
(Alternativt kunne vi regnet ut kovariansen ut fra korrelasjonskoeffisienten mellom mors og fars høyde og brukt formelen som åpner for korrlasjon).

c) Vi bruker F-test for lik varians. F=6,875/5,549=1,239. Kritisk verdi for tosidig F-test på 5 %-nivå med 196 frihetsgrader i teller og nevner er 1,32. Vi kan dermed ikke forkaste en hypotese om lik varians for kvinner og menn.

d) Testobservatoren er $T=r\cdot\sqrt{\frac{n-2}{1-r^2}}=0{,}101\sqrt{\frac{197-2}{1-0{,}101^2}}=1{,}418$. Kritisk grense for tosidig T-test på 5 %-nivå med 195 frihetsgrader er 1,97. Følgelig kan vi ikke forkaste en nullhypotese om ingen korrelasjon. Positiv korrelasjon kan blant annet skyldes at vi foretrekker partnere som ikke skiller seg for mye fra oss selv i høyde, men korrelasjonen er ikke sterk nok til å være signifikant. Det er den heller ikke om vi er villige til å utelukke at sann korrelasjon er negativ og dermed bruke ensidig test som har kritisk grense 1,65.

e) Vi har at $\hat{\beta}=\hat{\rho}\frac{SE(fars\_høyde)}{SE(mors\_høyde)}=0{,}101\cdot\frac{\sqrt{6{,}875}}{\sqrt{5{,}549}}=0{,}112$.

Oppgave 3

a) Vi ser at både båndbredde og rom er signifikante faktorer. (Båndbredde kun på 10 %-nivå.) I tillegg ser vi at interaksjonsleddet er signifikant slik man vil forvente dersom 2,4 GHz-båndet har dårligere hastighet når avstanden er kort, men bedre rekkevidde. Fra den deskriptive delen av analysen ser vi at gjennomsnittlig nedlastingshastighet er bedre med 2,4 GHz, men at dette i sin helhet skyldes at 5 GHz-båndet virker svært dårlig i kjelleretasjen (rom 4). I de øvrige rommene har 5 GHz-båndet litt bedre hastighet.

b) I analyse 2 er kjelleretasjen (rom 4) utelatt. Vi ser da at det ikke er signifikante forskjeller verken mellom de øvrige rommene eller mellom 2,4 GHz-båndet og 5GHz-båndet. Ut fra «teorien» er dette litt overraskende, men vi har få målinger og høy varians.

c) ANOVA forutsetter uavhengige, normalfordelte observasjoner og lik varians på tvers av gruppene. Vi ser at den observerte variansen varierer sterkt mellom de fire rommene. Det er derfor grunn til å tro at forutsetningen om lik varians ikke er oppfylt. Man kan også problematisere om uavhengighetsantagelsen er oppfylt når målingene skjer i rask rekkefølge siden tilfeldige forstyrrelser kan tenkes å berøre mer enn én måling.