Løsningsforslag/sensorveiledning – SFB12003 Metodekurs II HiØ 2016H.

Oppgave 1 (10%)

a) Undersøkelseseffekt innebærer at resultatene påvirkes av metoden (måten undersøkelsen gjennomføres på). Måtene å samle inn data på ved kvantitative undersøkelser er spørreskjema via post, web-basert spørreskjema, standardisert telefonintervju og personlig standardisert intervju. Her kan f.eks. intervjueffekt og/eller konteksteffekt nevnes og i hvilken grad disse kan påvirke resultatene ved de ulike datainnsamlingsmetodene

b) Oppdragsforskning innebærer at forskningen skjer på bakgrunn av en forespørsel fra noen. Etiske utfordringer ved at oppdragsgiver kan forsøke å unngå å publisere resultater som ikke passer med oppdragsgivers ønsker er relevant å trekke inn her.

Oppgave 2 (20%)

a) Enheter i studien er et representativt utvalg av individer fra Sverige og USA

b) Avhengig variabel: Er folk flest til å stole på? Med verdiene Ja og Nei
Uavhengig variabel: Land med verdiene Sverige og USA
Uavhengig variabel: Kjønn med verdiene Mann og Kvinne

c) «Folk flest er til å stole på»: Ordinal, fordi Ja-svar er mer positive til folk flest enn de som svarer nei (kan rangeres) eller nominalnivå fordi det er vanskelig å rangere disse verdiene og gi en verdi høyere verdi enn en annen.
Land: Nominal
Kjønn: Nominal

d) Operasjonalisering innebærer å gjøre et fenomen målbart. Jo mer abstrakt et begrep er, jo flere spørsmål er ofte nødvendig for å måle fenomenet. Hvor godt eller relevant spørsmålet er for å fange fenomenet. Når det gjelder om dette spørsmålet er egnet for å gi svar på om amerikanere eller svensker er mest til å stole på, så kan det diskuteres. På den ene siden kan man si at «på seg selv kjenner man andre» og dermed kan det være et argument for at denne studien viser at svenskene er mer til å stole på. På den andre siden kan man si at dette ikke er et valid mål på andel personer man kan stole på i Sverige eller USA, for det er veldig uklart hva som menes med «folk flest» og i et stort land som USA er det også mer kriminalitet, som igjen er mye i mediene og som igjen gjør at mange føler seg utrygge. Kanskje er det heller hvor utrygge amerikanere føler seg i forhold til svensker? Eller noe helt annet. (Oppgaver som definerer operasjonalisering greit og argumenterer fornuftig for enten ja eller nei eller både ja og nei bør få full pott. Ikke helt klare fasitsvar her, men refleksjoner som viser at de forstått hva operasjonalisering handler om er det vi er ute etter her.)

Oppgave 3 (35%)

a)

$$b_2 = \frac{s_{XY}}{s_X^2} = \frac{\sum_{i=1}^{11}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{11}(X_i - \bar{X})^2} = \frac{-0{,}611}{1{,}274} = -0{,}4796$$

Dersom prisen for ett pund kaffe øker med en dollar, synker konsumet per person per dag med 0,4796 kopper eller ca. en halv kopp i gjennomsnitt.

$$b_1 = \bar{Y} - b_2\bar{X} = 2{,}206 - (-0{,}4796) \cdot 1{,}011 = 2{,}691$$

Dersom prisen på kaffe er null (gratis kaffe) vil konsumet være 2,69 kopper per dag per person i gjennomsnitt.

b)

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{11}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{11}(Y_i - \bar{Y})^2} = \frac{0{,}293}{0{,}442} = 0{,}663$$

66,3% av variasjonen i Y forklares av modellen

c)

Hypoteser: $H_0: B_2 = 0$ og $H_A: B_2 \neq 0$.

Kritiske verdier: $\pm t_{\frac{\alpha}{2}}(n-k) = \pm t_{\frac{0{,}01}{2}}(11-2) = \pm t_{0{,}005}(9) = \pm 3{,}250$

Testverdi:

$$\frac{b - H_0 verdi}{se(b)} = \frac{b_2}{se(b_2)}$$

Vi trenger altså beregne $se(b_2)$:

$$se(b_2) = \sqrt{var(b_2)}$$

$$var(b_j) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2} \cdot \frac{1}{1 - R_j^2}$$

$$var(b_2) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \cdot \frac{1}{1 - 0}$$

$$se(b_2) = \sqrt{var(b_2)} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Der

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}^2}{n-k}} = \sqrt{\frac{\sum_{i=1}^{11}(Y_i - \hat{Y}_i)^2}{n-k}} = \sqrt{\frac{0{,}149}{11-2}} = \sqrt{\frac{0{,}149}{9}} = \sqrt{0{,}0166} = 0{,}129$$

$$\hat{\sigma}^2 = 0{,}129^2 = 0{,}0166$$

$$se(b_2) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \sqrt{\frac{0{,}0166}{1{,}274}} = \sqrt{0{,}013} = 0{,}114$$

Testverdi:

$$\frac{b_2}{se(b_2)} = -\frac{0{,}4796}{0{,}114} = -4{,}207$$

Testverdien ligger i forkastningsområdet (større enn 3,250 eller mindre enn -3,250), så vi forkaster nullhypotesen. Vi har funnet støtte for alternativhypotesen om at prisen påvirker konsumet.

d) Høyere standardfeil vil si at den beregnede koeffisienten er mindre nøyaktig. Dette fører til en lavere testverdi som gjør at det skal mer til (høyere beregnet verdi for koeffisienten) for at vi forkaster nullhypotesen. Kritisk verdi påvirkes ikke av standardfeilen til den beregnede koeffisienten.

e) Dersom prisen øker med 1% reduseres konsumet med 0,25% (kan også ha med at dette er elastisiteten, men det er ikke et krav til å få riktig). $H_0: B_2 = -1$ og $H_A: B_2 \neq -1$. Testverdi blir:

$$\frac{b_2 - H_0 verdi}{se(b_2)} = \frac{-0{,}253 - (-1)}{0{,}0494} = \frac{-0{,}253 + 1}{0{,}0494} = \frac{0{,}747}{0{,}0494} = 15{,}121$$

Kritiske verdier: $\pm t_{0,005}(9) = 3{,}250$. Vi forkaster nullhypotesen da testverdi ligger i forkastningsområdet. Vi har funnet støtte for alternativhypotesen om at populasjonsparameteren er -1 (som vil si at kaffe er nøytralelastisk – ikke krav å ha med dette).

Oppgave 4

a) $b_1 = 35{,}78$: En bolig med størrelse 0 kvm, ingen soverom, ingen peis og ingen svømmebasseng vil ha en pris på $35782. Evt. kan dette tolkes som konstantleddet til regresjonsmodellen som kun har til hensikt å plassere regresjonslinjen for å minimere de kvadrerte avvikene.

$b_2 = 1{,}548$: Ett ekstra rom gir en prisøkning på i gjennomsnitt $1548, alt annet likt (dersom vi kontrollerer for antall soverom og hvorvidt boligen har peis og svømmebasseng).

$b_3 = -4{,}773$: Ett ekstra soverom gir i gjennomsnitt en prisreduksjon på $4773, alt annet likt.

$b_4 = 6{,}904$: En bolig med peis har i gjennomsnitt en pris som er $6904 høyere, alt annet likt.

$b_5 = 51{,}511$: En bolig med svømmebasseng har i gjennomsnitt en pris som er $51511 høyere, alt annet likt.

b) $H_0: B_5 = 0$ og $H_A: B_5 \neq 0$. P-verdien er 0,0273 og lavere enn signifikansnivået 0,10 som vil si at vi forkaster nullhypotesen. Vi har funnet støtte for at det er en forskjell i salgsprisen for boliger med og uten svømmebasseng. Alternativt kan vi bruke kritiske verdier $\pm\, t_{\frac{\alpha}{2}}(n-k) = \pm t_{0,05}(14-5) = \pm t_{0,05}(9) = \pm 1{,}833$. Vi ser at testobservator «t-ratio» 2,6305 er i forkastningsområdet, så vi forkaster nullhypotesen.

c) Heteroskedastisitet innebærer at modellen har ulik presisjon for ulike verdier av en forklaringsvariabel og er et brudd på en av forutsetningene ved minste kvadraters metode. Dette kan vi finne ved å se om det kvadrerte avviket avhenger av en av forklaringsvariablene. I modellen $\hat{u}_i^2 = B_1 + B_2 kvm_i + B_3 bedrms_i + B_4 firepl_i + B_5 pool_i + v_i$ kan vi teste $H_0: B_2 = B_3 = B_4 = B_5 = 0$ mot H_A: En eller flere av påstandene i H_0 er gale.
Testobservator/testuttrykk beregnes ved

$$F = \frac{(R_{ur}^2 - R_r^2)/m}{(1 - R_{ur}^2)/(n-k)} = \frac{(0{,}282 - 0)/4}{(1 - 0{,}282)/(14-5)} = \frac{0{,}0705}{0{,}798} = 0{,}0883$$

Kritisk verdi: $F_\alpha(Df_1, Df_2) = F_{0,1}(4{,}9) = 2{,}693.$

Vi beholder nullhypotesen da testobservator ikke er i forkastningsområdet. Feilleddet til den beregnede modellen (modell 1) er ikke heteroskedastisk.

d) Bolig A: 100 kvm, tre soverom, ingen peis, svømmebasseng.
Bolig B: 100 kvm, tre soverom, ingen peis, ikke svømmebasseng.
Differanse i anslag på salgspris for bolig A $\hat{Y}'$ og bolig B $\hat{Y}$:

$$\begin{aligned}
\hat{Y}' - \hat{Y} &= (b_1 + b_2 kvm'_i + b_3 bedrms'_i + b_4 firepl'_i + b_5 pool'_i + b_6 pool'_i \cdot kvm'_i) \\
&\quad -(b_1 + b_2 kvm_i + b_3 bedrms_i + b_4 firepl_i + b_5 pool_i + b_6 pool_i \cdot kvm_i) \\
&= (b_1 + b_2 100 + b_3 3 + b_4 0 + b_5 1 + b_6 \cdot 1 \cdot 100) \\
&\quad -(b_1 + b_2 100 + b_3 3 + b_4 0 + b_5 0 + b_6 \cdot 0 \cdot 100) \\
&= b_5(1-0) + b_6(100-0) = b_5 - 100 b_6 = -86{,}1 - 100 \cdot 0{,}803 = -3{,}1
\end{aligned}$$

Bolig C: 200 kvm, tre soverom, ingen peis, svømmebasseng.
Bolig D: 200 kvm, tre soverom, ingen peis, ikke svømmebasseng.
Differanse i anslag på salgspris for bolig C $\hat{Y}'$ og bolig D $\hat{Y}$:

$$\hat{Y}' - \hat{Y} = b_5 - 200b_6 = -86{,}1 - 200 \cdot 0{,}803 = 79{,}9$$

Vi ser at i denne modellen har hvorvidt boligen har svømmebasseng eller ei ulik effekt avhengig av størrelsen på boligen. Dette kalles en samspillseffekt (interaction effect) og gir ulikt stigningstall for hvordan størrelsen på boligen påvirker prisen.

e) Forslag til ny modell:
$price_i = B_1 + B_2 kvm_i + B_3 bedrms_i + B_4 firepl_i + B_5 pool_i + B_6 kvm_i^2 + B_7 kvm_i^3 + u_i$
altså et polynom. Alternativt kan man bruke logaritmer eller inverse funksjoner. Hvorvidt det er en ikke-lineær effekt kan her testes ved en F-test der $H_0: B_6 = B_7 = 0$.

f) 1-kuttsmodellen innebærer å fjerne alle variablene som ikke har en signifikant effekt. Dette kan vi gjøre ved å se på p-verdien til koeffisientene. For et 1% signifikansnivå er kun kvm signifikant, noe som innebærer at modellen da blir:

$$price_i = B_1 + B_2 kvm_i + u_i$$

På et 10% signifikansnivå er også dummyvariabelen for svømmebasseng signifikant, noe som gir modellen:

$$price_i = B_1 + B_2 kvm_i + B_5 pool_i + u_i$$

Vi kan også velge å fjerne konstantleddet da det heller ikke er signifikant.
Modellene vi ender opp med har ikke med de ikke-relevante variablene og vi har med dette forsøkt å fjerne/redusere inkluderingsproblemet med Modell 1.

Veldig kort:

1. ..
2. …
3. ..
   a. $b_1 = 2{,}691$ $b_2 = -0{,}4796$
   b. 66,3%
   c. $\frac{b_2}{se(b_2)} = -\frac{0{,}4796}{0{,}114} = -4{,}207$, $\pm t_{\frac{\alpha}{2}}(n-k) = \pm t_{\frac{0{,}01}{2}}(11-2) = \pm t_{0{,}005}(9) = \pm 3{,}250$
4. ..
   a. ..
   b. P-verdi<sign.nivå: Forkast H0
   c. Brudd på forutsetning ved at feilledd avh. av nivå på forklaringsvariabel.
   $F = 0{,}0883$ Kritisk verdi: $F_\alpha(Df_1, Df_2) = F_{0{,}1}(4{,}9) = 2{,}693.$
   d.