SOS1120 KVANTITATIV METODE – EKSAMEN HØST 2011 - Sensorveiledning

OPPGAVE 1

1-1: Figur 1 inkluderer 6 variable: 5 variable måler respondentenes bruk av henholdsvis kino, idrettsarrangement, folkebibliotek, museum og teater/musikal/revy. Den siste variabelen er historisk tid.

1-2: Teksten forteller oss at figuren er basert på prosentandel som har svart ja til at de har gått på kino, idrettsarrangement, folkebibliotek, museum og teater/musikal/revy de siste 12 månedene. Hver av disse variablene har derfor kun to verdier slik de er fremstilt her: ja og ikke-ja (den siste verdien kan inneholde alle som svarte nei, samt de som ikke husker, og de som eventuelt svarer at de vet ikke). Dersom noen skriver at vi ikke vet hvor mange verdier denne variabelen har, fordi det kun er oppgitt i teksten at figuren er basert på andelen som har svart ja til bruk av hvert kulturtilbud, så er det også greit. Uansett er disse fem variablene på nominalnivå.

Historisk tid er i utgangspunktet en kontinuerlig variabel på høyeste målenivå, dvs forholdstallsnivå. Variabelen måler fem ulike tidspunkt med faste avstander mellom verdiene, bortsett fra siste intervall hvor avstanden mellom tidspunktene er 1 år lengre. Noen vil kanskje derfor argumentere med at variabelen er på ordinalnivå. Men det riktige svaret er forholdstallsnivå.

1-3: Figuren viser hvor stor andel av respondentene i de fem utvalgene som deltok på ulike kulturarrangement i 1991, 1994, 1997, 2000 og 2004. Tabellen omfatter respondenter i alderen 9-79 år. En god besvarelse bør kommentere både nivå på forbruk av kulturtilbud og endring over tid. Det er fint om studentene sammenligner nivået på forbruket av ulike kulturtilbud (høyere andel som gikk på kino enn på museum) og sammenligner endringer over tid, som f eks at andelen respondenter som har gått på kino ser ut til å ha økt i perioden, mens andelen som har besøkt museum i mindre grad er endret.

1-4: Læreboken omtaler tidsdesign sammen med analysetyper av longitudinelle data i kapittel 7. De vanligste design er gjentatte tverrsnittundersøkelser, ulike typer paneldata (prospektive undersøkelser), og retrospektive undersøkelser (erindringsdata). En god besvarelse vil nevne at paneldata er gjentakende undersøkelser av de samme enhetene (individene), gjentatte tverrsnittundersøkelser er undersøkelser av forskjellige utvalg (ikke samme individer) som antas å være representative for den samme populasjonen og som det derfor gir mening å sammenligne på gruppenivå/aggregert nivå; mens erindringsdata er én undersøkelse av samme individ, som bes å huske bakover i tid om det man ønsker å få informasjon om. Dersom noen har tid til å nevne fordeler og ulemper med ulike design er det veldig bra, men ettersom dette ikke er spurt om er det ikke grunnlag for å gi trekk til dem som ikke skriver noe om dette.

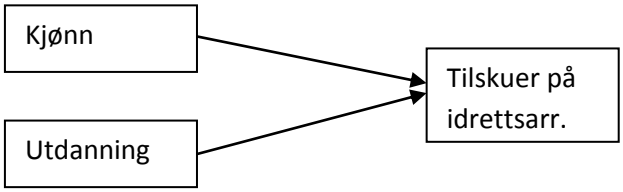
OPPGAVE 2

2-1: Avhengig variabel er deltakelse på idrettsarrangement de siste 12 måneder, med verdiene ”Tilskuer” og ”Ikke tilskuer”. Uavhengig variabel er kjønn, med verdiene ”kvinner” og ”menn”. Tabellen omfatter respondenter i alderen 20-79 år, som deltok i undersøkelsen i 2004.

Prosentdifferansen mellom menn og kvinner er 13 prosentpoeng. 62 prosent av mennene og 49 prosent av kvinnene har svart at de har vært tilskuer på et idrettsarrangement de siste 12 månedene.

OPPGAVE 3

3-1: Modell for årsakssammenhengen mellom variablene i Tabell 2:



3-2: Effekten av kjønn på bruk av kulturtilbud (her: deltakelse på idrettsarrangement) kontrollert for utdanning er 15 prosentpoeng.

Effekten av utdanning på bruk av kulturtilbud (her: deltakelse på idrettsarrangement) kontrollert for kjønn er 17 prosentpoeng.

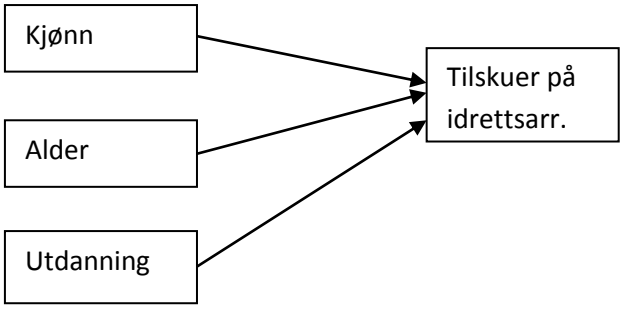
Samspill er hhv – 9, eller 9, avhengig av hvilket utgangspunkt man velger.

En god besvarelse viser gangen i utregningen, og gir en kort tolkning av effektene og samspillet.

3-3: Det er statistisk sett ikke mulig å si noe om sammenhengen i befolkningen mellom kjønn, utdanning og bruk av kulturtilbud (her: deltakelse på idrettsarrangement) basert på Tabell 1 og 2. Vi mangler informasjon om antall personer i de ulike undergruppene (antall kvinner med lav utdanning, antall kvinner med høy utdanning, antall menn med lav utdanning, antall menn med høy utdanning) og kan derfor ikke regne ut den vanlige statistiske signifikanstesten som ligger til grunn for statistisk generalisering (Kji-kvadrat-testen).

OPPGAVE 4

4-1: Modell av årsakssammenhengen mellom variablene i Tabell 3:



Her kan studentene enten skrive formelen for regresjonsligningen som ligger til grunn for i Modell 3 er:

$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + e_i$

Eller den mer spesifikke regresjonsmodellen i Modell 3:

$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i}$

(dette skal være Y-hatt, og X-hatt'er – har ikkje tid til å finne ut av dette nå, men dere skjønner kva som menes ☺)

4-2: Konstantleddet i Modell 1 er den gjennomsnittlige verdien på Y for menn (ettersom menn er kodet 0); dvs at menn i snitt har gått nesten 12 ganger på idrettsarrangement de siste 12 månedene. Konstantleddet er signifikant, med en standardfeil på knappe 2,4 og en t-verdi på 5.

Regresjonsanalysen er basert på det samme utvalget som i Tabell 1 og 2; og de yngste her er 20 år. Statistisk sett er konstantleddet i Modell 2 den gjennomsnittlige verdien på Y for menn som har verdien 0 på aldersvariabelen, men utvalget omfatter bare respondenter i alderen 20 – 79 år, derfor har konstantleddet ikke substansiell mening. Det er mulig noen studenter blir forvirret av dette. Men det er ikke uvanlig at regresjonslinjen (evt regresjonsplanet i trivariat regresjon) ikke skjærer Y-aksen for det utvalget man har definert, slik at konstantleddet blir meningsløst. Det er mulig noen studenter antar at aldersvariabelen er kodet om, slik at de yngste i utvalget (20 år) er gitt verdien 0 i alder. Det er ikke gitt noen informasjon som tilsier at vi har kodet om aldersvariabelen, så dette er feil. Jeg synes uansett vi skal være milde i bedømmingen av svarene her.

Konstantleddet i Modell 3 er den gjennomsnittlige verdien på Y for menn som har verdien 0 på aldersvariabelen og som også har verdien 0 på utdanningsnivåvariabelen; dvs. vi har samme diskusjon som over med hensyn til hva 0 på aldersvariabelen betyr. Dette betyr igjen at konstantleddet ikke gir substansiell mening.

4-3: Alle modellene er basert på data som ble samlet inn i 2004, og omfatter personer i alderen 20-79 år på intervjutidspunktet. Modell 1 viser at menn i snitt går knappe 12 ganger på idrettsarrangmenter i løpet av de siste 12 månedene (konstantleddet), mens regresjonskoeffisientenf or kjønn er -3,7, noe som betyr at kvinner i gjennomsnitt går 8,3 ganger (3,7 ganger sjeldnere) enn menn på idrettsarrangement i løpet av de siste 12 månedene. Kjønnsvariablen er signifikant, men t-verdien er ikke så veldig høy (-2,4).

I Modell 2 er regresjonskoeffisienten knyttet til aldersvariabelen liten (-0,134), kontrollert for kjønn. Dette betyr ikke nødvendigvis at alder har en liten effekt på Y. Aldersvariabelen måler fortløpende år mellom 20 og 79. Menn på 20 år går i snitt går 15,2 ganger på idrettsarrangement (17,9 - 0,134*20); mens kvinner på 20 år går ca 11,6 ganger i året (vi trekker fra regresjonskoeffisienten for kvinner, som er -3,55 i Modell 2). En god besvarelse bør regne ut forventet verdi på Y for et par aldersgrupper, for eksempel for de som er 30 eller 40 eller 50 år. Eventuelt for dem som er 79 år. Vi ser at kjønnsvariabelen fremdeles er signifikant, og koeffisienten har endret seg noe, men relativt lite fra Modell 1. Vi ser også at aldersvariabelen er signifikant.

Tolkningen av modell 3 blir tilsvarende, men nå er det tre X-variable, og regresjonskoeffisientene måler effekten av hver av de tre på Y, kontrollert for de to andre X-variablene. Den nye variabelen i Modell 3 er utdanningsnivå. Regresjonskoeffisienten knyttet til utdanningsnivå ikke signifikant. Tolkningen er derfor at når vi kontrollerer for kjønn og alder har utdanning ingen betydning for hvor ofte folk går på idrettsarrangement. Der fint om studentene regner ut et par (to-tre) eksempler på hvilke verdier på Y som forventes for ulike verdikombinasjoner på X'ene.

Modelltilpasningene (gjengitt i egen tabell) viser at ingen av de tre modellene fanger opp særlig mye av variasjonen i Y, dvs ingen av de tre variablene vi har inkludert gir et betydelig bidrag til vår forståelse av Y. Forklart varians (adjusted R square) er 0,03 i Modell 1, 0,08 i Modell 2, og 0,09 i Modell 3. Det viktigste bidraget ser ut til å være introduksjon av aldersvariabelen i tillegg til kjønn i Modell 2. Men vi bør være forsiktige med å konkludere hvilke av X-variablene som har størst betydning for Y, fordi dette er avhengig av i hvilken rekkefølge vi trekker dem inn.