Sensorveiledning - PSY2014 V2022

Oppgave 1: Popularitet hos barn (del 1)

I denne oppgaven skal vi studere om språkforståelse (SPRÅK) er assosiert med popularitet (POPULARITET) hos skolebarn, i et utvalg på n=200.

1. Modell 1 er en lineær regresjonsmodell der POPULARITET er avhengig variabel, og SPRÅK er eneste uavhengig variabel.

- Hvordan vil du oppsummere sammenhengen mellom POPULARITET og SPRÅK på bakgrunn av modell 1?

Her er det ønskelig at det trekkes frem:

- **Tolkningen av estimatet av** b_1: Stigningstallet for variabelen SPRÅK blir estimert til $\hat{b}_1 = 1.0$.
- **Signifikansverdien til** b_1: $\hat{b}$ er statistisk signifikant (t-verdi til >7.6, og p-verdi <0.05).
- **Forklart varians**: R^2 kan regnes ut ved $R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{619.26}{182.40+619.26} = 0.2275$. I overkant av 22% av variansen i popularitet kan forklares av variabelen SPRÅK.
- **Konfidensintervallet** (kan gjerne defineres), viser at hypoteser for verdier av b_1 der koeffisienten er <0.76 eller >1.2898 kan forkastes. Dette intervallet rommer ikke 0, så igjen ser vi at nullhypotesen kan forkastes på et 0.05 nivå.

- Hva er den forventede skåren på POPULARITET for et barn som har skåren 10 på variabelen SPRÅK?

$$\begin{aligned} \hat{POPULARITET} &= \hat{b}_0 + \hat{b}_1 \\ &= 10.0395 + 1.0251 \cdot SPRÅK \end{aligned}$$

Forventet popularitet for et barn med skåren 10 på variabelen SPRÅK blir:

$$E(POPULARITET|SPRÅK = 10) = 10.0395 + 1.0251 \cdot 10 = 20.2905 \approx 20.3$$

Læreboken bruker notasjonen $E(X|Y)$, men vi har også ofte skrevet dette som $\hat{Y}$, så det er selvsagt helt i orden at kandidatene bruker denne notasjonen.

Oppgave 2: Popularitet hos barn (del 2)

Vi skal nå bygge videre på modell 1 fra oppgave 1. I studien ble det også samlet data om den sosioøkonomiske statusen til barnas familie (SES, dvs. foreldrenes grad av utdannelse og inntekt), samt barnas grad av personlighetstrekket introversjon (INTROVERSJON). Det ble så kjørt to regresjonsmodeller. I Modell 2a var SPRÅK og SES uavhengige variabler, mens i Modell 2b var SPRÅK og INTROVERSJON uavhengige variabler.

I den vedlagte utskriften kan du se at stigningstallet for SPRÅK i de to nye modellene har endret seg i forhold til hva det var i Modell 1. Hvordan kan du forklare denne endringen?

I denne oppgaven ønskes det at kandidaten kan gjøre rede for at hvorfor estimatet for SPRÅK synker i modell 2a, mens det øker i modell 2b. Bakgrunnen for endringen kan leses ut fra korrelasjonstabellen, og forstås også ved å se på fortegnet til koeffisientene i utskriften.

Modell 2a

I denne modellen er SES en *konfunderende variabel*. Det å vokse opp i en familie med høy SES ar assosiert med bedre språkforståelse (positiv korrelasjon mellom SES og språk), og høyere popularitet (igjen positiv korrelasjon). Dersom SES ikke blir kontrollert for i modellene vil det føre til et kunstig høyt estimat på sammenhengen mellom SPRÅK og POPULARITET. Følgelig synker $\hat{b}_1$ fra 1.0 til 0.4 når SES inkluderes i modellen.

Modell 2b

I denne modellen øker den estimerte koeffisienten til SPRÅK når INTROVERSJON inkluderes. Her er INTROVERSJON en *supressorvariabel*. INTROVERSJON er positivt korrelert med språk, men negativt korrelert med popularitet. I dette utvalget er barn med godt språk mer introverte, og introversjon er negativt assosiert med popularitet. Effekten av SPRÅK og INTROVERSJON virker i motsatt retning, og effektene vil delvis kanselleres. *Kontrollert* for INTROVERSJON er det en større sammenheng mellom SPRÅK og POPULARITET. F.eks. for to barn som har samme grad av introversjon, men skiller seg en enhet på SPRÅK, er den mer språksterke forventet å skåre 1.59 enheter høyere på POPULARITET.

Oppgave 3: Tilfredshet på arbeidsplassen

I oppgave 3 skal vi studere i hvilken grad tilfredshet på arbeidsplassen henger sammen med opplevd stress i arbeidsdagen. Etter at data er samlet inn fra 300 ansatte i ulike bedrifter, blir følgende to lineære regresjonsmodeller kjørt.

```
modell_3 <- lm(tilfredshet~stress)
modell_4 <- lm(tilfredshet~stress + I(stress^2))
```

1. I vedlegget finner du plot av residualene fra modell_3 og modell_4. På hvilke måter bryter residualene fra modell 3 med antagelsene i en lineær regresjonsanalyse, og hvorfor er det i lys av dette hensiktsmessig å kjøre modell 4?

- I figuren øverst til venstre ser vi en tydelig struktur i residualene fra `modell_3`. Residualene later til å følge en annengradsfunksjon. De er store og positive for lave verdier av stress, synker i verdi for økende grad av stress, for så å snu ved et stressnivå på rundt 5, og øker så igjen. Dette mønsteret bryter med sentrale antagelser om fordelingen av feil-termene (epsilon) i lineær regresjon. Feiltermene er antatt å være uavhengige og trukket fra samme normalfordeling, og dette skal gjenspegles i residualene.
- I residualene fra `modell_4` ser vi derimot ingen tydlige brudd i antagelsene (disse antagelsene kan gjerne beskrives i noe mer detalj). Etter at annengradsleddet er inkludert, later ingen antagelser til å være brutt.

2. Basert på modell 4, hva er forventet nivå av tilfredshet for en person som rapporterer en stress-skåre på 10? Hvordan vil du beskrive forholdet mellom stress og tilfredshet på arbeidsplassen, utifra estimatene i modell 4?

- Sammenhengen mellom stress og tilfredshet er ikke lineær, men beskrives godt av en annengradsfunksjon. Både førsteordens og annenordens-leddene er positive, så tilfredshet på arbeidsplassen vil i dette utvalget øke mer og mer ettersom grad av stress øker.

$$\begin{aligned} \hat{tilfredshet} &= \hat{b}_0 + \hat{b}_1 \cdot stress + \hat{b}_2 \cdot stress^2 \\ &= 5.244271 + 0.293062 \cdot 10 + 0.146957 * 10^2 \\ &= 22.87059 \approx 22.9 \end{aligned}$$

Oppgave 4: Navigasjon

I denne oppgaven skal vi se på forholdet mellom navigasjon (evnen til å finne veien til et ønsket mål), IQ, og selvtillit.

1. I vedlegget (Oppgave 4 del 1) finner du R-utskrift fra en regresjonsanalyse der navigasjon er avhengig variabel, mens IQ og selvtillit er uavhengige variabler. Hvilket av de tre regresjonsplanene korresponderer til denne under figuren? Begrunn svaret.

Her var det kommet inn feil variabelnavnene i R-utskriften, da disse var hentet fra et tidligere utkast. Spatial_hukommelse i utskriften skal erstattes med IQ, og Verbal_hukommelse med selvtillit. Dette ble komunnisert via Canvas til kandidatene, og eksamenstiden ble utvidet med 15 minutter. Kandidatene var i forkant informert om at de burde være oppmerksom på eventuelle beskjeder på Canvas, men det er likevel ikke alle som har fått med seg korreksjonen før etter at eksamen var avslutett. Dette bør tas med i betraktning når denne oppgaven sensureres.

Gitt endringene i variabelnavn er det figur A som korresponderer til R-utskriften. Dersom man holder IQ konstant men øker selvtillit, ser vi knapt noen forventet endring i Navigasjon. Holder vi selvtillit konstant men øker IQ, stiger forventet navigasjon med omtrent en enhet per IQ-poeng.

2. Visuell hukommelse er hukommelse for ting som må huskes i visuell form. I vedlegget, del 2, finner du utskrift fra kjøring av følgende fire R-kommandoer:

```
summary(lm(Navigasjon~IQ))
summary(lm(visuell_hukommelse~IQ))
summary(lm(Navigasjon~visuell_hukommelse))
summary(lm(Navigasjon~IQ+visuell_hukommelse))
```

Forklar hva vi tester med i denne sekvensen av modeller. Hva blir din samlede konklusjon av alle fire regresjonsanalysene?

Sekvensen med modeller går gjennom den klassiske fremgangsmåten for å vurdere hvorvidt en variabel fungerer som en *mediator* (kan gjerne defineres). I denne oppgaven vil jeg kalle `IQ` for eksponeringen, `Navigasjon` som utfallet, og `visuell_hukommelse` som en mulig mediator. (Det forventes ikke at kandidatene bruker samme terminologi).

- I modell 1 testes det om eksponeringen er assosiert med utfallet. (Er det noen effekt å mediere?)
- I modell 2 tester vi om eksponeringen er assosiert med mediatoren. (Er det belegg for at mediatoren er påvirket av eksponeringen?)
- I modell 3 tester vi om mediatoren er assosiert med utfallet. (Er det belegg for at utfallet er påvirket av mediatoren?)

Dersom `visuell_hukommelse` skal kunne mediere effekten av `IQ`, kreves det at den uavhengige variabelen i alle de tre analysene over er signifikant. I siste skritt, Modell 4, er både eksponeringen og mediatoren inkludert i analysen, og man vurderer om det er en betydelig nedgang i den estimerte koeffisienten til `IQ` når `visuell_hukommelse` også inkluderes i modellen. Denne nedgangen tas da til inntekt for at mediatoren er av betydning.

I våre analyser ser vi derimot at prediktoren i modell 3 *ikke* er signifikant, og vi kan således ikke hevde at høyere nivåer av `visuell_hukommelse` fører til signifikant høyere nivåer av `Navigasjon`. Tilsvarende ser vi ingen nedgang i den estimerte koeffsienten til `IQ` i modell 4. Denne sekvensen av modeller gir derfor *ikke* støtte til påstanden om at effekten av IQ på Navigasjon medieres av visuell hukommelse.

Oppgave 5: Avbrytelser og møteform

I denne oppgaven skal vi studere forhold som henger sammen med antall ganger den som snakker under et møte blir avbrutt av andre møtedeltagere.

- Faktoren TYPE er kodet 1 dersom et gitt møte holdes fysisk, 2 om det holdes over telefon (kun lyd), eller 3 om møtet holdes som videosamtale over Zoom, Teams, eller liknende.
- Faktoren ERFARING er kodet "Uerfaren" dersom gruppen ikke har samarbeidet tidligere, og "Erfaren" dersom gruppen har jobbet sammen overt tid.

1. Forklar kort begrepene "innen-gruppe varians" og "mellom-gruppe varians", og hvordan forholdet mellom dem kan brukes til å vurdere nullhypotesen in en variansanalyse.

> Se forelesning 7 for en gjennomgang av disse begrepene, eventuelt avsnitt 12.3 i læreboken (Agresti). Det viktige poenget er at populasjonsvariansen under H0 kan estimeres enten gjennom mellom-gruppe variansen, eller gjennom innen-gruppe variansen. Ratioen av de to (MS_b / MS_w) gir i F-verdien, som under H0 har en forventet verdi på 1.0, mens der høyere verdier gir støtte *mot* nullhypotesen.

2. I vedlegget finner du utskrift fra en enveis variansanalyse over faktoren TYPE. Fyll inn de manglende verdiene i den vedlagte tabellen, og konkluder med hensyn på resultatene.

> ```
> > summary(ANOVA1<-aov(AVBRYTELSER~TYPE, data=DAT))
>              Df Sum Sq Mean Sq F value   Pr(>F)
> TYPE          2   1131   565.4    43.3 1.65e-15 ***
> Residuals   147   1920    13.1
> ---
> Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> ```
>
> - Frihetsgradene til faktoren `TYPE` er gitt av antall grupper (G) minus 1. (Frighetsgraden er ikke G, da utregningen av SS_b forutsetter gjennomsnittet for hele utvalget (grand mean) også er utregnet på de samme dataene).
> - $MS_b = \frac{SS_b}{df_b} = \frac{1131}{2} = 565.4.$
>
> Besvarelsen bør presisere hva nullhypotesen til en ANOVA er; her at gjennomsnittsnivået til variabelen *abrytelser* er lik i alle gruppene (møtetypene) i populasjonen. Under H0 skulle vi forvente en F-verdi på 1.0, med en samplingfordeling som følger en F-fordeling gitt av staistiskkenes frihetsgrader. I analysen over er F-verdien 43.3, og dette er sterkt signifikant. Det er *svært* usannsynlig å observere gjennomsnittsverdiene vi i de ulike gruppene dersom utvalget vårt var trukket fra en populasjon der snittet på avbrytelser var likt i de tre møtetypene

3. I neste omgang blir en toveis anova gjennomført, med faktorene TYPE og ERFARING. Konkluder med hensyn på resultatene fra analysen, og forklar hvorfor F-verdien til faktoren TYPE har endret seg fra enveis-analysen.

> I utskriften ser vi en signifikant *hovedeffekt* av både `TYPE` (p< 2e-16) og `ERFARING` (p< 2e-16). Sett over nivåene av erfaring er det forskjeller i antall avbrytelser på tvers av de ulike møteformene. På samme måte er det ulik gjennomsnittlig antall avbrytelser for grupper med lite kontra mye erfaring med å jobbe sammen. Disse hovedeffektene er tydelige i linjediagrammet, og dette bør påpekes.
>
> Til slutt ser vi en signifikant interaksjonseffekt `TYPE:ERFARING`, (p=0.000394). Hva en interaksjon er kan gjerne defineres i besvarelsen. Interaksjonseffekten kan sees i linjediagrammet gjennom at linjene ikke er parallelle, men at det særlig for videosamtaler er forskjeller på gjennomsnittlig antall avbrytelser over nivåene på faktoren erfaring.

Oppgave 6: Søsken og sosial angst

I denne oppgaven vil vi undersøke i hvilken grad det å være et enebarn (dvs. å ikke ha søsken) er forbundet med sosial angst som voksen. En skåre på 0 i variabelen `SOSIAL_ANGST` indikerer ingen sosial angst, mens en skåre på 1 indikerer sosial angst.

1. Hvorfor er lineær regresjon dårlig egnet til å undersøke dette forholdet?

- Lineær regresjon forutsetter at den avhengige variabelen er kontinuerlig. I denne oppgaven er den avhengige variabelen `SOSIAL_ANGST` dikotom/binær, altså målt med to nivåer (0 og 1). En lineær modell vil her kunne resultere i umulige forventede verdier (f.eks. negative verdier, eller verdier over 1.0).

2. I vedlegget finner du en krysstabell av variablene i data fra et utvalg på 760 personer. Vis at chi-kvadratverdien er 9.19.

Vi kan begynne med å finne forventet frekvens (E) i alle cellene.

$$E_{1,1} = 110 * 250/760 = 36.18421$$

$$E_{2,1} = 110 * 510/760 = 73.81579$$

$$E_{1,2} = 650 * 250/760 = 213.8158$$

$$E_{2,2} = 650 * 510/760 = 436.1842$$

Vi kan så regne ut kjikvadrat verdien med formelen

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{50 - 36.18421)^2}{36.18421} + \frac{60 - 73.81579)^2}{73.81579} + \frac{200 - 213.8158)^2}{213.8158} + \frac{450 - 436.1842)^2}{436.1842}$$

$$= 5.27512 + 2.585843 + 0.8927139 + 0.4376049$$

$$= 9.191282$$

3. Hva vil du konkludere med vedrørende forholdet mellom det å være et enebarn og sosial angst?

- Nullhypotesen vi skal teste ved hjelp av Kjikvadrat-statistikken er at variablene ENEBARN og SOSIAL_ANGST er uavhengige.
- Tabellen har (C-1)*(R-1) = 1 frihetsgrad. Den kritiske verdien for å forkaste nullhypotesen ser vi i tabellen er 3.84. Kjikvadratverdien er større enn den kritiske verdien, og vi kan derfor forkaste nullhypotesen. Vi kan konkludere med at forekomsten av sosial angst ikke er like stor i gruppen med enebarn som i den med søsken.
- For å få mer innsikt i hvordan variablene er assosierte kan vi se på residualene og stolpediagrammet. De standardiserte residualene kan lese som z-fordelte variabler, og følgelig er verdier mindre enn -2 eller større enn 2 å regne som store/av særlig interesse. Ut ifra residualene ser vi at det er vesentlig flere enebarn med sosial angst enn det vi forventer under H0 (og tilsvarende færre uten sosial angst).

Til slutt kunne det være rimelig å vurdere om antagelsene for valide kjikvadrat-verdier kunne være brutt.