Sensorveiledning SOS4020 Ny ordning - høst 2004

a) En standardfeil måler graden av statistisk usikkerhet knyttet til et estimat – i dette tilfellet et (aritmetisk) gjennomsnitt. Standardfeilen er standardavviket i sannsynlighetsfordelingen til parameterestimatet.
Et konfidensintervall er en øvre og nedre grense for en parameter, konstruert slik at intervallet vil innholde den sanne parameterverdien med en valgt sannsynlighet. Når utvalget (som her) er stort, vil sannsynlighetsfordelingen ligge nær normalfordelingen og et 95% intervall konstrueres ved å multiplisere standardfeilen med 1.96.

$$SE(\hat{m}) = 3.4 \big/ \sqrt{984} = 0.108$$

Øvre grense: 13.4 + 1.96*0.108 = 13.61
Nedre grense: 13.4 − 1.96*0.108 = 13.19

b) Koeffisientene testes med t-test, og er signifikante på 5% nivået dersom testobservatoren er større enn 1.96 i tallverdi (to-halet test), siden antallet observasjoner er stort.

Kjønn: $t = 0.04 \big/ 0.22 = 0.18$, ikke signifikant

Ungdomsskole A: $t = 0.99 \big/ 0.29 = 3.41$, signifikant

Privat skole: $t = 0.59 \big/ 0.40 = 1.48$, ikke signifikant

Videregående skole: $t = 0.52 \big/ 0.28 = 1.86$, ikke signifikant

Det er ingen påviselig kjønnseffekt, kontrollert for skoletype, med hensyn til skoleholdninger. Elevene ved ungdomsskole A har mer positive skoleholdninger enn elevene ved ungdomsskole B (referanseskolen). Derimot er det ikke signifikante forskjeller mellom referanseskolen på den ene siden og henholdsvis privat skole og videregående skole på den annen side.
Ungdomsskole B har lavest skåre i gjennomsnitt, mens ungdomsskole A har høyest skåre.

c)

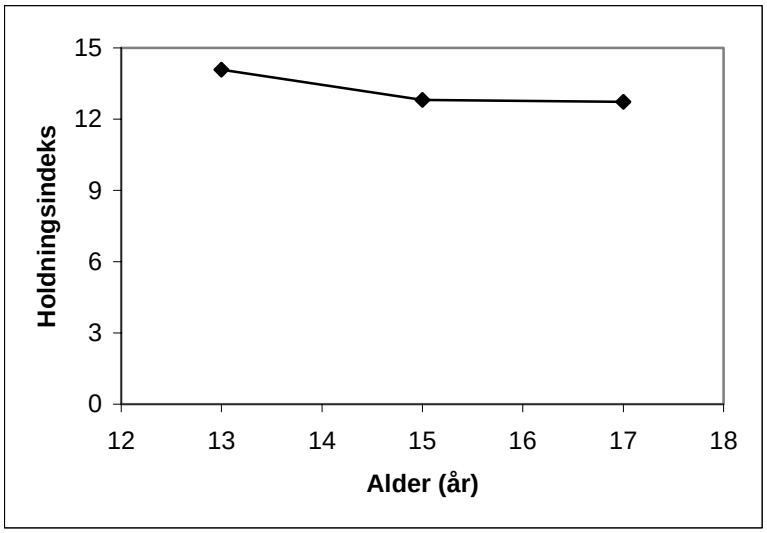
Alder: $t = -0.94 \big/ 0.29 = -3.24$

Alder kvadrert: $t = 0.15 \big/ 0.05 = 3.00$

Begge koeffisienter er signifikante.
Hensikten med andregradsleddet er å ta vare på eventuell kurvelinearitet. Siden koeffisienten til andregradsleddet er statistisk signifikant forskjellig fra null, må nullhypotesen om at sammenhengen er lineær forkastes.
Man kan også innse at siden andregradleddet er positivt, står vi overfor en u-formet sammenheng og et minimumspunkt når aldersvariabelen er $X^* = -(-0.94) \big/ (2*0.15) = 3.1$, dvs. i aldergruppen 13 + 3.1 = 16.1 år.

d) Predikert (forventet) verdi for gutter (kjønn=0) i ungdomsskole B (0 på alle dummyvariable) er:

13 år (alder=0): 14.09 + (-0.94)*0 + 0.15*0*0 = 14.09
15 år (alder=2): 14.09 + (-0.94)*2 + 0.15*2*2 = 12.81
17 år (alder=4): 14.09 + (-0.94)*4 + 0.15*4*4 = 12.73



Kurvene for andre skoleslag blir parallelle med kurven for ungdomsskole B og vil gjennomgående ligge på et høyere nivå. Det forutsettes ikke at kandidaten har gjort utregningen for alle skoleslag.

Kandidaten vil kanskje påpeke et problem med den substansielle tolkbarhet av kurvene, idet hver enkelt av disse kurvene (med mulig unntak av kurven for privatskole) for noen aldersgrupper vil referere til grupper med svært få (eller ingen) observasjoner: Det er få 17-åringer i grunnskolen og ingen 13- og 15-åringer i videregående skole.

De positive skoleholdningene **avtar** med andre ord med stigende alder blant de yngste. Mellom 15- og 17-åringene er det imidlertid liten forskjell. Enkelte kandidater vil sannsynligvis innse at laveste holdningsverdi gjenfinnes i aldersgruppen 16 år, enten ved å beregne predikert verdi også for denne gruppen, eller som angitt ovenfor. I så fall vil de også innse at kurven er økende fra 16 til 17 år.

e) Økning i standardfeilen når nye variable tas inn i modellen tolkes i læreboken som tegn på kolinearitet. Det er et tegn på at det er forholdsvis høy korrelasjon mellom variabelen ”Videregående skole” og de to aldersvariablene som er tatt inn i modellen. Grunnen til dette er at det bare er de eldste som går på videregående.

f) Variabelen ”Sosial status” er ikke signifikant ($t = -0.20$), mens ”God foreldrekontakt” er klart signifikant ($t = 12.6$). Det er med andre ord ingen klare tegn til at barn fra lavstatushjem har andre holdninger til skolen enn barn fra høystatushjem (når en som her kontrollerer for kjønn, alder, foreldrekontakt og skoleslag). Derimot er det en klar sammenheng mellom graden av foreldrekontakt og skoleholdninger, kontrollert for de øvre variablene i modellen: Barn som sier de har god kontakt med sine foreldre har gjennomgående atskillig mer positive holdninger enn barn med dårlig foreldrekontakt. En legger merke til at R^2 øker fra under 2% til over 14% når denne variabelen tas inn i modellen. Det ser med andre ord ut til at gode relasjoner til foreldre har en gunstig innvirkning på de unges holdninger til skolen.

g) Det er ikke grunnlag for å hevde at forskjellen mellom de to ungdomsskolene er resultat av ulikheter mht. sosial status og elevenes foreldrekontakt, siden parameterverdien til ungdomsskole A ikke har forandret seg nevneverdig og fortsatt er signifikant forskjellig fra null i modell nr. 3.

h) Dersom sosial status bare påvirker skoleholdninger blant de i videregående skole, og ikke blant andre, vil det foreligge samspill mellom variablene "Sosial status" og "Videregående skole". Én måte å undersøke dette på vil bestå i å gjøre partielle analyser, der en først ser på de som går i videregående skole separat, og deretter de øvrige. Hvis hypotesen er riktig vil en finne en effekt av sosial klasse blant elevene i videregående, men ikke blant de andre. Den andre tilnærmingen vil bestå i å bruke et samspillsledd, definert som produktet av de nevnte variablene. Et signifikant samspillsledd vil peke i retning av forskjellig effekt av sosial status i ulike skoleslag.

i) $Odds = \frac{Andel}{1 - Andel} = \frac{0.205}{1 - 0.205} = 0.258$

Odds er med andre ord forholdet mellom sannsynligheten for å ha skulket mer enn fem ganger og sannsynligheten for ikke å ha gjort det. En logit er en logaritmisk omkodet odds og dette er et alternativt mål på hvor vanlig forekommende et fenomen er. Hensikten med omkodingen til logit er å få et mål for den avhengige variabelene som er lineært assosiert med den uavhengige variabelen. Ulempen ved logits som mål, er at det ikke har noen enkel intuitiv tolkning.

j) Exp(b) er antilogaritmen (eksponensial-funksjonen) til logistiske parametere.
For konstantleddet er denne størrelsen lik oddsen for at referansegruppen (dvs. de med null på holdningsvariabelen) skal ha egenskapen. Oddsen for at de med verdi 0 på holdningsvariabelen skal ha skulket skolen mer enn fem ganger er med andre ord 1.91. Dette svarer til en andel på 1.91/(1 + 1.91) = 0.66, eller 66%. Skulking er med andre ord atskillig mer utbredt blant de med laveste skåre på holdningsvariabelen.
For den uavhengige variabelen er antilogaritmen til den logistiske parameterverdien lik oddsraten. Den forteller med andre ord med hvilken multiplikativ faktor oddsen endres når den uavhengige variabelen stiger med en enhet. Her er oddsraten 0.854 og det vil si at oddsen synker med omtrent 15% (=100*(1 − 0.854)) når holdningsmålet øker med éen enhet. Skulk er med andre ord mindre utbredt blant de med høy skåre på holdningsvariabelen (=positive skoleholdninger).
Det fremgår av Wald-testen for variabelen "Holdninger til skolen" at sammenhengen er statistisk signifikant − testobservatoren er stor og signifikanssannsynligheten liten. Sannsynligheten for å få den observerte logistiske parameterverdien (eller en som i tallverdi er enda større) bare som resultat av tilfeldigheter, er mindre enn 0.001, dvs. mindre enn 1 promille.

k) Hosmer-Lemeshow testen baserer seg på en sammenligning av observerte og predikerte antall personer, og er generelt et mål på modellens tilpasning til data. I dette tilfellet, med bare én uavhengig variabel med mange verdier, er den et mål på om den logistiske kurven gir en god beskrivelse av den empiriske sammenhengen. Hosmer-Lemeshow testen viser at vi ikke kan forkaste nullhypotesen om at sammenhengen følger en logistisk kurve, idet signifikanssannsynligheten er 0.13. Sannsynligheten for å få det avviket vi her har bare som resultat av tilfeldigheter er med andre ord 13%.