Sensorveiledning - Våren 2015. SOS1120

Av Øyvind Wiborg

Generelt:
Svar utenfor {} er tilstrekkelig for full poengpott. Svar som inkluderer elementer innenfor {} styrker helhetsvurderingen av kandidat. Avrundinger i flere desimaler som leder til avvikene utregninger aksepteres også som full pott. Det skal ikke gis trekk for følgefeil i oppgaver som baseres på gale utregninger i tidligere oppgaver. Skjønnsmessig vurdering gis i slike tilfeller.

Bruk skåringsskjema vedlagt i excelformat. Det anbefales at spesielt gode svar fargemarkeres med grønn. Slik at det er lettere å oppsummere helhetsinntrykket om studenten ligger og vipper mellom to karakterer.

Nedre grenser for følgende karakterer er:
* A: 38 (maksimal poengscore er 42)
* B: 34
* C: 25
* D: 20
* E: 13
* F: Under 13

Oppgave 1

a) Hvilke målenivå befinner variablene «degree», «coninc», «region», «hrs1» seg på? (1 poeng)

SVAR:

degree = ordinalnivå **0.25**
coninc = forholdstallsnivå **0.25**
region = nominalnivå **0.25**
hrs1 = forholdstallsnivå **0.25**

b) Begrunn kort hvorfor du mener «region» befinner seg på det målenivået (1 poeng)

SVAR: Variabelen er på nominalnivå fordi verdikategoriene er gjensidig utelukkende og uttømmende. Verdiene kan derimot ikke rangeres.

{ Denne kategoriske variabelen har heller ikke faste intervaller eller noe absolutt nullpunkt. }

c) Hva ville sentraltendensmålet *modus* ha fortalt oss om du hadde brukt dette målet på variabelen «region»? (0.5 poeng)

SVAR:
Modus ville ha angitt den regionen som forekommer hyppigst.
{Alternativt: modus angir her den regionen hvor det bor flest folk}

d) Kan vi bruke *median* på variabelen for region? Ja / Nei (0.5 poeng)

SVAR: NEI

{For å bruke median som et mål på sentraltendens kreves det at variabelen er på minst ordinalnivå. Dette betyr at verdiene/verdikategoriene skal kunne rangeres fra høy til lav. «Region» er kun på nominalnivå. Det gir ingen mening å rangere «landsdeler»}

Tabell 1: Frekvenstabell av utdanningskategorier. N=26039

rs highest degree	Freq.	Percent	Cum.
lt high school	2,687	10.32	10.32
high school	13,820	53.07	63.39
junior college	1,832	7.04	70.43
bachelor	5,084	19.52	89.95
graduate	2,616	10.05	100.00
Total	26,039	100.00	

e) Hvilken utdanningskategori er modus i tabell 1? (1 poeng)

SVAR: «High school» (evt. «videregående skole»)
{Dette er verdien som forekommer hyppigst/oftest i tabellen: 53.07%}

f) Tolk *medianen* i tabell 1(1 poeng)

SVAR: Ifølge medianen er «high school» den mest sentrale verdi i tabellen hvor den midterste verdien er gitt innenfor den kumulative prosenten.

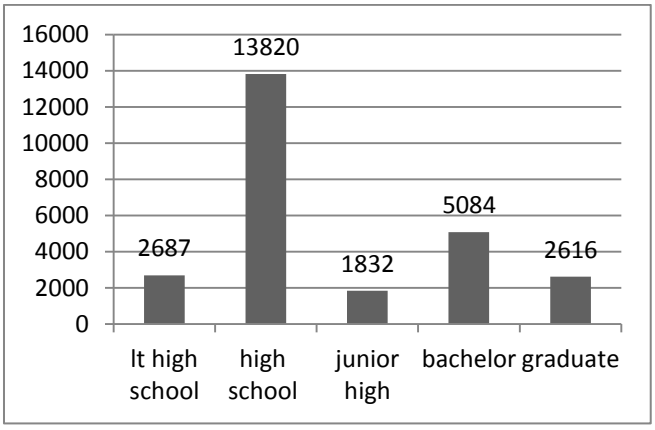
{Medianen angir den midterste verdien i rekken av alle de rangerte verdiene. I tabellen finner man medianen ved å kikke på den kumulative prosenten. Medianen angis av kategorien hvor den kumulative prosenten først innholder den miderste prosentilen 50%}

g) Hvor stor prosentandel av utvalget har fullført bachelorgrad eller høyere ifølge denne tabellen? (1 poeng)

SVAR: 19,52+10,05 = 29,57%

h) Tegn et histogram basert på tabell 1. Marker stolpene med kategorinavn. Bruk antall observasjoner på y-aksen. (1 poeng)

SVAR:

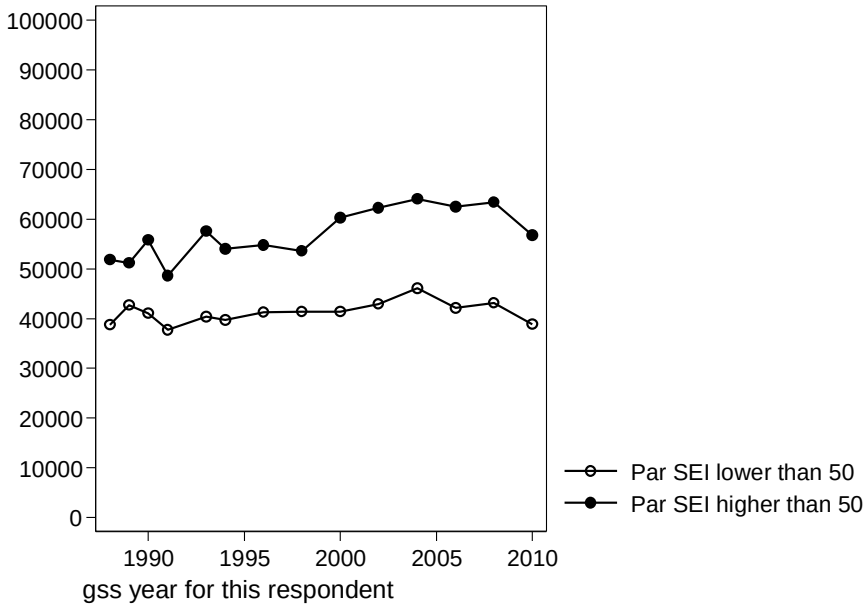


{Det kreves ikke her at man har tallverdien på toppen av stolpene. Mulig uklarhet: oppgaven nevner «histogram» og ikke «stolpediagram» (som er det mest riktige for kategoriske variabler). Jeg har ikke lagt stor substansiell vekt på skillet mellom histogram/stolpediagram i forelesningene. Dersom studenten påpeker dette, teller det bare positivt i helhetsinntrykket. Men det viktigste poenget med oppgaven er at studenten forstår/klarer å fremstille fordelingen av verdikategoriene i tabellen grafisk}

Oppgave 2

Anta at figuren (figur 1) nedenfor er representativ for den arbeidsføre delen av USA. *Beskriv kort* hva linjene i figuren nedenfor forteller oss om inntektsutviklingen avhengig om de har fra foreldre med høy eller lav yrkesposisjon. (2 poeng)

FIGUR1: Gjennomsnittlig årlig inntekt (i dollar) for personer med foreldre som har lav og høy yrkesstatus («parents' sei»: Par SEI). Personer mellom 30-70 år.



SVAR: Linjene viser at den gjennomsnittlige inntekten er relativt stabil (/svakt økende) i årene 1985 til 2010) for begge sosioøkonomiske grupper

{Svar i samme dur aksepteres også: Trenden er (relativt stabil/) svakt økende for begge grupper. Forskjellen mellom de som har foreldre med høy og lav sosial bakgrunn (parsei) ser ikke ut til å avta over tid. }

Oppgave 3

Tenk deg at en forsker stiller seg opp foran en kino klokken 13.00, en helt tilfeldig valgt dag. Kinoen har nettopp vist en barnefilm om Snurre Sprett. Folk strømmer ut av kinoen. Han spør de ti tilfeldig første personene som kommer ut av kinoen om de har lyst til å være med i en spørreundersøkelse. Anta at alle sier ja til å være med.

a. Er dette er et sannsynlighetsutvalg hvor analyseresultatene kan generaliseres til resten av befolkningen? Ja/nei (1 poeng)

SVAR: NEI

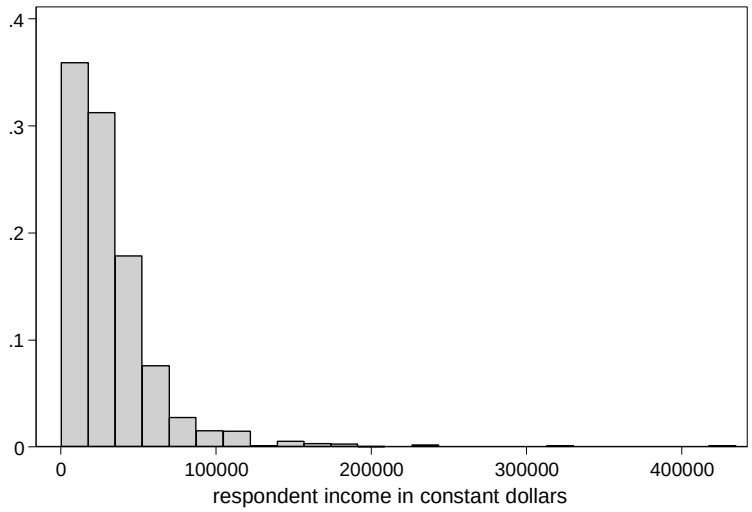
b. Begrunn kort svaret ditt i a. (2 poeng)

SVAR: Dette er et utvalg hvor forskeren ikke kjenner sannsynligheten for at individene trekkes fra populasjonen. Et sannsynlighetsutvalg krever at denne sannsynligheten er kjent.

{Svar langs liknende resonnementer aksepteres også: Utvalget er basert på slumpmessig trekning, og kan være skjevt og lite representativt for populasjonen. Her er det for eksempel rimelig å tro at de som kommer ut fra kinoen er i hovedsak barn og barneforeldre i en bestemt aldersgruppe.}

Oppgave 4

FIGUR 2: Histogram av inntektsfordeling i USA. Personer 30-70 år.



4.1 Inntektsfordelingen i figuren over er (1 poeng) :

a. høyreskjev
b. venstreskjev
c. normalfordelt
d. t-fordelt
e. uniform

4.2 Hvis *medianen* i denne inntektsfordelingen er 35602 $, hva er det mest sannsynlig *gjennomsnittet* for inntekt av følgende alternativer (1 poeng) :

a. 44680 $
b. 35602 $
c. 12101 $

Oppgave 5

Tabell 2: Krysstabell mellom inntekt (incDV) og utdanningsnivå (degreeDV). Dette er variabler som er omkodet til to kategorier basert på inntektsvariabelen «coninc» og utdanningsvariabelen «degree».

```
. ta degreeDV incDV , row V chi nof
```

RECODE of degree (rs highest degree)	RECODE of rincome (respondents income) <25000	>=25000	Total
low educ	70.99	29.01	100.00
high educ	37.81	62.19	100.00
Total	62.37	37.63	100.00

Pearson chi2(1) = 3.1e+03 Pr = 0.000
Cramér's V = 0.3004

5.1 Anta at utdanning påvirker inntekten. Regn ut prosentdifferansen i forhold hva som er antatt årsak. Vis utregningen.(1 poeng)

SVAR:
62,19%-29.01% = **33.18 %-poeng**

{evt: 37,81%-70.99% = **-33.18 %-poeng}**

5.2 Gi en kort tolkning av denne prosentdifferansen (1 poeng)

SVAR: Begge tolkninger aksepteres:

Tolkning 1 (%-poeng): Det er 33.18%-poeng flere som er tjener mer eller lik 25000 dollar blant de med høy enn lav utdanning i USA

Tolkning 2 (styrkemål): En prosentdifferanse på 33.18% forteller oss at det er en moderat(/noe sterk) sammenheng mellom utdanning og inntekt .

{Spesielt godt helhetsinntrykk om studenten har foretatt begge tolkninger. Det aksepteres også at studentene foretar tolkning 1 ut i fra den første kolonnen i tabellen (de som tjener mindre enn 25000 dollar i året}

5.3 Tolk korrelasjonsmålet Cramers V (1 poeng)

SVAR: Korrelasjonsmålet Cramers V(her på 0,30) forteller oss at det er en moderat(/noe sterk) sammenheng mellom utdanning og inntekt .

{Selv om standardiserte korrelasjonsmål som Cramers V går fra 0 (ingen korrelasjon) til 1 (perfekt korrelasjon), er det sjeldent at vi korrelasjoner nærmere 1 i samfunnsvitenskap. Dette fordi det ofte er tilfeldig støy (folk gjør tilfeldige ting) / eller at det forekommer målefeil i utvalg/variablene.}

5.4 Formuler en nullhypotese og en forskningshypotese for sammenhengen mellom utdanning og inntekt. (1 poeng)

H0: Det er ingen sammenheng mellom utdanning og inntekt i den arbeidsføre amerikanske befolkningen (/populasjonen)

HA: Det er en sammenheng mellom utdanning og inntekt i den arbeidsføre amerikanske befolkningen (/populasjonen)

5.5 Utfør en moderne kjikvadrattest. La signfikansnivået (alpha) være 0,05. Signifikanssannsynligheten (p-verdien) er oppgitt som «Pr =0,000». (2 poeng)

SVAR:
{I en moderne test kan vi sammenlikne den (observerte) signifikanssannsynligheten med det (kritiske) signifikansnivået, for å avgjøre om vi kan forkaste nullhypotesen}.

{Vi har allerede formulert forskning og nullhypotese for sammenhengen mellom utdanning og inntekt (i populasjonen). Signifikansnivået er også allerede angitt. }

Siden signfikanssannsynligheten (p=0.000) er lavere enn signfikansnivået (a=0,05) kan vi forkaste nullhypotesen med stor grad av sikkerhet. I forhold til det valgte signifikansnivået kan vi konkludere at vi tar feil i færre en 5% tilfeller når det gjelder å feilaktig forkaste en sann nullhypotese.

{Dette betyr at vi får (indirekte) støtte for forskningshypotesen om at det er en (statistisk signifikant)sammenheng mellom utdanning og inntekt i den (arbeidsføre) amerikanske befolkningen (/populasjonen).}

Oppgave 6
Nedenfor ser du at tabellen oppgir beskrivende statistikk for prestisjeskåre for arbeidstimer per uke.

Tabell 3: Deskriptiv statistikk for antall arbeidstimer per uke (hrs1).
Mean = gjennomsnitt, SE(mean) = standardfeil, N = antall observasjoner.

```
. tabstat hrs1 , stat(mean sem N )
```

variable	mean	se(mean)	N
hrs1	42.57777	.1117816	15835

a. Finn kritisk t-verdi i vedleggstabellen for et signifikansnivå på 99,9% (alfa= 0.001). Frihetsgrader er gitt ved df = N-1. (1 poeng)

SVAR:
{alfa er allerede gitt til 0.001}
Df = 15835-1= 15834,
Kritisk t-verdi er da : 3.291

{Vi finner kritisk t-verdi i tabellen bak ved å kikke nederst i kolonnen hvor alfa er 0.001}
.

b. Beregn (estimér) et 99,9% konfidensintervall for gjennomsnittet av arbeidstimer i utvalget basert på formelen (1 poeng):

$$KI_{99,9\%} = mean \pm t*SE$$

KI_nedre: 42.58 - 0.11*3.291 = 42.21799
KI_øvre: 42.58 +0.11*3.291 = 42.2942

Konfidensintervallet(KI) = {42.22 , 42.94}

{Det er tilstrekkelig at de viser utrekningen for KIøvre og nedre}

c. Gi en kort tolkning av konfidensintervallet (1 poeng)

SVAR:
enten Det sanne populasjonssnittet for gjennomsnittlig arbeidstimer per uke (hrs1) vil med 99% sikkerhet befinne seg innenfor 42,22 - 42,94 timer

eller I 99% av utvalgene i samplingfordelingen vil vi finne det sanne populasjonsnittet for gjennomsnittlig arbeidstimer per uke (hrs1) innen intervallet **42.22 - 42.94 timer**

{Svarene må likne på minst et av disse to alternativene}

Oppgave 7

Nedenfor ser du en korrelasjonstabell mellom flere av variablene. Alle tallene viser pearsons-r som er korrelasjonsmål for to kontinuerlige variabler av gangen.

Tabell 4: korrelasjonstabell. Pearsons-r

```
. corr coninc sei parsei region
(obs=24191)
```

	coninc	sei	parsei	region
coninc	1.0000			
sei	0.3915	1.0000		
parsei	0.2230	0.2956	1.0000	
region	-0.0230	0.0169	0.0444	1.0000

a. Tolk korrelasjonen mellom barnas sosiale yrkesstatus(sei) og inntekt (coninc). (1 poeng)
En Pearsons r på 0.39 viser at det er en (relativt sterk/) moderat positiv sammenheng mellom yrkesstatus og inntekt.

{Dette betyr at jo høyere yrkesstatus, desto mer tjener man}

b. Hvorfor viser alle korrelasjonene langs diagonalen i tabellen *r=1* ? (1 poeng)

Diagonalen viser perfekt korrelasjonen, fordi variablene er her korrelert med seg selv.

c. Hvilke(n) av variablene hører ikke hjemme rent statistisk sett i denne korrelasjonstabellen? (2 poeng)
Svar: «region»

{Region passer ikke inn her fordi den er på nominalnivå. Pearsons r krever kontinuerlige variabler på intervall eller forholdstallsnivå}

Oppgave 8 (2 poeng)

Noen hevder at meritokrati i moderne samfunn *fremmer* sosial reproduksjon på tvers av generasjoner i moderne samfunn. Ifølge dette resonnementet overføres talent, utdanningspreferanser og jobbpreferanser, enten sosialt eller genetisk fra foreldregenerasjon til barna. I et «perfekt» meritokrati vil familiebakgrunn dermed ha en fullstendig indirekte effekt på barnas inntekt. Familiebakgrunn vil virke gjennom at barn av foreldre med høy status oftere får høy utdanning og dermed høyere lønn enn barn av foreldre med lav status. Rent statistisk sett betyr dette at etter kontroll for barnas yrkesposisjon og utdanning, vil familiebakgrunn dermed ikke ha noen gjenværende direkte effekt på barnas inntekt.

Tegn et årsak-virkningsdiagram som viser hvordan foreldres sosiale posisjon (parsei50) antas å påvirke inntekt (coninc) *fullstendig* gjennom egen utdanning (educ13) og yrkesprestisje (sei50).

Parsei50→educ13→sei50→coninc

{Full pott så lenge tidsrekkefølgen her er etablert. Aksepteres også i tillegg; direkte effekter fra: educ13 → coninc, eller parsei50 → sei50}

(Studenten kan også bruke beskrivelser av variablene fremfor variabelnavn)

Oppgave 9

Vi ønsker her å undersøke sammenhengen mellom forelderens sosiale posisjon (parsei) og barnas inntekt oppgitt i US-dollars ($$$). Vi inkluderer kontrollvariabler i alle modellene.

Modell 1:

```
. reg coninc parsei50 age30, beta
```

Source	SS	df	MS		
Model	1.9270e+12	2	9.6350e+11	Number of obs	= 25337
Residual	3.5858e+13	25334	1.4154e+09	F(2, 25334)	= 680.72
				Prob > F	= 0.0000
				R-squared	= 0.0510
				Adj R-squared	= 0.0509
Total	3.7785e+13	25336	1.4914e+09	Root MSE	= 37622

coninc	Coef.	Std. Err.	t	P>|t|	Beta
parsei50	458.3265	12.52584	36.59	0.000	.2273166
age30	22.77181	14.28682	1.59	0.111	.0099021
_cons	48303.74	321.8364	150.09	0.000	.

9.1 Tolk konstantleddet i modell 1 (1 poeng)

SVAR:

I USA tjener (arbeidsføre amerikanere/) personer som er 30 år og som har foreldre med 50 seiskåre, i snitt 48303,7 dollar per år.

9.2 Tolk regresjonskoeffisienten for alder (age30) (1 poeng)

SVAR:

For hvert år eldre en person er, tjener man i snitt 22,8 dollar mer, kontrollert for (/når vi tar hensyn til/ holder konstant etc) foreldrenes sosiale posisjon (Parsei50).

[*** 0.5 poeng trekkes dersom studenten har glemt å nevne «kontroll for» ***]

Modell 2

```
. reg coninc parsei50 age30 educ13years sei50, beta
```

Source	SS	df	MS		
Model	6.9663e+12	4	1.7416e+12	Number of obs	= 24133
Residual	2.9021e+13	24128	1.2028e+09	F(4, 24128)	= 1447.93
				Prob > F	= 0.0000
				R-squared	= 0.1936
				Adj R-squared	= 0.1934
Total	3.5987e+13	24132	1.4913e+09	Root MSE	= 34681

coninc	Coef.	Std. Err.	t	P>|t|	Beta
parsei50	144.1102	12.7465	11.31	0.000	.0714973
age30	26.82744	13.9323	1.93	0.054	.0114513
educ13years	2839.341	97.82583	29.02	0.000	.2174337
sei50	483.5553	14.53315	33.27	0.000	.2415356
_cons	47670.97	321.2127	148.41	0.000	.

9.3 Fortolk den (ujusterte) R^2 i modell2 (1 poeng)

SVAR:

Til sammen forklarer alle de uavhengige variablene i modellen 19,36% av variasjonen i inntekt (coninc)

{Variabelnavn istedenfor «alle uavhengige variablene» aksepteres også. Alternativ tolkning1: Modellen forklarer/predikerer 19,36% av variasjonen i inntekt. Alternativ tolkning2: Modellen reduserer feilprediksjonen i inntekt med 19,36%}

9.4 Hvilken av variablene i modell 2 ser ut til å være minst viktig for barnas inntekt? (Oppgi tallet du vurderer dette ut i fra) (1 poeng)

SVAR:

Alder har minst innvirkning på den avhengige variabelen(/inntekt). Dette kan vi se ut i fra at alder har den laveste standardiserte regresjonskoeffisent/beta (beta=0.011)

[0.5 poeng i trekk dersom ikke betakoefisienten med tall ikke er nevnt]

{Den standardiserte regresjonskoefisenten lar oss sammenlikne variablenes relative betydning på den avhengige variabelen. Den gjør oss i stand til å sammenlikne betydningen av uavhengige variabler med ulik enhetsformat (e.g. kroner, alder etc,) og ulike fordelinger.}

9.5 Begrunn kort hvorfor regresjonskoeffisenten til foreledrenes sosiale yrkestatus (parsei50) synker i modell 2 i forhold til modell 1? (2 poeng)

SVAR:

Regresjonskoefisenten synker i modell2 når det kontrolleres for barnas egen utdanning og yrkesstatus. (1 poeng). En rimelig grunn til at dette er at betydningen av foreldrenes yrkesposisjon på barnas inntekt virker (delvis) gjennom barnas egen utdanning og sosial yrkesposisjon. (1 poeng)

[Full pott om begge poenger er med]

{Aksepteres også: «Personer med høy sosial bakgrunn, velger i større grad høyere utdanning, som igjen gjør at de får høy yrkesstatus med høy inntekt»}

9.6 Predikér årlig inntekt for en person med 14 års utdanning, som er 45 år gammel, som har gjennomsnittlig prestisjeskåre (SEI=50), og som har foreldre med prestisjeskåre PARSEI= 51. Vis utregning (2 poeng).

Modell 2:

Likning (1)

Inntekt = 47670.97 + 2839.341****educ13years*** + 26.82744****age30***+ 483.5553****sei50*** + 144.1101****parsei50***

Vi plugger inn verdiene for prediksjon (likning 2):

Inntekt = 47670.97 + **(**2839.341*1)+ (26.82744*15)+ (483.5553*0) + (144.1102*1)

= 51056.8328 dollar

[Vi krever ikke her at den første likningen (likning 1) er satt opp. Bare likning 2 med utrekning og predikert svar]

[NB: HER ER DET DESSVERRE SNEKET SEG INN EN UKLARHET: Det er ikke angitt hvilken modell 1,2, eller 3] som skal benyttes. Men modell 1 kan ikke brukes siden den ikke inneholder alle variablene. De som bruker modell 2 eller 3 får full pott]

Oppgave 10

Modell 3:

```
. reg coninc parsei50 age30 educ13years sei50 year2000 parsei50Xyear2000
```

Source	SS	df	MS		
Model	7.0587e+12	6	1.1765e+12	Number of obs	= 24133
Residual	2.8929e+13	24126	1.1991e+09	F(6, 24126)	= 981.14
				Prob > F	= 0.0000
				R-squared	= 0.1961
				Adj R-squared	= 0.1959
Total	3.5987e+13	24132	1.4913e+09	Root MSE	= 34627

coninc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
parsei50	63.55626	17.69913	3.59	0.000	28.86487	98.24765
age30	20.9517	13.93013	1.50	0.133	-6.352225	48.25562
educ13years	2822.469	97.74828	28.87	0.000	2630.876	3014.061
sei50	481.5806	14.51235	33.18	0.000	453.1355	510.0257
year2000	2917.458	448.6771	6.50	0.000	2038.023	3796.893
parsei50Xyear2000	148.1438	23.3335	6.35	0.000	102.4087	193.879
_cons	46245.7	383.8942	120.46	0.000	45493.25	46998.16

I modell 3 kontrolleres det for tid (før og etter 2000). Det er dessuten inkludert et samspillsledd mellom år og foreldrenes yrkesposisjon (parsei50Xyear2000).

a. Hvor mye mer betyr foreldrenes sosiale yrkes status for barnas inntekt etter år 2000, selv etter vi har kontrollert for barnas utdanning og yrkesstatus? (2 poeng).

   SVAR: 148,14 dollar

   {Dette kan leses direkte av samspillsleddet}

b. Ta utgangspunkt i likningen:

$$\widehat{Inntekt} = 46245.7 + 63{,}6 * \text{parsei50} + 2917{,}5 * \text{year2000} + 148 * \text{parsei50Xyear2000}$$

Utled to regresjonslikninger basert på dummyvariabelen for tid (year2000): før og etter år 2000.(3 poeng).

Før 2000:

$$\widehat{Inntekt} = 46245.7 + 63.6 * \text{parsei50} + 2917.5 * 0 + 148 * \text{parsei50} * 0$$

$$\boldsymbol{\widehat{Inntekt} = 46245.7 + 63.6 * \text{parsei50}}$$

Etter 2000:

$$\widehat{Inntekt} = 46245.7 + 63.6 * \text{parsei50} + 2917.5 * 1 + 148.1 * \text{parsei50} * 1$$

$$\widehat{Inntekt} = (46245.7 + 2917.5) + (148.1 + 63.6) * \text{parsei50}$$

$$\boldsymbol{\widehat{Inntekt} = 49163.2 + 211.7 * \text{parsei50}}$$

{nb: Selv om tallene er hentet fra modell3, ta hensyn til at det er feilaktig brukt «,» fremfor «.» i den opprinnelige likningen. Dette kan ha skapt forvirring hos studentene}