NTNU
Det skapende universitet

Institutt for sosiologi og statsvitenskap

Sensorveiledning til oppgave og skriftlig eksamen i SOS3003 våren 2016

Karakteren i SOS3003 skal vurderes ut fra en individuelt skrevet oppgave og en seks timers skriftlig eksamen. I den obligatoriske oppgaven får hver student tildelt en avhengig variabel fra datasett European Social Survey, og skal gjennomføre en regresjonsbasert analyse for å forklare hvilke sosiale prosesser og mekanismer som med på å generere verdien på denne avhengige variabelen. På den skriftlige eksamenen skal studentene besvare tre deloppgaver på seks timer, og under eksamen har de ikke tilgang til litteratur eller andre hjelpemidler. De to eksamensdelene skal tildeles lik vekt i karaktervurderingen, men det skal ikke offentliggjøres delkarakterer for de to delen. Studentene på master i sosiologi og master i statsvitenskap gjennomførte dette kurset i andre semester av masterstudiet, og undervisningen foregikk fra februar til mai.

Sensorveiledning til eksamensoppgaven i SOS3003 Anvendt statistisk dataanalyse i samfunnsvitenskap

Kravene til innleveringsoppgaven ble lagt ut på it's learning mandag 7. mars 2016, og innleveringsfristen ble satt til fredag 20. mai 2016. Oppgaveteksten var formulert slik:

SOS3003 Anvendt statistisk dataanalyse
Kravene til innleveringsoppgaven i SOS3003 våren 2016

Semesteroppgaven i SOS3003 skal leveres i to eksemplarer med siste frist fredag 20. mai klokka 14:00. Karakteren på semesteroppgaven vil utgjøre 50 prosent av den samlede karakteren i emnet SOS3003, og semesteroppgaven må derfor leveres med kandidatnummer og ikke med navn eller studentnummer.

1. De grunnleggende kravene til semesteroppgaven
Semesteroppgaven skal skrives som et vitenskapelig paper. Den skal være minimum 14 sider og maksimum 20 sider (Times New Roman 12 pt font, 1,5 linjeavstand) inkludert tabeller og eventuelle figurer. Forside og referanseliste skal ikke telles med når du regner antallet sider. Teksten og analyseresultater skal være et individuelt arbeid. Oppgaven skal baseres på studentens egne analyser av data fra European Social Survey (ESS) fra 2014, 2012, 2010, 2008, 2006, 2004 eller 2002, der hver student velger en unik kombinasjon av avhengig variabel, årstall og land. Vi fraråder studentene fra å sammenligne data fra flere år eller fra flere land. Analysene og diskusjonen i paperet må samsvare med de detaljerte krav som er beskrevet nedenfor.
2. Kravene til analysen

Analysen skal vise studentens kompetanse i å estimere og tolke regresjonsmodeller enten basert på minste kvadratsums metode (OLS) eller logistisk regresjon. De som velger å bruke OLS-regresjon må enten velge seg en kontinuerlige variabel, en variabel på ordinvalnivå som kan behandles som en kontinuerlig variabel (minst fem rangerte kategorier), eller konstruere sin egen skala eller indeks basert på flere variabler, og bruke dette som sin avhengige variabel. De som velger logistisk regresjonsanalyse må enten velge en dikotom variabel, eller omkode en kategorisk eller kontinuerlig variabel til to grupper med verdiene 0 og 1.

3. Oppbyggingen av paperet

Papiret bør bestå av følgende deler, i følgende rekkefølge, og med følgende anbefalte lengde:

a) En forside med en beskrivende tittel, et sammendrag (abstract) på maksimalt 100 - 150 ord (sammendraget skal skrives med enkel linjeavstand, og skal vise problemstillingen, presenterer dataene og metodene som brukes, og oppsummere de viktigste funn ene), og opplysninger om kandidatnummer, emnekode og semester.

b) En innledning på 2 - 4 sider som beskriver den overordnede problemstillingen, og hvorfor du mener det er interessant å besvare denne problemstillingen. Dette bør begrunnes vitenskapelig med utgangspunkt i relevant litteratur og tidligere forskning. Her skal du også presentere en eller flere hovedhypoteser for analysen, og forklare hvorfor du forventer den eller de sammenhengene du beskriver i hypotesen(e). Her skal du også diskutere hvilke andre uavhengige variabler du må ha med i modellen som kontrollvariabler.

c) En metodedel på 2 - 3 sider som beskriver datasettet som brukes (utvalg, og populasjon), den avhengige variabelen, og de uavhengige variablene som skal brukes i analysen. Vær mest mulig konkret i denne beskrivelsen, og begrens bruken av store tabeller og grafiske fremstillinger av variablene. Til slutt i metodedelen skal du skrive at de data som er benyttet i denne oppgaven er hentet fra European Social Survey, ESS (European Social Survey 2014), at disse dataene er i anonymisert form er stilt til disposisjon gjennom Norsk samfunnsvitenskapelig datatjeneste (NSD), og at NSD er ikke ansvarlige for analysen av dataene eller de tolkningene som er gjort.

d) En regresjonsanalyse på 8 -10 sider som presentere en startmodell, en diskusjon av hvordan du kan forbedre denne modellen, og en presentasjon av en justert regresjonsmodell. I denne delen må du derfor vise minst to regresjonsmodeller, en innledende modell med de uavhengige variablene som er presentert i innledningen, og en justert modell der du viser hvordan eventuelle polynomer, samspill eller dummykodinger kan forbedre den første modellen. Du skal samtidig vurdere i hvilken grad den justerte modellen tilfredsstiller modellforutsetningene for den estimeringsteknikken du bruker, og eventuelt justere modellen ytterligere slik at den i størst mulig grad kan tilfredsstille disse forutsetningene.

e) Skriv så en avsluttende drøfting av funnene i forhold til det du forventet i innledningen, og en poengtert konklusjon på 1 – 2 sider.

f) Det skal settes inn referanser i teksten, og til slutt skal du vise en alfabetisk referanseliste. Denne bør ha enkel linjeavstand. Du må bruke en av referansestilene APA, Harvard eller Chicago. Referanseføringen bør konsekvent følge den valgte stilen. Du bør begrense bruken av fotnoter, sluttnoter og eventuelle appendiks til et minimum, og det er ikke nødvendig å legge ved innholdsfortegnelse, tabell- eller figurliste.

Generell informasjon til sensorene:

Veiledningen av oppgavene har foregått i PC-øvingene, og har disse har vært ledet av fjerdesemesters masterstudenter. Etter som de fleste studentene har brukt ESS-data fra Norge, så har jeg ikke satt noe krav om at de skal bruke desingvekter i analysen. De studentene som har slått sammen data fra flere naboland har heller ikke blitt pålagt å bruke populsjonsvekter eller kontroll for eventuell klyngestuktur i disse dataene. Analysene skal bedømmes ut fra studentenes forståelse av den regresjonsmodellen (OLS eller logistisk regresjon) som de har valgt å bruke i sin analyse, og de ferdighetene de viser i anvendt dataanalyse.

Sensorveiledning til eksamensoppgaven i SOS3003 Anvendt statistisk dataanalyse i samfunnsvitenskap

Den skriftlige eksamenen ble gitt fredag 10. juni, og bestod av tre delspørsmål.

Generell informasjon:

I vårsemesteret 2016 har vi hatt følgende bøker på pensum:

Skog, Ole-Jørgen (2009). *Å forklare sosiale fenomener. En regresjonsbasert tilnærming*. Oslo: Gyldendal Akademisk.

Christophersen, Knut-Andreas (2013). *Introduksjon til statistisk analyse. Regresjonsbaserte metoder og anvendelser*. Oslo: Gyldendal Akademisk.

Kohler, Ulrich & Frauke Kreuter (2012). *Data Analysis Using Stata* (3. utg) College Station, Tex.: Stata Press.

Engelsktalende studenter har blitt anbefalt å bytte ut bøkene til Skog og Christophersen med den tidligere pensumboka «Regression with Graphics» av Larence Hamilton. Jeg har også inntrykk av at mange av studentene har bytte ut Stata-boka Kohler & Kreuter med Alan Acock (A Gentle Introduction to Stata) eller Tor Midtbø (Stata. En entusiastisk innføring). Selv mener jeg at Midtbøs er den beste, men den dekker dessverre bare en del av det jeg går igjennom i SOS3003. Jeg antar også at mange av studentene har lagt stor vekt på Kristen Ringdals «Enhet og mangfold», som var pensum på bachelorkurset i metode, og at Ringdals beskrivelse av forutsetningene for OLS finnes igjen i besvarelsene.

De tre eksamensoppgavene skal telle en tredjedel av den samlede karakteren på eksamensbesvarelsen.

Oppgave 1

Vedlegg 1 gjengir en loggfil fra statistikkprogrammet Stata, og viser en analyse av data fra det norske utvalget av European Social Survey fra 2014. Beskriv oppbyggingen av modellene og estimatene fra modell 1 og modell 2 i vedlegg 1 (s. 4-10).

Veiledning til sensor:
I den vedlagte loggfilen viser jeg en stegvis oppbygging av en OLS-modell med en skale for institusjonell tillit som avhengig variabel. Skalaen er konstruert som en enkelt additiv indeks basert på reliabilitetsmålet Chronbachs alpha på 0,858. Begge regresjonsmodellene er estimert med de to kontinuerlige uavhengige variablene alder målt i antall år (agea) og utdanningslengde mål i antall år utdanning (eduyrs), og de tre kategoriserte uavhengige variablene kjønn (gndr), politisk interesse (polintr), og region (regions). I modell 1 blir de to kontinuerlige variablene estimert som lineære effekter, mens de tre kategoriserte variablene blir estimert som dymmysett med den første verdien i variabelen som referansekategori. Målet med deloppgave 1 er å avdekke hvor godt studenten har forstått hva som skiller tolkningen av en lineær uavhengig variabel og en dummykodet uavhengig variabel.

Modell 1 har en negativ lineær effekt av alder, som viser at den institusjonelle tilliten svekkes med økende verdi på aldersvariabelen, og en lineær positiv effekt av utdanning, som viser at den institusjonelle tilliten øker med økende utdanningsnivå. Begge disse effektene er statistisk signifikante både på 5% og 1%-nivå. Parameterestimatet for dummyen female viser ingen statistisk signifikant forskjell i institusjonell tillit mellom kvinner og menn. I tolkningen av koeffisientene til de

dummykodede variablene regions og polintr er det viktig at studenten forstår at koeffisientene måler forskjellen mellom gjennomsnittsverdien for hver enkelt dummy og gjennomsnittet for referansekategorien, etter kontroll for de andre uavhengige variablene i modellen. Ut fra variabelpresentasjonen innledningsvis i Statautskriftet så bør alle se at referansekategorien for regionsvariabelen er «Eastern Norway» og at referansekategorien for politisk interesse er «Very interested». Hvis de ønsker å si noe om de kategoriske variablenes samlede effekt på den avhengige variabelen så må de på F-testene under testparm i.regions og testparm i.polintr som begge viser at dummysettene har statistisk signifikant effekt på 5%-nivået på den institusjonelle tilliten.

I modell 2 har jeg utvidet modellen ved å estimere en kurvelineær alderseffekt (agea og c.agea#c.agea), og et samspill mellom kjønn og politisk interesse (i.polintr#c.female). Etter som alder har en statistisk signifikant negativ effekt og alder kvadrert har en statistisk signifikant positiv effekt kan vi anta den institusjonelle tilliten minsker for hver alderskategori blant de yngste informantene men at det negative andregradsleddet fører til at denne nedgangen stadig blir mindre og kanskje til og med bli positiv. De som forstår at det betingede marginalplottet for alder viser denne kurvelineære effekten bør honoreres for dette, og de som bruker Skogs forenklede derivasjonsformel (s. 285) og finner at bunnpunktet i kurven er 63,1 år [-0,0368826/(2*0,0002921)=63,1335159] bør få god karaktermessig uttelling for dette. Modellestimatene viser også at samspillet mellom politisk interesse og kjønn er en statistisk signifikant forbedring av modellen, men det er vanskelig å beskrive effekten av samspillet ut fra koeffisientene. De som ser at de kan bruke det betingede effektplottet for kjønn og politisk interesse i forhold til institusjonell tillit bør honoreres for dette. I dette plottet ser vi tydelig at den institusjonelle tilliten blant menn går jevnt ned for hver nedadgående kategori i politisk interesse, mens det blant kvinnene bare er små forskjeller i institusjonell tillit mellom de som er veldig interessert, ganske interessert og de som ikke er særlig interessert i politikk, mens de kvinnene som ikke er interessert i politikk også har lavere institusjonell tillit.

Oppgave 2

Beskriv de viktigste forutsetningene for OLS-regresjon, og drøft i hvilken grad disse forutsetningene er oppfylt ut fra testene i vedlegg 1 (s. 11-14).

Veiledning til sensor:

De ulike pensumbøkene opererer med litt ulike krav til OLS-modellen. De forutsetningene som trekkes fram i pensumboka til Skog kan oppsummeres slik:

at regresjonskurven er en rett linje

at restleddet er normalfordelt, homoskedastisk og uavhengig mellom observasjonene

at sammenhengen mellom den uavhengige og den avhengige variabelen ikke er spuriøs, som vil si at restleddet i modellen er ukorrelert med de uavhengige variabelen.

Kristen Ringdal (2013) formulerer to hovedtyper av forutsetninger for OLS-modellen:

1) At modellen er riktig spesifisert
    a. Alle relevante x-variabler er tatt med, og irrelevante er eliminert. Alle X-variablene oppfattes som faste, det vil si at de er uten målefeil.
    b. Sammenhengene mellom X-variablene og Y er lineære.
    c. Modellen er additiv, det vil si at det ikke er samspill (statistisk interaksjon) mellom X-variablene.
2) Regresjonsmodellen bygger også på fire forutsetninger om residualene og en om sammenhengen mellom X-variablene.
    a. Residualene har et gjennomsnitt på 0 i populasjonen.

b. Residualene har lik varians for alle X-variablene, homoskedastisitet.
c. Residualene er ukorrelerte med hverandre og med X-variablene.
d. Residualene er normalfordelte.
e. X-variablene må ikke være perfekt korrelerte, verken parvis, eller gruppevis.

Skog har en grundig drøfting av mulige konsekvenser av brudd på disse forutsetningene, og fremhever at kravet om en godt spesifisert modell, som vil si kravet om at modellen ikke har skjulte spuriøse effekter, er det viktigste, mens kravet om normalfordelt residual er den minst viktige av disse forutsetningene. I Statautskriftet har jeg testet modellspesifikasjonen i modell 2 med følgende tester:

linktest
I linktesten estimeres den predikerte verdien av Y (_hat) og den kvadrerte verdien av denne prediksjonen (_hatsq) i forhold til den opprinnelige avhengige variabelen Y. Hvis modellen klarer å beregne omtrent like gode prediksjoner for alle verdiene på den avhengige variabelen så vil _hat få en signifikant koeffisient mens _hatsq ikke blir signifikant. Hvis derimot _harsq blir signifikant så tyder det på at presisjonen for de predikerte verdiene er varierer utover skalaen, og at modellen kanskje mangler en viktig forklaringsvariabel. I Statautskriften er _hat statistisk signifikant mens _hatsq ikke er statistisk signifikant, og det tyder på at modell 2 er godt spesifisert.

ovtest
Stata beskriver «Ramsey's regression spesification error test» som en test på om modellen er riktig spesifiser, men Tor Midtbø hevder i sin bok at dette er en test på linearitetsforutsetningen. I praksis er det likevel vanskelig å fastslå om signifikante avvik i denne testen skyldes brudd på linearitetsforutsetningen (Skogs første gruppe) eller manglende forklaringsvariabel (Skogs tredje gruppe), så vi bør godta begge disse beskrivelsene. Resultatet av ovtesten viser at det ikke er signifikant forskjell mellom modell 2 og en perfekt modell uten manglende variabler.

hettest
Stata beskriver hettesten som Breusch-Pagan (BP-test) og Cook-Weisbergs test av heteroskedastisitet. I denne testen brukes de kvadrerte residualene som avhengige variabel, og nullhypotesen om homoskedastisitet avvises dersom forklaringsvariabelen påvirker residualene. I Statautskriftet ser vi at kjikvadratet er statistisk signifikant, og det tyder på at det er heteroskedastisitet i modell 2. De studentene som har brukt denne testen i sine egne analyser ser sannsynligvis også at et kjikvadrat på rundt 30 tyder på en forholdsvis svak heteroskedastisitet, og at konsekvensene av dette forutsetningsbruddet får mindre innvirkning på modellen jo større utvalg vi har. Grunnen til dette er både at standardfeilene blir mindre jo for hver enhet og sannsynligheten for at forskjellen mellom den estimerte modellen og en perfekt modell er statistisk signifikant øker jo større utvalget er. Jeg har derfor lagt til tre residualplot som alle tyder på at residualene har omtrent samme varians for alle de predikerte verdiene (rvfplot) og for alle verdiene på de to kontinuerlige variablene (rvpplot agea og rvpplot eduyrs).

residualen
Histogrammet viser fordelingen til residualen fra modell 2 i forhold til den teoretiske normalfordelingen. Her ser vi at det er forholdsvis små avvik mellom residualen og normalfordelingen, men at residualen har en svak tendens til å være venstreskjev. Skog argumenterer for at kravet om normalfordelt residual er det minst alvorlige bruddet på forutsetningene i OLS, og at dette kravet kanskje ikke burde ha vært med blant forutsetningene.

vif
Variance Inflatin Factor (VIF) tester om det er multikolinearitet mellom de uavhengige variablene i modell 2. Kravet om at vi ikke kan ha full multikolinearitet, som vil si at vi estimerer effekten av den samme uavhengige variabelen to ganger i samme modell, er nevnt som en av forutsetningene for OLS i Ringdals bok, mens Skog ikke inkluderer dette i beskrivelsen av forutsetninger. I Stata er det også umulig å estimere en modell med full multikolinearitet. Skog likevel flere omfattende drøftinger av kolinearitetsproblemer knyttet til sterk korrelerte uavhengige variabler, både når det gjelder det at kolinearitet mellom to uavhengige variabler vil øke verdien på koeffisientene og standardfeilene for disse variablene, og hvordan vi må ta hensyn til kolinearitetene når vi skal tolke effektene av koeffisientene. I Statautskriften ser vi at det er forholdsvis høy kolinearitet mellom alder og alder kvadrert og mellom kjønn og samspillet mellom kjønn og politisk interesse. Det viktige med VIF-tabellen er at vi tar hensyn til dette under tolkningen ved at vi ikke kan tolke effekten av alder og effekten av alder kvadrert uavhengig av hverandre, og at kjønnseffekten som måles i koeffisienten for female ikke kan tolkes uavhengig av de andre variabelleddene der kjønn inngår.

Oppgave 3
Vi foretrekker ofte å bruke OLS-regresjon når vi skal analysere en kontinuerlig avhengig variabel i et sannsynlighetsutvalg. Hva er problemene med å bruke OLS-regresjon når vi skal analysere data med en dikotom (todelt) avhengig variabel, strukturerte data med klyngestruktur og dermed avhengighet mellom observasjonene, og når vi har tidsseriedata som for eksempel månedlige registreringer av andelen arbeidsløse i forhold til den totale arbeidsstyrken?

Dette er den klart vanskeligste oppgaven på denne eksamenen. Temaet drøftes ikke samlet i pensum, og det er heller ikke gitt på tidligere eksamener. Dette er derfor primært en test på i hvilken grad studentene klarer å drøfte forutsetningene til OLS utover det som går på å drøfte om en OLS-modell er riktig spesifisert. Besvarelsene på oppgave 3 bør derfor brukes for å justere ned den samlede karakteren på besvarelsen, men bør ikke i seg selv være et grunnlag for å gi strykkarakteren F. Det vil ai at de som skal få karakteren A også må ha en god besvarelse av denne deloppgaven, mens de som får karakteren E kan vise store svakheter i denne deloppgaven så lenge de to andre deloppgavene er besvart tilfredsstillende. I vektleggingen av de tre deloppgavene er det likevel viktig at hver oppgave telle en tredjedel av den samlede karakteren.

I oppgaveteksten nevnes det tre modeller som er vanskelig å estimere riktig i OLS.

Modeller med dikotom avhengig variabel.
Under lesingen av logistisk regresjon bør studentene ha fått med seg at de viktigste grunnene til at vi ikke bruker OLS når vi skal estimere en regresjonsmodell med en dikotom (todelt) avhengig variabel er at vi da sannsynligvis vil få problemer med (1) heteroskedastisitet og (2) at modellen vil predikere verdier utenfor intervallet 0 til 1. Dette er også fremstilt som de to viktigste grunnene til at vi ofte foretrekker å bruke logistisk regresjon og ikke OLS når vi har en dikotom avhengig variabel.

Data med klyngestruktur
Problemene med klyngestruktur blir drøftet inngående i kapittel 10 (Flernivåanalyse: ISS og random konstantledd, s. 108-117)) og i kapittel 11 (Flernivåanalyse: Random slope og kryssnivåsamspill, s. 118-129) i Christophersens bok. I tillegg la vi opp en egen firetimers forelesning om flernivåanalyser. Temaet har derimot ikke blitt tatt opp på dataøvingene. Hovedproblemet med strukturerte data med klyngestruktur er at avhengigheten mellom enhetene øker standardfeilene til estimatene, og at OLS-regresjon, som forutsetter at enhetene er tilfeldig trukket, vil undervurdere standardfeilene i sine estimat. Det er flere mulige løsninger på dette problemet, og på forelesningene har vi vist hvordan vi

kan beregne interklassekorrelasjonen ICC (også kalt rho) for å måle eventuell klyngestruktur, og hvordan vi kan estimere ulike flernivåmodeller (random intercept model, random slope model og modeller med kryssnivåsamspill).

Tidsseriedata

Skog presenterer problemet med autokorrelasjon i tidsseriedata allerede i kapitlet om regresjonsanalysens forutsetninger (kap. 9 s. 250-252), og han utdyper dette nærmere i et kapittel 12 om regresjonsanalyse av tidsseriedata (s. 324-248). Problemene med tidsseriedata utdypes enda mer i kapittel 14 (tidsserieanalyse s. 159-168) og kapittel 15 (paneldataanalyse s. 169-178) i Christophersens bok. I tillegg la vi opp en egen firetimers forelesning om analyse av tidsseriedata, men temaet ble ikke behandlet på dataøvingene. Hovedproblemet med tidsseriedata i OLS-modeller er at vi ofte får autokorrelasjon i restleddet. I forelesningene har jeg vist hvordan vi kan måle eventuell autokorrelasjon ved lag 1 med Durbin-Watsons d, og hvordan vi kan forsøke å kontrollere for denne autokorrelasjonen ved å estimere effekten av en lagget avhengig variabel sammen med de uavhengige variablene i OLS-modellen. I tillegge har jeg også vist en del grafiske hjelpemidler i Stata som kan brukes til å identifisere enda mer problematiske autokorrelasjonsmønster.

Trondheim 7.06.2016

Arild Blekesaune

Vedlegg 1

```
. * TASK 1
. * Dependent variable
. tab1 trstprl trstlgl trstplc trstplt trstprt

-> tabulation of trstprl
```

Trust in country's parliament	Freq.	Percent	Cum.
0. No trust at all	19	1.33	1.33
1. 1	16	1.12	2.45
2. 2	29	2.03	4.48
3. 3	51	3.57	8.05
4. 4	79	5.53	13.58
5. 5	149	10.43	24.00
6. 6	184	12.88	36.88
7. 7	321	22.46	59.34
8. 8	338	23.65	83.00
9. 9	142	9.94	92.93
10. Complete trust	101	7.07	100.00
Total	1,429	100.00	

```
-> tabulation of trstlgl
```

Trust in the legal system	Freq.	Percent	Cum.
0. No trust at all	13	0.91	0.91
1. 1	8	0.56	1.47
2. 2	25	1.75	3.22
3. 3	47	3.29	6.50
4. 4	42	2.94	9.44
5. 5	141	9.86	19.30
6. 6	136	9.51	28.81
7. 7	258	18.04	46.85
8. 8	385	26.92	73.78
9. 9	245	17.13	90.91
10. Complete trust	130	9.09	100.00
Total	1,430	100.00	

-> tabulation of trstplc

Trust in the police	Freq.	Percent	Cum.
0. No trust at all	17	1.19	1.19
1. 1	7	0.49	1.67
2. 2	17	1.19	2.86
3. 3	31	2.16	5.02
4. 4	40	2.79	7.82
5. 5	107	7.47	15.28
6. 6	132	9.21	24.49
7. 7	276	19.26	43.75
8. 8	395	27.56	71.32
9. 9	267	18.63	89.95
10. Complete trust	144	10.05	100.00
Total	1,433	100.00	

-> tabulation of trstplt

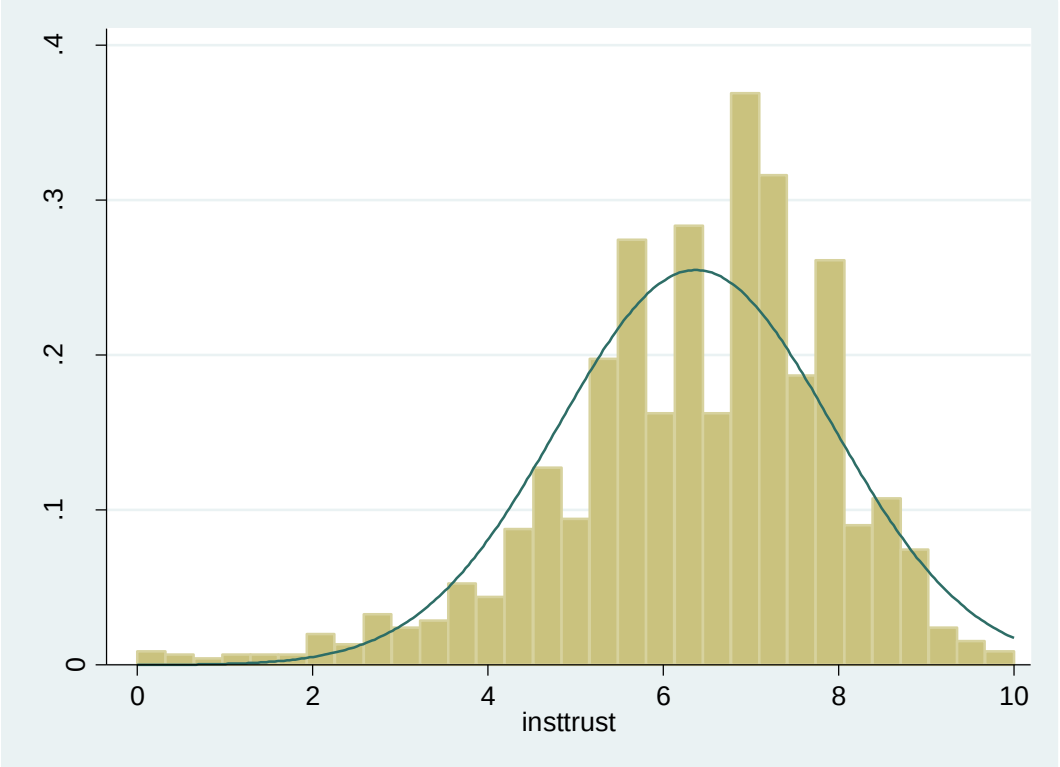
Trust in politicians	Freq.	Percent	Cum.
0. No trust at all	31	2.17	2.17
1. 1	29	2.03	4.20
2. 2	79	5.53	9.73
3. 3	114	7.98	17.70
4. 4	175	12.25	29.95
5. 5	310	21.69	51.64
6. 6	289	20.22	71.87
7. 7	256	17.91	89.78
8. 8	115	8.05	97.83
9. 9	20	1.40	99.23
10. Complete trust	11	0.77	100.00
Total	1,429	100.00	

-> tabulation of trstprt

Trust in political parties	Freq.	Percent	Cum.
0. No trust at all	26	1.83	1.83
1. 1	19	1.34	3.17
2. 2	72	5.07	8.23
3. 3	116	8.16	16.40
4. 4	179	12.60	28.99
5. 5	339	23.86	52.85
6. 6	272	19.14	71.99
7. 7	231	16.26	88.25
8. 8	125	8.80	97.04
9. 9	29	2.04	99.09
10. Complete trust	13	0.91	100.00
Total	1,421	100.00	

```
. alpha trstprl trstlgl trstplc trstplt trstprt

Test scale = mean(unstandardized items)

Average interitem covariance:     2.090183
Number of items in the scale:            5
Scale reliability coefficient:      0.8484

. generate insttrust=(trstprl+trstlgl+trstplc+trstplt+trstprt)/5
(25 missing values generated)

. summarize insttrust
```

Variable	Obs	Mean	Std. Dev.	Min	Max
insttrust	1411	6.371368	1.564346	0	10



```
. * Continous independent variables
. summarize agea eduyrs
```

Variable	Obs	Mean	Std. Dev.	Min	Max
agea	1436	46.76671	18.68344	15	104
eduyrs	1434	13.85216	3.717394	0	30

```
. * Categorical independent variables
. tab1 female regions polintr

-> tabulation of female
```

RECODE of gndr (Gender)	Freq.	Percent	Cum.
Male	764	53.20	53.20
Female	672	46.80	100.00
Total	1,436	100.00	

```
-> tabulation of regions
```

Region	Freq.	Percent	Cum.
Eastern Norway	738	51.39	51.39
Agder and Rogaland	180	12.53	63.93
Western Norway	248	17.27	81.20
Trøndelag	140	9.75	90.95
Northern Norway	130	9.05	100.00
Total	1,436	100.00	

```
-> tabulation of polintr
```

How interested in politics	Freq.	Percent	Cum.
1. Very interested	142	9.89	9.89
2. Quite interested	570	39.69	49.58
3. Hardly interested	595	41.43	91.02
4. Not at all interested	129	8.98	100.00
Total	1,436	100.00	

```
. * Model 1
. regress insttrust female agea eduyrs i.regions i.polintr
```

Source	SS	df	MS		
Model	260.574021	10	26.0574021	Number of obs	= 1410
Residual	3189.81121	1399	2.2800652	F(10, 1399)	= 11.43
				Prob > F	= 0.0000
Total	3450.38523	1409	2.44881848	R-squared	= 0.0755
				Adj R-squared	= 0.0689
				Root MSE	= 1.51

insttrust	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	.0217259	.0817042	0.27	0.790	-.13855	.1820018
agea	-.0087248	.0022769	-3.83	0.000	-.0131914	-.0042583
eduyrs	.0469327	.011629	4.04	0.000	.0241206	.0697448
regions						
Agder and Rogaland	-.1766634	.1268619	-1.39	0.164	-.4255235	.0721967
Western Norway	-.0885928	.112733	-0.79	0.432	-.3097368	.1325512
Trøndelag	.1572054	.1401817	1.12	0.262	-.1177836	.4321944
Northern Norway	-.4279995	.1459855	-2.93	0.003	-.7143735	-.1416254
polintr						
2. Quite interested	-.3296141	.1451707	-2.27	0.023	-.6143898	-.0448384
3. Hardly interested	-.5882118	.1499914	-3.92	0.000	-.882444	-.2939795
4. Not at all interested	-1.259279	.1992371	-6.32	0.000	-1.650115	-.8684434
_cons	6.663423	.2739974	24.32	0.000	6.125932	7.200913

```
. testparm i.polintr

 ( 1)  2.polintr = 0
 ( 2)  3.polintr = 0
 ( 3)  4.polintr = 0

       F(  3,  1399) =   15.48
            Prob > F =    0.0000

. testparm i.regions

 ( 1)  2.regions = 0
 ( 2)  3.regions = 0
 ( 3)  4.regions = 0
 ( 4)  5.regions = 0

       F(  4,  1399) =    3.15
            Prob > F =    0.0138
```

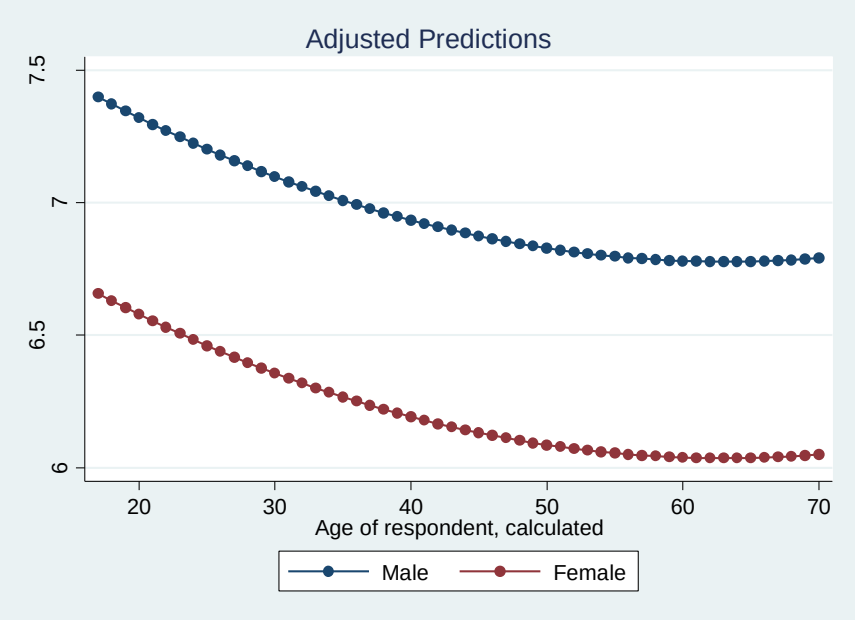
```
. * Model 2
. regress insttrust female agea c.agea#c.agea eduyrs i.regions i.polintr i.polintr#c.female
```

Source	SS	df	MS		
Model	302.513525	14	21.6081089	Number of obs	= 1410
Residual	3147.87171	1395	2.25653886	F(14, 1395)	= 9.58
				Prob > F	= 0.0000
Total	3450.38523	1409	2.44881848	R-squared	= 0.0877
				Adj R-squared	= 0.0785
				Root MSE	= 1.5022

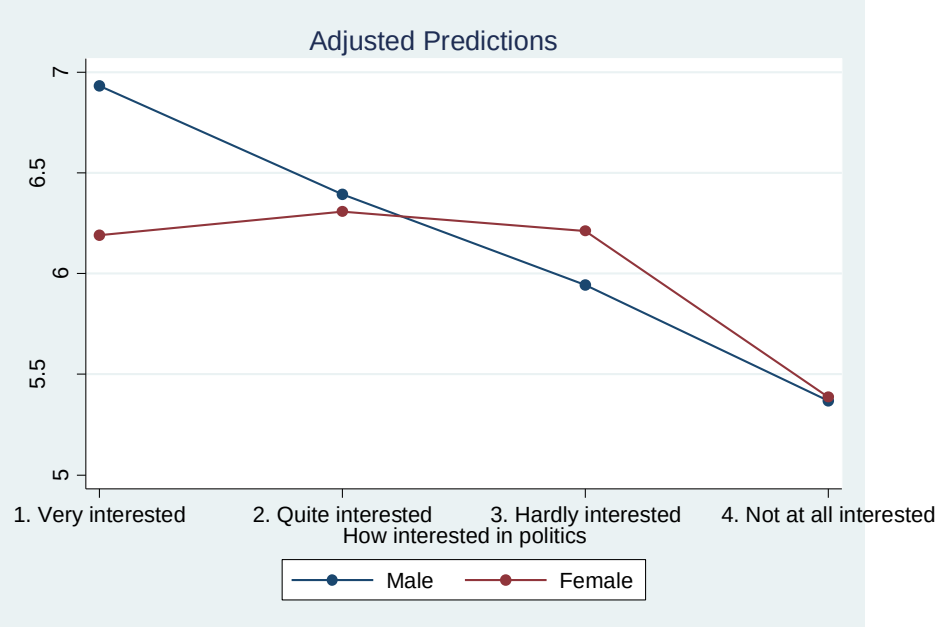
insttrust	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.7422764	.2708637	-2.74	0.006	-1.27362	-.2109323
agea	-.0368826	.0112304	-3.28	0.001	-.0589129	-.0148522
c.agea#c.agea	.0002921	.0001155	2.53	0.012	.0000655	.0005187
eduyrs	.0560367	.0121948	4.60	0.000	.0321145	.0799589
regions						
Agder and Rogaland	-.1656987	.1263012	-1.31	0.190	-.4134594	.082062
Western Norway	-.0888885	.112281	-0.79	0.429	-.3091462	.1313692
Trøndelag	.1803555	.1398191	1.29	0.197	-.0939228	.4546337
Northern Norway	-.4323831	.1452908	-2.98	0.003	-.717395	-.1473711
polintr						
2. Quite interested	-.5406308	.1805272	-2.99	0.003	-.8947648	-.1864967
3. Hardly interested	-.9912959	.1871873	-5.30	0.000	-1.358495	-.624097
4. Not at all interested	-1.565529	.2589123	-6.05	0.000	-2.073429	-1.05763
polintr#c.female						
2. Quite interested	.6571547	.2991533	2.20	0.028	.0703158	1.243994
3. Hardly interested	1.012231	.2981446	3.40	0.001	.4273709	1.597091
4. Not at all interested	.7611896	.3862881	1.97	0.049	.0034213	1.518958
_cons	7.381691	.3366289	21.93	0.000	6.721338	8.042045

```
. testparm agea c.agea#c.agea

 ( 1)  agea = 0
 ( 2)  c.agea#c.agea = 0

       F(  2,  1395) =   11.18
            Prob > F =    0.0000

. testparm i.polintr#c.female

 ( 1)  2.polintr#c.female = 0
 ( 2)  3.polintr#c.female = 0
 ( 3)  4.polintr#c.female = 0

       F(  3,  1395) =    4.21
            Prob > F =    0.0057
```

```
. * Conditional effect plot from Model 2
. quietly: margins, at(eduyrs=(10) regions=(1) polintr=(1) agea=(17/70) female=(0 1))

. marginsplot, noci

  Variables that uniquely identify margins: agea female
```




```
. do "C:\Users\arible\AppData\Local\Temp\STD0e000000.tmp"

. quietly: margins, at(eduyrs=(10) agea=(40) regions=(1) polintr=(1 2 3 4) female=(0 1))

. marginsplot, noci

  Variables that uniquely identify margins: polintr female
```



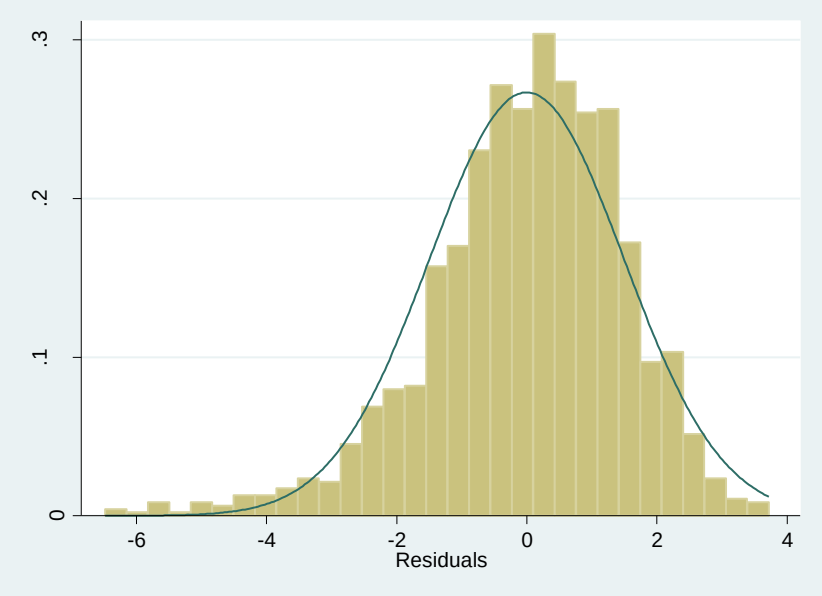
```
. * TASK 2
. * Link test for model specification
. linktest
```

Source	SS	df	MS		
Model	306.600519	2	153.30026	Number of obs	= 1410
Residual	3143.78471	1407	2.23438857	F(2, 1407)	= 68.61
				Prob > F	= 0.0000
				R-squared	= 0.0889
				Adj R-squared	= 0.0876
Total	3450.38523	1409	2.44881848	Root MSE	= 1.4948

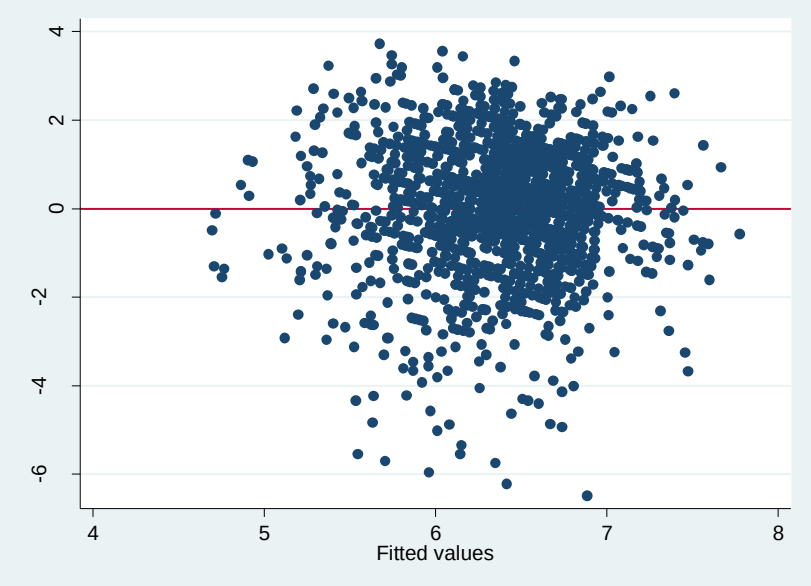
insttrust	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_hat	3.066591	1.530443	2.00	0.045	.0643943	6.068787
_hatsq	-.1645522	.1216692	-1.35	0.176	-.4032248	.0741204
_cons	-6.451812	4.801933	-1.34	0.179	-15.87153	2.967907

```
. * Ramsey's regression specification error test
. ovtest

Ramsey RESET test using powers of the fitted values of insttrust
       Ho:  model has no omitted variables
                 F(3, 1392) =      1.44
                   Prob > F =      0.2304

. * Breusch-Pagan (1979) and Cook-Weisberg (1983) test for heteroskedasticity
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of insttrust

         chi2(1)      =    32.26
         Prob > chi2  =   0.0000
```

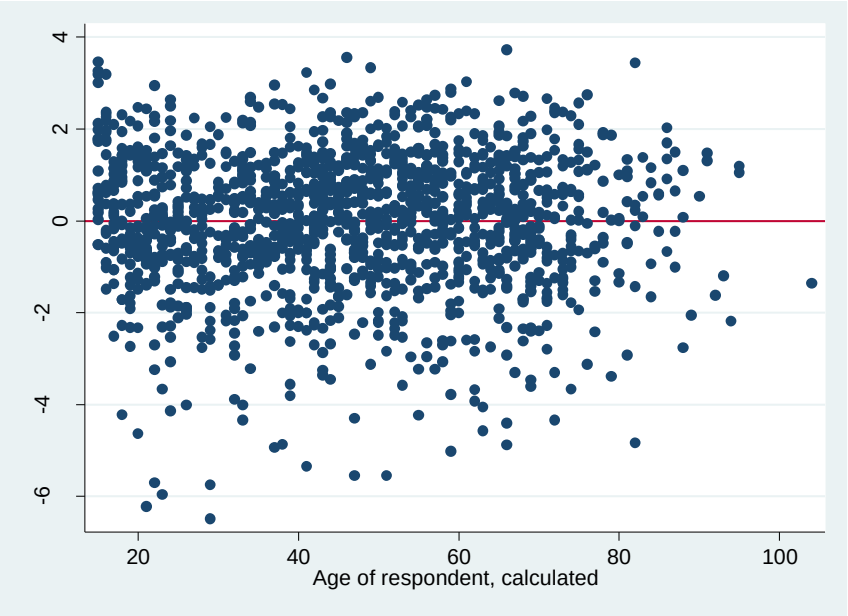
```
. * Tests of residual from Model 2
. predict residual, residual
(26 missing values generated)

. histogram residual, normal
(bin=31, start=-6.4846292, width=.32928309)
```



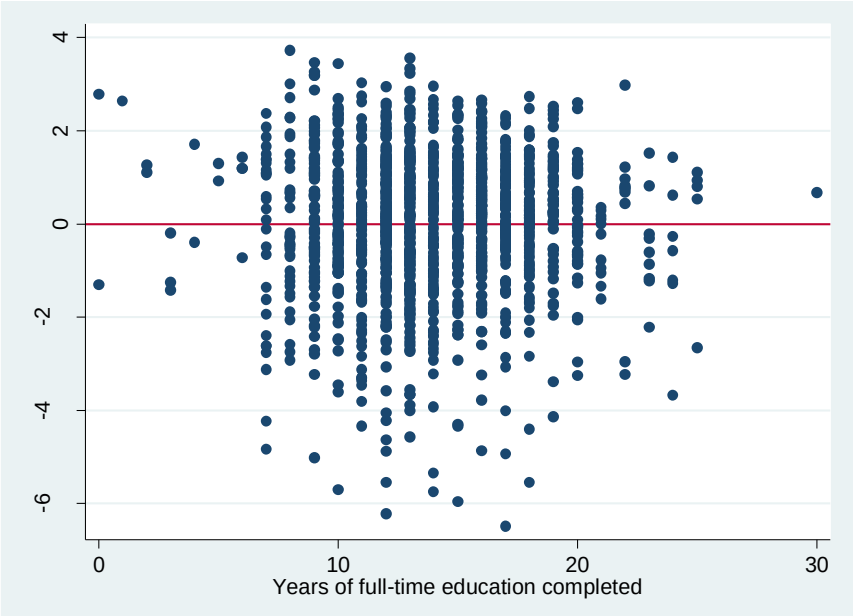

```
. * Residual-versus-fitted plot from Model 2
. rvfplot, yline(0)
```



```
. * Residual-versus-predictor plots from Model 2
. rvpplot agea, yline(0)
```




```
. rvpplot eduyrs, yline(0)
```



```
. * Test of collinearity in Model 2
. vif
```

Variable	VIF	1/VIF
female	11.41	0.087626
agea	27.18	0.036798
c.agea#		
c.agea	27.38	0.036521
eduyrs	1.28	0.781119
regions		
2	1.09	0.913926
3	1.11	0.898778
4	1.09	0.921232
5	1.08	0.925029
polintr		
2	4.89	0.204610
3	5.33	0.187781
4	3.26	0.306610
polintr#		
c.female		
2	8.21	0.121833
3	9.33	0.107228
4	3.74	0.267505
Mean VIF	7.60	

```
.
end of do-file
```