NTNU
Det skapende universitet

Institutt for sosiologi og statsvitenskap

Sensorveiledning til oppgave og skriftlig eksamen i SOS3003 våren 2017

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, på grunnlag av prioriterte ønsker, en avhengig variabel fra datasettet European Social Survey, og skal gjennomføre en regresjonsbasert analyse for å forklare hvilke sosiale prosesser og mekanismer som er med på å generere verdien på denne avhengige variabelen. Analysen bør derfor vurderes ut fra hva studenten har klart å gjøre med sin avhengige variabel, og at sensor må ta hensyn til at noen har fått tildelt mer problematiske variabler enn andre. Det er derfor ikke et mål å estimere en perfekt modell, men å gjøre det beste ut av den variabelen de har valgt. På den skriftlige eksamenen skal studentene besvare tre deloppgaver på seks timer, og studentene blir oppfordret til å ta med kalkulator på eksamen, men de har ikke tilgang til pensumlitteraturen eller andre skriftlige 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 Blackboard learn onsdag 1. mars 2017, og innleveringsfristen ble satt til tirsdag 16. mai 2017. Oppgaveteksten var formulert slik:

SOS3003 Anvendt statistisk dataanalyse
Kravene til innleveringsoppgaven i SOS3003 våren 2017

Semesteroppgaven i SOS3003 skal leveres i pdf-format på epost til ISS innen klokka 14:00 tirsdag 16. mai. I tillegg må alle levere inn et arkiveringseksemplar av semesteroppgaven utskrevet på papir til instituttkontoret ved ISS før 30. mai. 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 lineær regresjon eller logistisk regresjon. De som velger å bruke lineær OLS-regresjon må enten velge seg en kontinuerlige variabel på intervall- eller forholdstallsnivå, 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 variabel med kun to verdier, 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 funnene), 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 (årstall), at disse dataene er stilt til disposisjon i anonymisert form gjennom Norsk senter for forskningsdata (NSD), og at NSD er ikke ansvarlige for analysen av dataene eller de tolkningene som er gjort.

d) En analysedel med en regresjonsanalyse på 8 -10 sider som presenterer 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 transformasjoner, 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 dine 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 på PC-øvingene, og har disse har vært ledet av en fjerdesemesters masterstudent og en stipendiat i statsvitenskap. I dette semesteret har vi for første gang vist hvordan vi kan bruke designvekene (dweight) og poststratifiseringsvektene (pspweght) i

ESS7. Vektig er også beskrevet på side 331 til 333 i den nye pensumboka til Mehmetoglu & Jakobsen (2017). Alle datasettene i ESS har designvekter men det er bare ESS6 og versjon 2 av ESS7 som har poststratifiseringsvekt. I de nordiske utvalgene av ESS6 og ESS7, som de aller fleste bruker, er designvekten satt til 1 for alle enhetene, og designvekting vil derfor ikke påvirke resultatene for disse utvalgene. De som analyserer data fra et av de nordiske landene bør derfor bruke poststratifiseringsveken pspweght, som vekter opp for skjevheter mellom nettoutvalgene og populasjonene for en del sentrale variabler som kjønn, aldersgruppe, utdanningsnivå og region. De har brukt vektig bør honoreres for dette.

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 tirsdag 30. mai, og bestod av tre delspørsmål.

Generell informasjon:

I vårsemesteret 2017 har vi hatt følgende bøker på pensum:

Skog, Ole-Jørgen (2004). *Å forklare sosiale fenomener. En regresjonsbasert tilnærming*. Oslo: Gyldendal Akademisk.
Midtbø, Tor (2012). *Stata. En entusiastisk innføring*. Oslo: Universitetsforlaget.
Mehmetoglu, Mehmet og Jakobsen, Tor Georg (2017). *Applied Statistics Using Stata. A Guide for the Social Sciences*. Sage Publication Ltd.

Engelsktalende studenter har blitt anbefalt å bytte ut boka til Skog med den tidligere pensumboka «Regression with Graphics» av Larence Hamilton. Jeg har også inntrykk av at mange av studentene har brukt Alan Acock (A Gentle Introduction to Stata). 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 tolk estimatene fra modell 1 og modell 2 i vedlegg 1, og drøft i hvilken grad forutsetningene for OLS-regresjon er oppfylt ut fra testene i vedlegg 1 (s. 5-16).

Veiledning til sensor:

Målet med deloppgave 1 er å avdekke studentens kunnskaper og ferdigheter når det gjelder å tolke OLS-modeller, og er en konkret test av om studenten har forstått hva som skiller tolkningen av en lineær uavhengig variabel og en dummykodet uavhengig variabel.

I den vedlagte loggfilen viser jeg en OLS-modell med en skala for institusjonell tillit som avhengig variabel. Skalaen er konstruert som en enkelt additiv indeks basert på reliabilitetsmålet Chronbachs alpha på 0,858. De to første OLS-modellene er estimert med de to kontinuerlige uavhengige variablene alder målt i antall år (agea) og samlet husholdsinntekt målt i desiler (hinctnta). Labelen for variabelen hinctnta ble dessverre ikke presentert i oppgaveteksten, men alle fikk muntlig beskjed om hva variabelen målte under eksamen. Modell 1 har i tillegg de sju kategoriserte uavhengige variablene kjønn (female), utdanningsnivå (edlvdno), politisk interesse (polintr), partivalg (prtvtbno), hovedaktivitet (mnactic), region (landsdel) og beskrivelse av bosted (domicil). Kodingen på de fleste uavhengige variablene er presentert i tabellene, mens dummyen female kun er presentert som en dummykoding av variabelen gndr, og da bør alle skjønne at female er kodet med 1 for kvinne og 0 for menn. I modell 1 er alle dummysettene estimert med den første kategorien som referansekategori. I presentasjonen av de uavhengige variablene har tabellene ulik N, og og de som problematiserer dette vil kunne se at 6,27 prosent av utvalget blir utelatt fra modell 1 på grunn av missing på en eller flere variabler. De som drøfter dette som et mulig skjevhetsproblem bør honoreres.

Modell 1 viser at verken alder, husholdsinntekt eller kjønn har statistisk signifikant effekt på institusjonell tillit. For å se den samlede effekten av hvert enkelt dummysett bør vi først se på testparm-estimatene for effektene for hvert av de seks dummysettene. Her ser vi at heller ikke variablene utdanningsnivå, landsdel og bosted har statistisk signifikant effekt på p lik 0,05, men at landsdel og utdanningsnivå ligger helt på grensen til statistisk signifikant. De som drøfter disse grensetilfellene i forhold til mulige feil av type I og type II bør honoreres. Dummysettene for politisk interesse, partivalg og hovedaktivitet er alle statistisk signifikante, og vi må forvente at studentene også kommenterer enkeltkoeffisientene fra disse dummysettene. Her ser vi at den institusjonelle tilliten synker jevnt med synkende politisk interesse. For dummyene for partivalg ser vi ingen signifikante koeffisienter, men det har sammenheng med at modell 1 bruker partiet Rødt som referansegruppe, og at Rødt ikke utmerker seg i forhold til de andre partiene når det gjelder institusjonell tillit. Hvis noen argumenterer for at kategoriene kodet 66, 77 og 88 bør kodes om til missing, og ikke ser at dette i så fall vil medføre systematiske skjevheter på grunn av at vi da står igjen med et utvalg av partivelgere, så bør trekke på grunn av manglende forståelse. For variabelen hovedaktivitet bruker vi de med betalt arbeid som referanse, og da ser vi at de som er under utdanning har signifikant høyere tillit enn referansegruppen, mens de som er arbeidsløse uten å søke arbeide og gruppen andre har signifikant lavere tillit enn referansegruppen.

Modell 1 har en forklart varians på 17,23 prosent, mens den justerte R^2 viser at denne bare blir 14,17 prosent når vi kontrollerer for antall frihetsgrader i modellen. De som kommenterer dette bør honoreres.

De ulike metodebøkene som brukes i sosiologi og statsvitenskap opererer med litt ulike krav til OLS-modellen. De forutsetningene som trekkes fram i pensumbøkene til Skog (2004) og Midtbø (2012) 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) som de fleste har hatt som pensum på bachelorkurset i metode, formulerer to hovedtyper av forutsetninger for OLS-modellen:

- At modellen er riktig spesifisert
    - 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.
    - Sammenhengene mellom X-variablene og Y er lineære.
    - Modellen er additiv, det vil si at det ikke er samspill (statistisk interaksjon) mellom X-variablene.
- Regresjonsmodellen bygger også på fire forutsetninger om residualene og en om sammenhengen mellom X-variablene.
    - Residualene har et gjennomsnitt på 0 i populasjonen.
    - Residualene har lik varians for alle X-variablene, homoskedastisitet.
    - Residualene er ukorrelerte med hverandre og med X-variablene.
    - Residualene er normalfordelte.
    - X-variablene må ikke være perfekt korrelerte, verken parvis, eller gruppevis.

Både Skog (2004) og Midtbø (2012) har grundige drøftinger 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 (forutsetning 3) og at effektene av de uavhengige variabelen er estimert

slik at de fanger opp de empiriske sammenhengene (forutsetning 1), er det viktigste, mens kravet om normalfordelt residual (forutsetning 2) er den minst viktige av disse forutsetningene.

På side 11 i oppgaveteksten presenteres en del tester som kan brukes for å vurdere eventuelle brudd på forutsetningene til OLS. I linktest blir verdiene på den avhengige variabelen estimert ut fra den predikerte verdien for Y (_hat) og den kvadrerte predikerte verdien for Y (_hatsq) i modell 1, og der viser den signifikante effekten for _hat og den ikke-signifikante effekten for _hatsq at de predikerte verdiene fra modell 1 har en jevn lineær sammenheng med den opprinnelige Y-variabelen som måler institusjonell tillit. Ramseys RESET-test (ovtest), som tester nullhypotesen om at modell 1 ikke mangler viktige forklaringsvariabler, får en ikke-signifikant F-verdi, og det tyder på at modell 1 er riktig spesifisert. I beskrivelsen av denne testen hevder Midtbø (2012) at ovtest ikke er en generell test av modellspesifikasjonen, men av linearitetsforutsetningen, men her er det en tydelig uenighet mellom Midtbøs syn og Stata-dokumentasjonen for ovtest. De som har oppdaget denne uenigheten bør honoreres. Breusch-Pagan og Cook-Weisbergs test for heteroskedastisitet (hettest) viser at hypotesen om at modell 1 har konstant varians (homoskedastisitet) må forkastes, men her vil de som har forstått denne testen sannsynligvis konkludere med at et kjikvadrat på under 50 fra et utvalg på 1346 enheter er en forholdsvis svak heteroskedastisitet. Histogrammet med residualen tyder heller ikke på at modellen bryter forutsetningen om normalfordelt residual. Samlet sett kan vi konkludere med at modell 1 sannsynligvis ikke har alvorlige brudd med forutsetningene for OLS, og at parameterestimatene sannsynligvis er forventingsrette. Men kan modell 1 forenkles?

I utskriften for modell 2 har det dessverre sneket seg inn en feil ved at det står OLD-modell og ikke OLS-modell, men alle studentene ble informert om denne feilen i løpet av eksamenen. I modell 2 har vi forenkles modellen ved å fjerne variablene uten signifikant effekt i modell 1, og har da forenklet modellen fra 48 til 35 estimerte parametere ved å fjerne alder, husholdsinntekt, kjønn, region og bosted. På side 12 beregnes den samlede effekten av disse fem variablene, og der viser vi at den samlede effekten av de fjernede variablene ikke har statistisk signifikant effekt i modell 1. Modell 2 estimerer effekten av fire dummysett, og har en forklart varians på 15,77 prosent, mens den justerte R^2 viser 13,61 prosent. Dette tyder på at modell 2 har omtrent samme forklaringskraft som modell 1. Modell 2 har også 57 flere enheter enn modell 1.

F-testene på side 14 viser at all de fire dummysettene i modell 2 har statistisk signifikant effekt på p lik 0,05, og i regresjonskommandoen ser vi at vi har endret referansekategori i alle de fire dummysettene. For utdanningsdummyene har vi brukt kategori 6 (avsluttet videregående yrkesutdanning) som referanse, og da ser vi at de med sjuårig folkeskole, videregående allmennfag og alle gruppene med høyere utdanning har signifikant større institusjonell tillit enn referansegruppen. Dummysettet for politisk interesse har nå de som ikke er politisk interesserte som referanse, og da ser vi igjen at den at den institusjonelle tilliten øker jevnt med økende politisk interesse. I partibarometeret bruker vi FrP som referanse, fordi dette er det partiet som har de mest avvikende verdiene på institusjonell tillit, og dette gjør at nesten alle partiene, unntatt Rødt, Kystpartiet, Andre og de som har svart vet ikke, har høyere institusjonell tillit enn FrP-velgerne. Hvis vi sammenligner de samlede testene for partibarometeret fra modell 1 og modell 2 ser vi at partivalg ikke har noe større effekt i modell 2 enn i modell 1, og at forskjellen i antallet signifikante koeffisienter kun har sammenheng med valget av referansekategori. De som ikke ser dette bør trekkes i karakter. I dummysettet for hovedaktivitet har vi brukt de under utdanning som referansekategori, og da får alle dummyene, unntatt de arbeidsløse som søker etter jobb, signifikant lavere institusjonell tillit enn referansegruppa. Hvis vi sammenligner de samlede testene for hovedaktivitet i modell 1 og modell 2 ser vi at denne variabelen har fått økt forklaringskraft etter at vi fjernet mange av variablene i modell 1. De som ser dette bør honoreres.

På side 15 i oppgaveteksten viser vi de samme testene som vi brukte i modell 1 for å vurdere eventuelle brudd på forutsetningene til OLS. Her viser linktest at de predikerte verdiene fra modell 2 har en jevn lineær sammenheng med den opprinnelige Y-variabelen. Ramseys RESET-test (ovtest), som tester nullhypotesen om at modell 1 ikke mangler viktige forklaringsvariabler, får en statistisk signifikant F-verdi, og det tyder på at modell 2 har utelatt viktige forklaringsvariabler, og at modell 2 dermed har et spesifikasjonsproblem som vi ikke hadde i modell 1. Breusch-Pagan og Cook-Weisbergs test for heteroskedastisitet (hettest) viser også at modell 2 har sterkere heteroskedastisitet enn modell 1. Også histogrammet med residualen tyder på at modell 2 har et større problem med uteliggere på venstre side enn modell 1, men også residualen fra modell 2 er forholdsvis normalfordelt. På side 16 gjengir jeg en metode som jeg har vist på en forelesning, men som ikke blir presentert i pensumlitteraturen, for å kunne vurdere hvor i modellen vi finne de største avvikene mellom predikert Y og Y. Denne testen bygger på Stata rvfplot, som blir presentert i Midtbø (2012, 108), men jeg har brukt absoluttverdien til residualen kombinert med en gradvis estimert regresjonslinje (lowess) for å få fram hvor vi har størst avvik. Diagrammet på side 16 viser at det er prediksjonene rundt verdien 5 som har størst residualverdi, og ut fra den antatte fordelingen for Y-variabelen tillit, som de flinkeste studentene bør kunne se for seg selv om fordelingen ikke er presentert, tyder dette på at modell 2 har størst problem med å predikere riktige verdier for de som har relativt lav institusjonell tillit. Det kan tyde på at modell 2 mangler variabler som forklarer hvorfor noen har litt under gjennomsnittlig høy institusjonell tillit. Samlet sett kan vi ut fra dette konkludere med at modell 1 kommer bedre ut enn modell 2 når det gjelder testene knytet til forutsetningene til OLS.

Oppgave 2
Beskriv estimatene fra modell 3 i vedlegg 1 (s. 17-18), og drøft i hvilken grad modell 3 tilfredsstiller forutsetningene for logistisk regresjon.

Veiledning til sensor:

Målet med deloppgave 2 er å avdekke studentens kjennskap til logistisk regresjon.

I denne oppgaven er den kontinuerlige variabelen for institusjonell tillit delt inn i to grupper, der skillet mellom de to gruppene er satt til 6,5, og det er estimert en logistisk regresjonsmodell med de samme X-variablene som ble brukt i modell 2. På forelesningene har jeg gått igjennom følgende tre metoder for å beskrive effektene i en logistisk regresjonsmodell:

- **Tolkning av de logistiske regresjonskoeffisientene**
    - *Ser på de logistiske koeffisientens fortegn, og på p-verdien for å vurdere om effekten er statistisk signifikant*
- **Tolkning av koeffisientene i oddsskalaen**
    - *Oddsratene er ikke oppgitt i oppgaveteksten, så de som ønsker å anvende denne metoden må bruke egen kalkulator for å beregne oddsraten som en eksponentialfunksjon (e^{coeff}) av den logistiske koeffisienten. Oddsratene vil gi bedre forståelse av styrken på sammenhengen enn de logistiske koeffisientene, men her er det viktig at studentene ser at OR måler forholdet mellom to odds – og da er det viktig at de skiller klart mellom odds, oddsrater og sannsynligheter.*
- **Omregning til sannsynligheter**
    - *På forelesningene har vi for det meste benyttet margins-kommandoene i Stata for å beregne betingede sannsynligheter, men i Mehmetoglu &*

Jakobsen (2017, s. 169-171) står det litt om hvordan vi kan beregne betingede sannsynligheter ut fra regresjonsligningen. De som klarer å beregne riktige betingede sannsynligheter har derfor jobbet grundig med pensumstoffet, og bør derfor honoreres mye.

Modell 3 ble estimert med de samme X-variablene og for det samme antallet enheter som modell 2, og de fleste vil forhåpentligvis sammenligne disse to modellene. McFaddens pseudo R^2 i modell 3 er på 0,09, men den er ikke direkte sammenlignbare med R^2 i modell 2. De som tolker den samlede modellstatistikken (Pseudo R2 og LR chi2) bør får honnør for det. Det blir ikke oppgitt mål for samlet effekt for dummysettene i modell 3, men alle bør kunne beskrive sammenhengene mellom de fire dummysettene og dummyen for institusjonell tillit ut fra p-verdiene (P>[z]) og koeffisientene i tabellen på side 17. For utdanningsdummyene har vi igjen brukt kategori 6 (avsluttet videregående yrkesutdanning) som referanse, og da ser vi at de med sjuårig folkeskole, videregående allmennfag og alle gruppene med høyere utdanning unntatt bachelor fra universitetet har signifikant større institusjonell tillit enn referansegruppen. Dummysettet for politisk interesse har ikke politisk interesserte som referanse, og da ser vi også her at den institusjonelle tilliten øker jevnt med økende politisk interesse. I partibarometeret bruker vi FrP som referanse, og dette gjør at nesten alle partiene, unntatt Rødt, Kystpartiet, MDG, Andre og de som har svart vet ikke, har høyere institusjonell tillit enn FrP-velgerne. I dummysettet for hovedaktivitet har vi brukt de under utdanning som referansekategori, og da får alle dummyleddene, unntatt de to gruppene med arbeidsløse, signifikant lavere institusjonell tillit enn referansegruppa. De som sammenligner modell 3 med modell 2 vil se at det er tre signifikante dummyledd i modell 2 som ikke får signifikant effekt på p lik 0,05 i modell 3. Ellers viser de to modellene omtrent det samme mønsteret, og det kan tyde på at resultatene er forholdsvis robuste. De som vurderer likheten mellom modell 2 og modell 3 bør honoreres.

Skog beskriver de viktigste forutsetningene for logistisk regresjon slik:

- at sammenhengen mellom variablene er S-formet og kan beskrives med den logistiske funksjonen, som vil si en rett linje på logit-skalaen.
- at de enkelte observasjonene er uavhengige av hverandre, men har ingen forutsetninger knyttet til restleddet.
- at sammenhengen mellom den uavhengige og den avhengige variabelen ikke er spuriøs.

På side 18 presenteres resultatene av Hosmer-Lemeshows goodness-of-fit-test for modell 3. I denne testen deles utvalget inn i ti ulike grupper, og så beregnes en kjikvadrattest for å se om modellen har sammen forklaringskraft for alle de ti gruppene. Det ikke-signifikante kjikvadratet på 5,13 med 8 frihetsgrader tyder på at det er liten forskjell mellom gruppene, og at modellen er godt spesifisert. Mehmetoglu & Jakobsen bruker også linktest for å vurdere om den logistiske regresjonsmodellen har omtrent samme forklaringskraft når vi sammenligner predikerte sannsynligheter og med sannsynligheten for å ha verdien 1 på den dikotome tillitsvariabelen, og linktesten viser at det er et klart lineært forhold mellom predikert og observert sannsynlighet.

Hosmer-Lemeshows goodness-of-fit-test og linktest viser derfor begge at modell 3 tilfredsstiller kravet om at sammenhengen mellom variablene er S-formet og kan beskrives med den logistiske funksjonen, som vil si en rett linje på logit-skalaen. Etter som det norske utvalget av ESS7 er basert på sannsynlighetsutvelging vil modellen også tilfredsstille kravet om at de enkelte observasjonene er uavhengige av hverandre. Den tredje forutsetningen kan ikke vurderes ut informasjonen i modell 3, men de som drøfter problemet med spuriøsitet ut fra testene av modell 1 og modell 2 vil i alle fall ha

et empirisk grunnlag for å vurdere dette. Alle bør kunne drøfte problemet med utelatte variabler på et generelt grunnlag.

Oppgave 3
Hva er problemene med å bruke OLS-regresjon når vi skal analysere paneldata med flere registreringstidspunkt for hver enhet, og data som har en klyngestruktur fordi vi har slått sammen tverrsnittsundersøkelser fra flere land i en samlet datamatrise? Beskriv noen metoder som kan brukes for å analysere slike data.

Veiledning til sensor:

Målet med denne deloppgaven er å teste studentenes kjennskap til mer avanserte datamodeller enn OLS og logistisk regresjon.

Grunnlaget for denne oppgaven er kapittel 9 (Multilevel Analysis) og kapittel 10 (Panel Data Analysis) i Mehmetoglu & Jakobsen (2017, 193-267). Det viktigste målet med oppgaven er å skille ut de som kjenner til problemene med å analysere slike data med OLS. De som ikke kjenner til dette bør trekkes i karakter. De som klarer å relatere problemene med paneldata og data med klyngestruktur til forutsetningen om uavhengig mellom observasjonene bør honoreres for dette. Og de som kjenner til de viktigste metodene for å behandle slike data bør honoreres sterkt.

Data med klyngestruktur
Det ble gitt en egen firetimers forelesning om flernivåanalyser, men temaet har 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 intraklassekorrelasjonen 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) der vi kontrollerer for klyngestrukturen i data.

Analyse av paneldata
Det ble også gitt en egen firetimers forelesning om paneldata, men heller ikke dette temaet ble fulgt opp med egen dataøving. Problemene med paneldata er at vi både har klyngestruktur og tidsserier i dataene. I dette semesteret har vi lagt hovedvekt på hvordan vi kan analysere paneldata med OLS-, between- og within-modeller (fixed effects), og hvordan vi tolker de estimerte parameterne fra disse tre modellen. I tillegg har vi vist hvordan vi kan analysere paneldata med lange tidsserier med Time-Series Cross-Section-Methods (TSCS), der vi kontrollerer for autokorrelasjon med en lagget avhengig variabel. På forelesningene har jeg ikke vist hvordan vi kan analysere paneldata med Random-Effekt-modeller, men også dette er dekket i pensumboka til Mehmetoglu & Jakobsen (2017).

Trondheim 30.05.2017

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

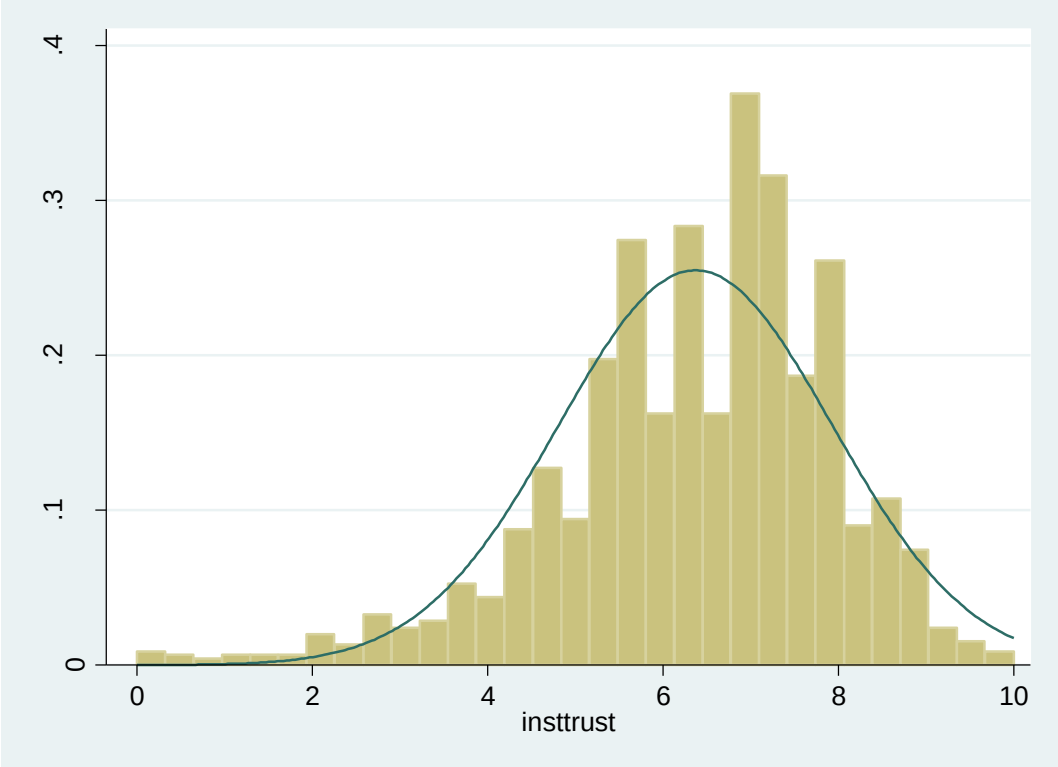
Trust in politicians	Freq.	Percent	Cum.
0. No trust at all	31	2.17	2.17
1. 1	29	2.03	4.20
2. 2	79	5.53	9.73
3. 3	114	7.98	17.70
4. 4	175	12.25	29.95
5. 5	310	21.69	51.64
6. 6	289	20.22	71.87
7. 7	256	17.91	89.78
8. 8	115	8.05	97.83
9. 9	20	1.40	99.23
10. Complete trust	11	0.77	100.00
Total	1,429	100.00	

-> 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
```

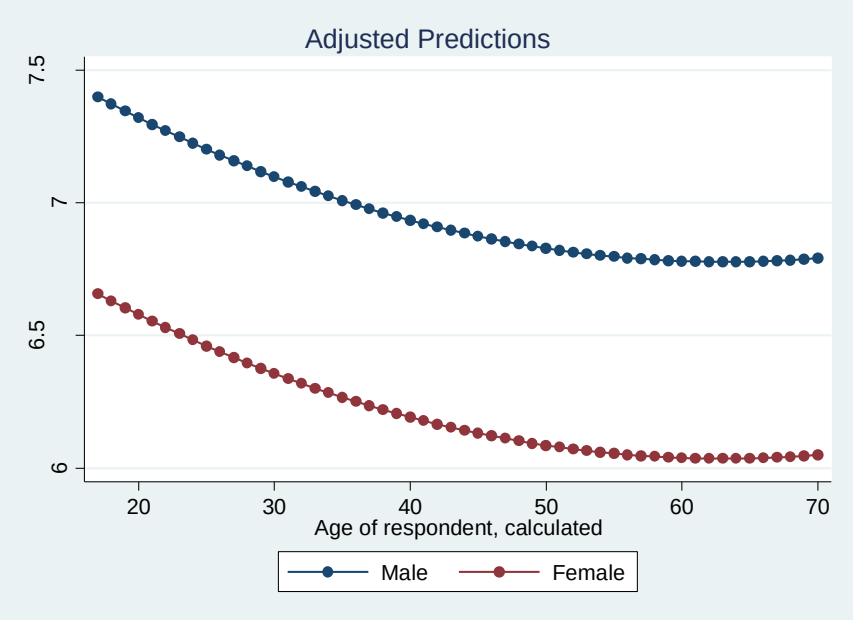
```
. * Model 2
. regress insttrust female agea c.agea#c.agea eduyrs i.regions i.polintr i.polintr#c.female
```

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

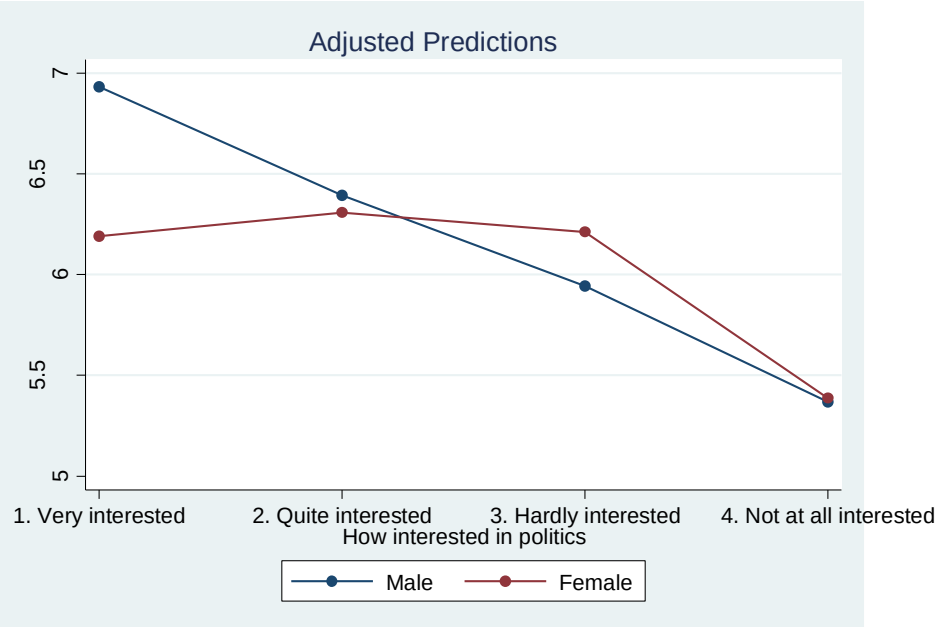
insttrust	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.7422764	.2708637	-2.74	0.006	-1.27362	-.2109323
agea	-.0368826	.0112304	-3.28	0.001	-.0589129	-.0148522
c.agea#c.agea	.0002921	.0001155	2.53	0.012	.0000655	.0005187
eduyrs	.0560367	.0121948	4.60	0.000	.0321145	.0799589
regions						
Agder 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
```



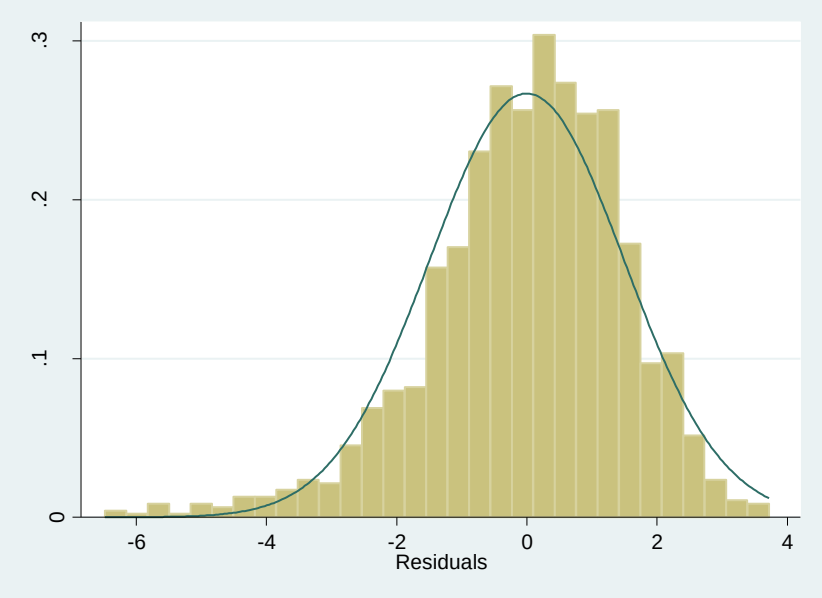
```
. * TASK 2
. * Link test for model specification
. linktest
```

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

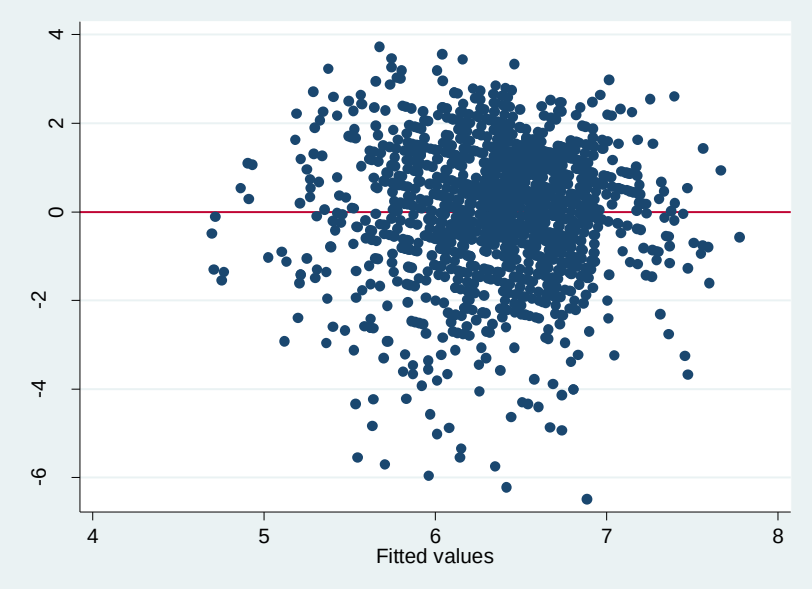
insttrust	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_hat	3.066591	1.530443	2.00	0.045	.0643943	6.068787
_hatsq	-.1645522	.1216692	-1.35	0.176	-.4032248	.0741204
_cons	-6.451812	4.801933	-1.34	0.179	-15.87153	2.967907

```
. * 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
```

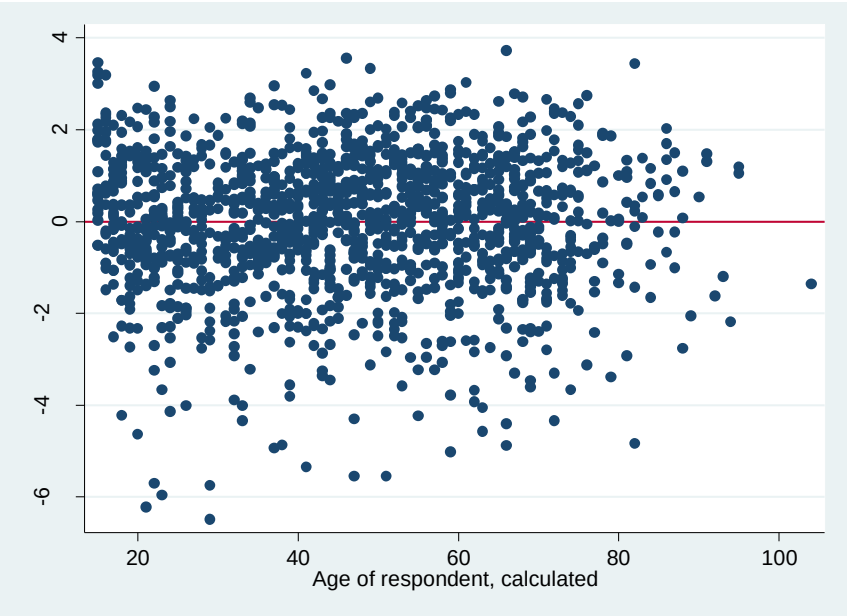
```
. * Tests of residual from Model 2
. predict residual, residual
(26 missing values generated)

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



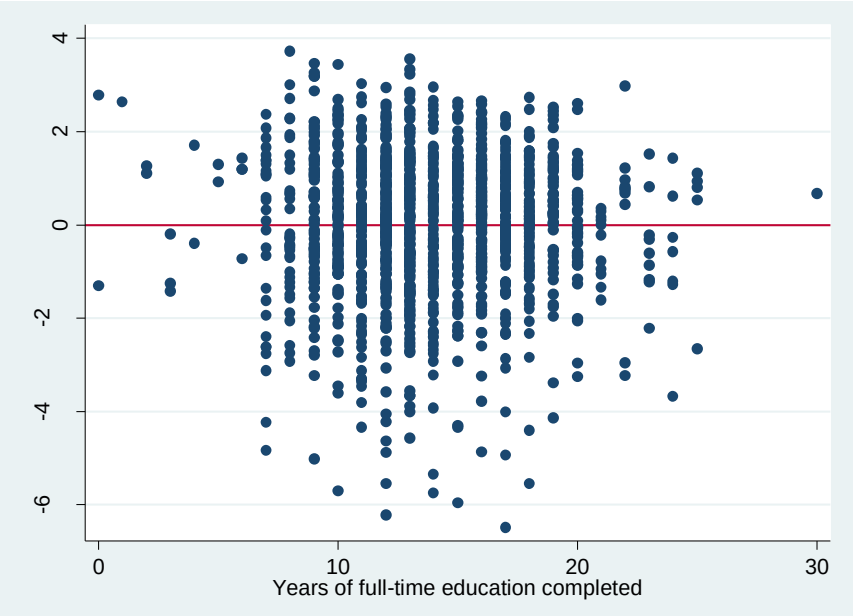

```
. * Residual-versus-fitted plot from Model 2
. rvfplot, yline(0)
```



```
. * 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
```