Forutsetninger OLS

Louisa Boulaziz - louisabo@uio.no

September 3, 2020

Her følger kort oppsummering regresjonsdiagnostistikk. Dette er bare veiledende og er ment som en introduksjon for de som skal bruke R for å løse oppgaver. Selv om jeg skriver hva forutsetningene for OLS består må du bruke Christophersen kap. 7 for å få en bedre og mer fullstendig gjennomgang av forutsetningene.

Forutsetninger/kritiske aspekter ved OLS-vurdering består av:

1. Ingen utelatt variabel skjevhet (no omitted variable bias, OVB)
2. Linær sammenheng mellom variablene
3. Ingen autokorrelasjon - observasjonen skal være uavhengige
4. Normalfordelte residualer
5. Homoskedastiske residualer
6. Ingen perfekt multikolinearitet
7. Manglene opplysninger (missing verdier, NA)

Nummer 1) Den er en teoretisk forutsetning.

Nummer 2) OLS forutsetter at sammenhengen er lineær. Det finnes riktig nok tweaks man kan gjøre, men da er det ikke standard OLS-regresjons. I R kan lineæritets-antagelsen kan testes grafisk. Bruk pakken car og kjør koden veresPlot(). Denne plotter residualene til den avhengige variabelen

mot residualene til den uavhengige variabelen vi er interessert i.

Nummer 3) Uavhengighet/ingen autokorrelasjon. Denne antagelsen testes enkelt med en Durbin-Watson test. Koden er durbinWatsonTest(). (Funker imdilertid kun på tidssereiedata). Pdwtest fra pakken plm funker på panel- og tidsseriedata.

Denne antagelsen holder dersom vi har et tilfeldig utvalg fra en populasjon, på et tidspunkt. Da vil observasjonene være statistisk uavhengige (alle observasjonene er trukket tilfeldig), og likt distribuert (alle observasjonene er trukket fra samme populasjon). Eks. På data med autokorrelasjon er statsbudsjettet overtid eller partiprogrammer til politiske partier over tid. Eller paneldata med land-år som observasjoner.

Nummer 4) Normalfordelte residualer. Residualene fra modellen er normalfordelt og har gjennomsnitt tilnærmet lik 0.

Koden er qqPlot(). Dette plottet plotter studentiserte residualer fra regresjonen vår mot kvantiler fra den kumulative normalfordelingen. Studentiserte residualer er en alternativ måte å standardisere på i beregning av varians for hver enkelt oberservasjon. Formålet med dette er at vi får statistisk uavhengighet mellom teller og nevner, noe som lar oss bruke residualene til statistiske tester.

Nummer 5) Homoskedastiske residualer. Variansen til residualene skal være konstante for ulike nivå av uavhengig variabel.

Vi kan teste for heteroskedastisitet ved hjelp av plot av studentiserte residualer mot standardiserte predikerte verdier fra modellen. Dette kan gjøres med spreadLevelPlot().

Nummer 6) Ingen perfekt multikolinearitet. Det skal ikke være en perfekt lineær sammenheng mellom et sett av de uavhenige variablene. Dette fører til at regresjonene ikke lar seg estimere og skyldes som regel at man har lagt inn dummyvariable for alle kategorier av en variabel. Vif() tester for multikolinearitet og er også en del av pakken car.

Nummer 7) Dette er hva du vil si om missing verdier. I R fjernes f.eks

disse automatisk når du kjører regresjon, men man må legge på na.rm=T i andre tilfeller. Mange missing verdier kan være problematisk, se Christophersen for hvordan du kan håndtere missing-verdier. Missing knyttes i første om gang til utvalget ditt.

Ekstra: inflytelsesrike obeservasjoner (plotte alle observasjonene og reglinjen). Man kan også bruke influenceIndexPlot(). Denne koden kombinerer informasjon om uteliggere og innflytelsesrike observasjoner.