Hvorfor OLS er den bedste estimator

- Gauss-Markov-teoremet og forudsætninger for OLS

Benjamin Egerod

22 februar 2016

Indledning – Antagelser er til for at blive brudt

"Alle statistiske modeller er forkerte – men nogle er brugbare" er en almindelige talemåde blandt folk, der arbejder med statistik[1]. Vi har igennem nogle uger arbejdet med OLS-estimatoren, som netop antager, at man kan beskrive virkeligheden med en meget simpel statistisk model. For at OLS skal kunne fungere, skal man derfor være villig til at gøre sig fem antagelser om, hvordan virkeligheden ser ud:

1) Data kommer fra en tilfældigt udvalgt stikprøve.
2) Sammenhængen, man vil undersøge, er lineær.
3) Der er ingen perfekt korrelation mellem ens uafhængig variable.
4) De uafhængige variable er eksogene.
5) Fejlledet er homoskedastisk.

Antagelserne vil – som talemåden siger – aldrig være opfyldt perfekt. Alligevel vil resultater fra OLS i langt de fleste tilfælde stadig fortælle os noget om virkeligheden, der gør os klogere substantielt set. At OLS er den mest udbredte teknik til at estimere sammenhængen mellem variable på skyldes langt hen ad vejen nogle utroligt brugbare egenskaber ved disse fem antagelser. For det første er antagelserne ofte ikke specielt urealistiske – og i situationer, hvor de ikke holder, kan man typisk håndtere det. For det andet har det vist sig, at hvis antagelserne tilnærmelsesvist er opfyldt, vil det være umuligt at finde på en bedre estimator. Det er netop, hvad det såkaldte Gauss-Markov-teorem siger – og emnet for det her undervisningsnotat.

I et senere afsnit vil jeg behandle antagelserne og herunder komme med en mere uddybende beskrivelse af, hvad de betyder. Det er vigtigt, at bemærke, at det ikke er præcis de samme antagelser, som vi er blevet præsenteret for i Agresti & Finlay (2014)[2]. Gauss-Markov-teoremet er dog den klassiske måde at introducere OLS på. På den måde skal dette notat ses som et supplement til Agresti & Finlay (2014). Formålet er ikke bare at lære at diagnosticere brud på antagelserne bag OLS, men også at sætte dem ind i en bredere forståelsesmæssig ramme. Desuden giver notatet et statistisk framework, hvori vi kan forankre en lang række af de koncepter, som vi møder i kursuslitteraturen på Metode 2 – f.eks. eksogenitet og kausalitet (King, Keohane & Verba 1994; Dunning 2008) og nødvendigheden af variation på de uafhængige variable (Gerring 2004).

Hvad det betyder, at OLS er 'bedst', er indtil videre rimelig uklart. Derfor vil jeg i næste afsnit kort skitsere Gauss-Markov-teoremet og relatere det til, hvordan vi bedømmer estimatorer i forhold til hinanden. Jeg bruger lidt tid på at diskutere kriterierne for at udpege den mest optimale estimator i en given situation. Dernæst behandler

[1] Det kommer fra et citat af den britiske statistiker George Box og har sin helt egen Wikipediaside: https://en.wikipedia.org/wiki/All_models_are_wrong, som indeholder nogle præcise og letfordøjelige Box-udsagn

[2] De udelader antagelse 3 og 4 og tilføjer en antagelse om normalfordelte residualer, som vi kommer ind på siden hen

jeg Gauss-Markov-antagelserne hver for sig. Jeg vil særligt fokusere på, hvilke konsekvenser brud på den enkelte antagelse har. Jeg vil løbende relatere det til nogle koncepter, vi har lært på Metode 1. I konklusionen påpeger jeg vigtigheden af at kombinere evnerne til at stille en praktisk diagnose med substantiel forståelse for ens data og den teoretiske baggrund for OLS.

Udover at være et supplement til kursuslitteraturen skal noten også spille sammen med forelæsningerne. Bl.a. for at gøre læsebyrden så lille som muligt vil illustrationerne i det her notat være rimelig abstrakte. Jeg vil komme med mere virkelighedsnære eksempler på forelæsningerne.

Gauss-Markov-sætningen

På Metode 1 brugte vi bl.a. meget tid på at vurdere, hvilken statistisk estimator, der var den korrekte, til at besvare et givent spørgsmål. Tænk bare på alle de overvejelser, der var i spil, når man vurderede, om man skulle bruge en t-test, pearsons r eller sågar en gamma. Og listen stopper ikke ved de estimatorer, vi lærte på Metode 1 – der findes bogstaveligt talt tusindvis af teknikker til at estimere sammenhænge mellem variable. Her på kurset bruger vi OLS i langt de fleste tilfælde. Og det er et generelt mønster – uagtet hvilket felt man bedriver sin forskning på, er OLS klart den mest anvendte estimator, når man laver multivariate statistiske analyse. Men når der findes så mange udmærkede estimatorer, hvordan er det så sket, at OLS har fået en så fremherskende position, som den har? Det er ikke ren vane, og her kommer Gauss-Markov-teoremet[3] ind i billedet.

Teoremet siger i korte træk, at så længe de fem såkaldte Gauss-Markov-antagelser er opfyldt, så vil OLS være den Bedste Lineære Unbiased Estimator (vi siger, at den er BLUE). Det betyder, at ud af alle de måder, som vi kunne udregne en sammenhæng mellem to eller flere variable på, så er OLS mindst lige så god som den næstbedste – hvis de fem antagelser er opfyldt. Det er et utroligt stærkt resultat. Det betyder, at vi i langt de fleste tilfælde kan tage OLS som vores udgangspunkt, undersøge om antagelserne i rimelig grad er opfyldt og finde måder at håndtere eventuelle brud på. Nogle gange vil man bruge andre estimatorer, men på Metode 2 fokuserer vi typisk på, hvordan vi kan 'rette op' på brudene på antagelserne (undtagelsen er ved binære afhængige variable, som vi behandler senere på kurset).

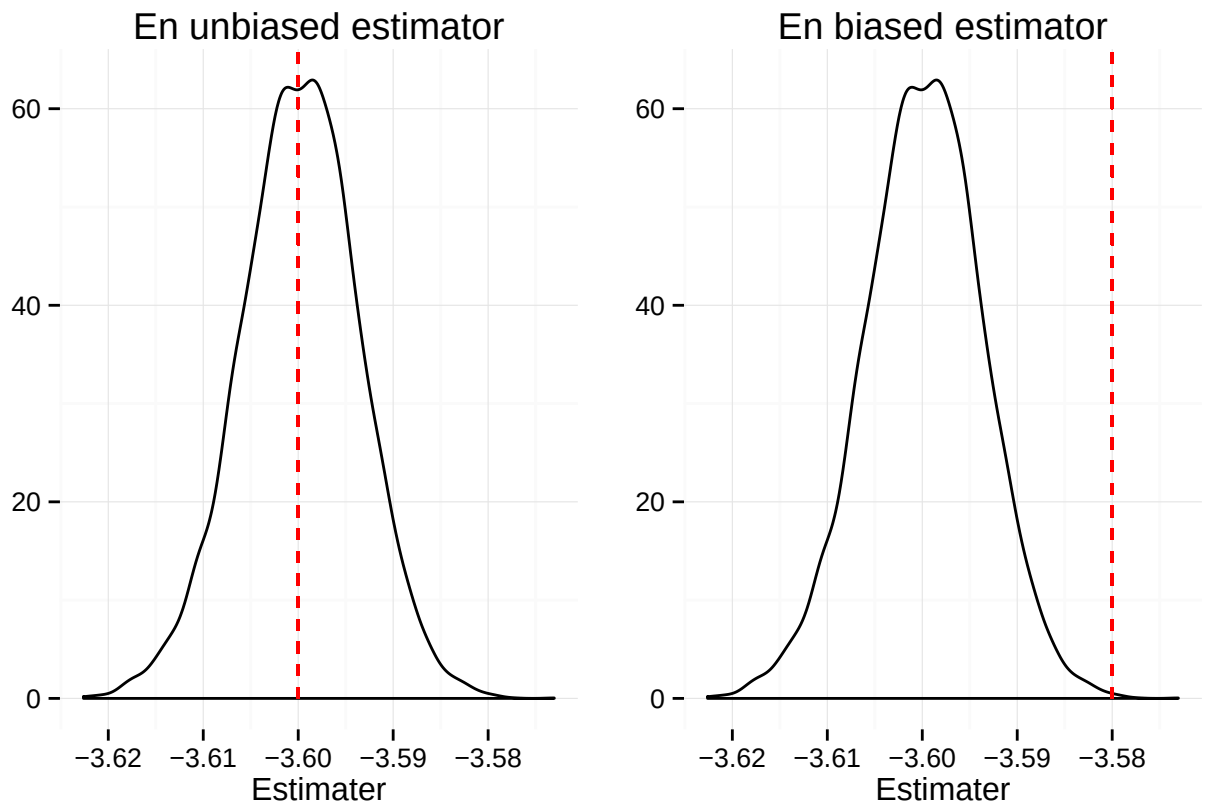
Bias og efficiens

Når man vælger, hvilken estimator man skal bruge, gør man det typisk ud fra begreberne bias og efficiens. Begge dele er koncepter, som vi (mere eller mindre indirekte) er stødt på i løbet af Metode 1. Bias er ofte det væsentligste kriterie, vi arbejder ud fra. Hvis man siger, at en estimator er unbiased, så betyder det, at gennemsnittet af dens stikprøvemålsfordeling er lig med gennemsnittet i populationen. På Metode 1 så vi f.eks. at vores bud på andelen af folk, der stemmer på Dansk Folkeparti, på tværs af mange stikprøver ikke vil ramme den faktiske andel i populationen, hvis særlige typer af DF-vælgere systematisk fravælger at deltage i meningsmålinger – vi vil opleve bias.

I figuren neden for giver jeg eksempler på, hvordan en biased og en unbiased estimator kan se ud – forestil jer, at vi bruger OLS-estimatoren. I venstreside er OLS unbiased – dens bud på sammenhængen mellem to variable har normalfordelt sig stort set præcist omkring populationens sande værdi, som er illustreret ved den røde linje. I højre side er OLS biased – dens stikprøvemålsfordeling har et gennemsnit, der er mere negativt end det sande (som igen

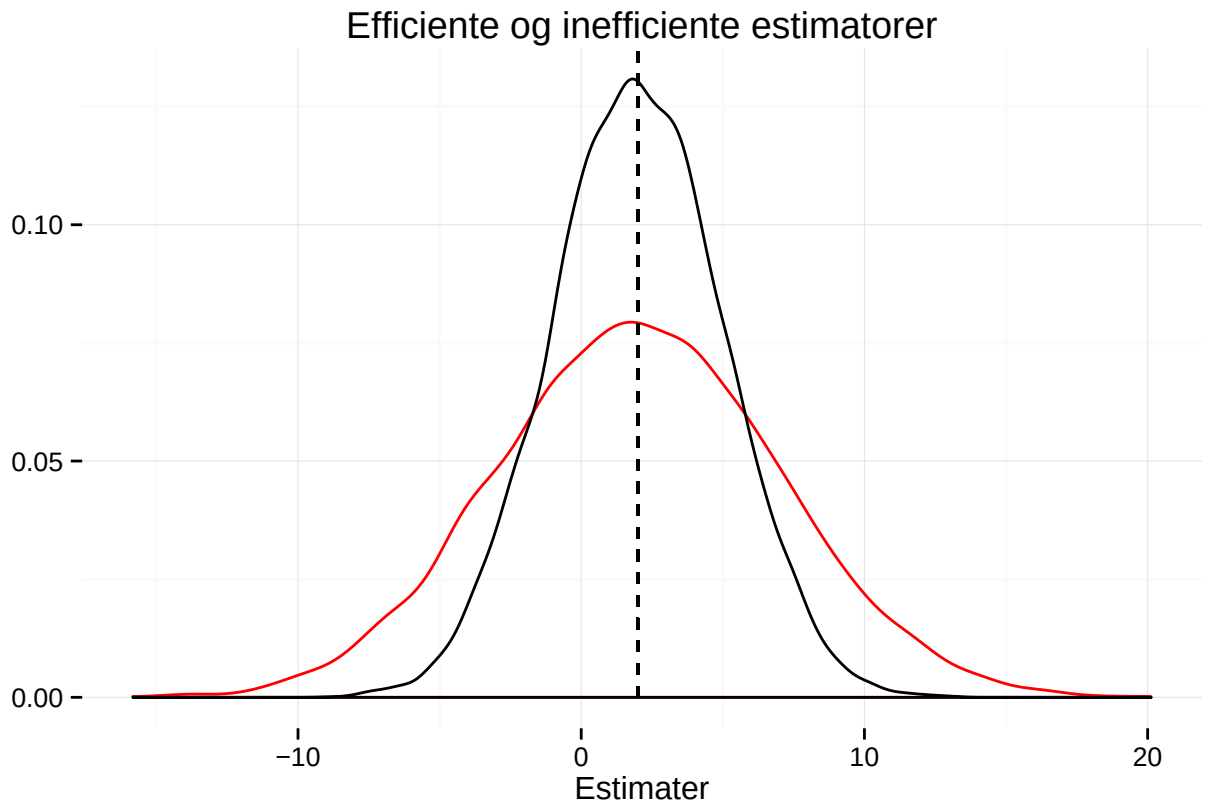
[3]Det kræver lidt matematik at bevise Gauss-Markov-teoremet, og behandlingen her vil i vidt omfang bruge intuition i stedet for. Matematikken er dog ikke vanvittigt kompliceret, så hvis det har interesse, kan jeg anbefale at kigge i f.eks. Wooldridge (2013)

er den røde linje). Som vi senere skal se, vil OLS være biased, hvis de fire første Gauss-Markov-antagelser ikke er opfyldt.



Det er ofte oplagt at bruge unbiasedness som sit første kriterie, når man skal vælge, hvilken estimator, der er den mest passende. Det giver mening at have som grundlæggende betingelse, at estimatoren i hvert fald i gennemsnit skal ramme rigtigt. Men det er ikke nok – for hver problemstilling vil der eksistere en lang række estimatorer, der er unbiased. Derfor bruger vi efficiens, som det andet kriterium. Det drejer sig om estimatorens præcision – jo mere spredning, der er i stikprøvemålsfordelingen omkring dets gennemsnit, desto mindre efficient vil den være (og desto større vil standardfejlen være). Det er derfor ikke nok at man i gennemsnit rammer rigtigt – for at en estimator skal være 'den bedste' skal man kunne forvente, at den over gentagne målinger vil give resultater, der ligger stabilt og tæt samlet omkring dette gennemsnit. Det er præcis hvad Gauss-Markov-teoremet siger, at OLS vil gøre – hvis antagelserne er opfyldt – den vil både ramme rigtigt i gennemsnit *og* være mindst lige så præcis som enhver tænkelig alternativ estimator.

I grafen neden for har jeg illustreret dette med to stikprøvemålsfordelinger over simuleret data – den røde fordeling er gamma-koefficienten (som vi mødte på Metode 1), mens den sorte er OLS. Her antager vi, at Gauss-Markov-antagelserne er opfyldt. Som det fremgår, rammer begge estimatorer i gennemsnit den sande værdi og er derfor unbiased. Men OLS' stikprøvemålsfordeling er langt mere koncentreret omkring gennemsnittet, og man kan derfor forvente langt mere præcise resultater. OLS er således i dette tilfælde en *bedre* estimator end gamma.



Forudsætninger for OLS regression

Indtil nu har jeg kun behandlet antagelserne, der ligger til grund for OLS, overfladisk. I dette afsnit vil jeg uddybe, hvad antagelserne betyder, og hvilke konsekvenser brud på dem har for vores resultater. Som nævnt tidligere er der fem væsentlige antagelser – 1. tilfældig udvælgelse, 2. linearitet, 3. ingen perfekt kollinearitet, 4. eksogenitet og 5. homoskedasticitet. Dertil kommer antagelsen om normalfordelte fejlled. Den bliver ofte anset som mindre vigtig, idet den ikke er nødvendig, når man har tilstrækkeligt mange observationer. Antagelserne har forskellig betydning for OLS' bias og efficiens, som jeg nu vil gennemgå.

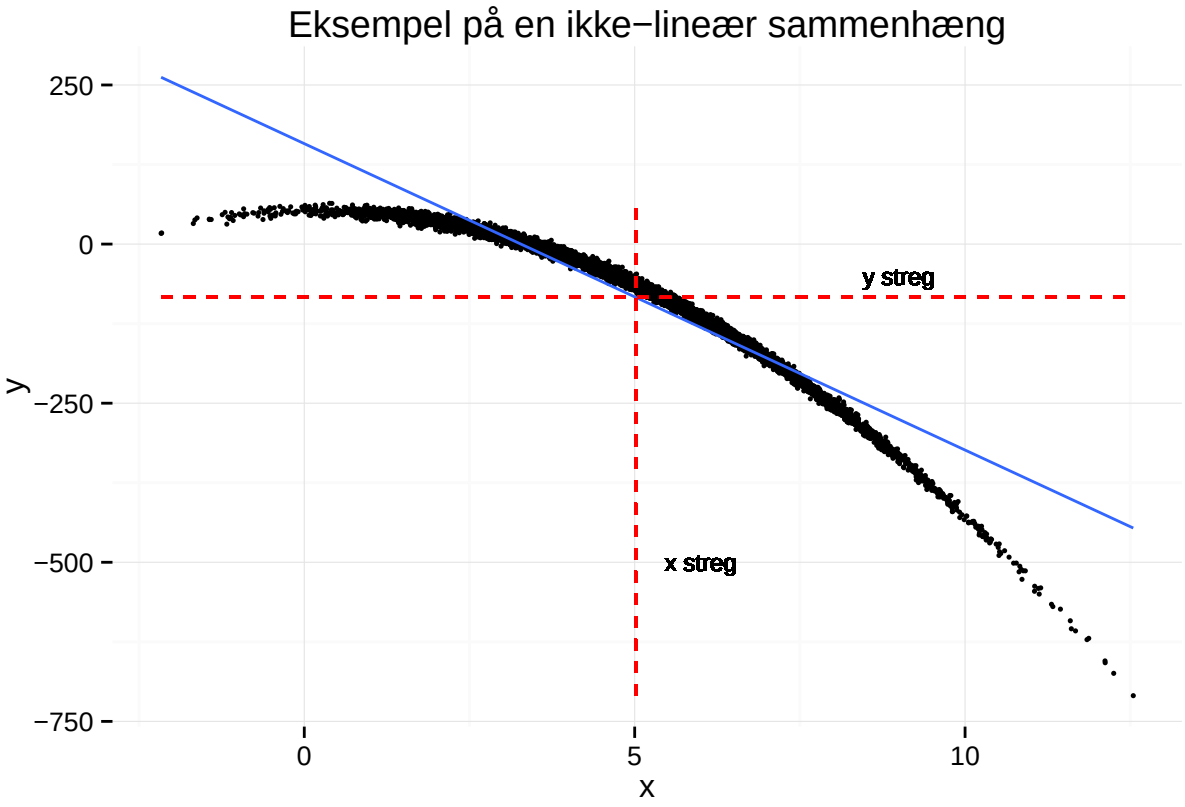
Tilfældig udvælgelse

Er en antagelse, som vi også har gjort os med alle de estimatorer, vi mødte på Metode 1. Når man arbejder med stikprøvedata er den nødvendig at gøre sig[4]: hvis man arbejder med en udvælgelsesmetode, der ikke er sandsynlighedsbaseret, vil man ikke kunne opnå repræsentativ data, og ens estimater vil i gennemsnit ramme en forkert værdi. Konsekvensen er derved, at OLS-estimatoren bliver biased. Det er vigtigt, at man husker på, at frafald i ens undersøgelse vil være et brud på denne antagelse, hvis det ikke sker tilfældigt – hvis man f.eks. arbejder med vælgeradfærd og en særlig type DF-vælgere falder fra, eller hvis man i et landestudie systematisk mangler data fra fattige lande, vil det give bias i ens resultater.

[4] Antagelsen indtager en noget anderledes betydning, når man har data på hele populationen, som kan være tilfældet, når man arbejder med registerdata, se f.eks. Thomsen (1997)

Linearitet

Linearitetsforudsætningen drejer sig specifikt om den form, som *sammenhængen* mellem den afhængige variabel og de uafhængige antager. En lineær sammenhæng er en, hvor den afhængige variabel vil ændre sig med den samme konstante værdi, når vi ændrer den uafhængige variabel. Det sker uanset, hvilken værdi variablene oprindeligt havde[5]. Et klassisk eksempel på en ikke-lineær sammenhæng er den, man finder mellem individers indkomst og alder – i takt med at man bliver ældre, vil man typisk opleve en øget indkomst. Dette gælder, indtil man forlader arbejdsmarkedet og går på pension, hvorefter ens indkomst vil falde. I figuren neden for har jeg illustreret en sådan kurvelineær sammenhæng, der skifter fortegn. Se Agresti & Finlay (2014) for yderligere behandling af ikke-linearitet – bl.a. af eksponentielle sammenhænge, transformation af variable og tolkning.



Det er vigtigt at holde sig for øje, at et brud på denne antagelse vil medføre bias – vores estimater vil ikke i gennemsnit ramme den sande sammenhæng. Men det er også relevant at påpege, at resultaterne ikke vil være lige dårlige for alle værdier af variablene – som det fremgår vil regressionslinjen ramme tendensen i data relativt godt for værdierne tæt på gennemsnittet af den afhængige og uafhængige variabel (i punktet $(\bar{x}, \bar{y})$, som er vist med de røde linjer). Omvendt giver regressionslinjen et meget dårligt bud på tendensen i data, når man bevæger sig væk fra gennemsnittene af de to variable.

[5]Man kalder det ofte antagelsen om linearitet i parametrene netop, fordi det drejer sig om sammenhængen mellem to variable og ikke variablene selv.

Ingen perfekt kollinearitet

Antagelsen omhandler forholdet mellem vores uafhængige variable. Nærmere bestemt betyder antagelsen, at ingen uafhængig variabel må være perfekt korreleret med de andre uafhængige variable. Hvis en uafhængig variabel kan forudsiges perfekt af en anden eller en kombination af andre uafhængige variable, så kan modellen simpelthen ikke estimeres. Hvis dette sker, kan det nemt genkendes – Stata undlader at estimere koefficient og standardfejl for en eller flere variable. Støder man ind i det problem, bør man omspecificere sin model ved at fjerne variable. Det er vigtigt at være opmærksom på, at man ikke risikerer spuriøsitet, når man fjerner en variabel, der er perfekt forudsagt af andre variable – al variation, som den ville kunne forklare i vores afhængige variabel, bliver allerede forklaret af de andre variable. Man oplever en særlig form for perfekt kollinearitet, hvis man ikke har noget variation i sin uafhængige variabel – det vil også medføre, at modellen ikke kan estimeres. Igen oplever vi, at variation i vores uafhængige variabel – Gerring (2004) påpeger – er nødvendig, hvis vi skal have nogen resultater.

Der er en subtil forskel mellem perfekt kollinearitet og multikollinearitet, hvor det sidste er det Agresti & Finlay (2014) behandler. Multikollinearitet er situationer, hvor ens uafhængige variable er korrelerede uden at være det fuldstændig. Multikollinearitet gør standardfejlene større, og OLS bliver derved mindre efficient, end den ellers ville have været. Det er dog vigtigt at holde sig for øje, at fravær af multikollinearitet ikke er en af Gauss-Markov-antagelserne, og OLS vil stadig være BLUE, selvom vores uafhængige variable er korrelerede. Det er også vigtigt, at man er bevidst om, at multikollinearitet ikke som sådan er et statistisk problem – OLS' standardfejl er ikke forkerte, når de uafhængige variable er korrelerede. Det er også, hvad der gør multikollinearitet til et fænomen, der ofte kan være ret svært at håndtere: det gør vores estimater mere upræcise, men så længe variablen også korrelerer med den afhængige variabel, vil det give bias i parameterestimatet at fjerne den (se næste afsnit), og det er således en nødvendig mangel på præcision.

Eksogenitet

Som vi tidligere har været inde på, kan man skrive en OLS-model på følgende måde:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon_i$$

Hvor x_1 til x_k er et antal uafhængige variable, og β_1 til β_k er korrelationen mellem hver af vores uafhængige variable og y. β_0 er konstanten og ϵ er et fejlled. Husk på, at fejlledet repræsenterer alle tænkelige variable, der påvirker vores afhængige variabel, men som vi ikke har taget med i regressionsmodellen. Det betyder også, at fejlledet er fundamentalt uobserverbart – hver del, som vi ønsker at observere af ϵ, skal operationaliseres som en variabel, der kan inddrages i vores analyse, og derved er det ikke længere en del af fejlledet. Desuden er der så mange faktorer, som påvirker enhver given afhængig variabel, og så meget tilfældighed, at det aldrig vil være muligt for os at indfange alt det, der vil være en del af fejlledet.

Antagelsen om eksogenitet er overtrådt, når mindst én uafhængig variabel er korreleret med noget, der er i vores fejlled. Det siger man typisk kan ske på tre måder:

(relevante) udeladte variable
Alle variable, som er korrelerede med den afhængige, men ikke indgår i regressionsmodellen, vil være en del af fejlledet. Hvis de udeladte variable er korrelerede med de uafhængige variable, som vi har i vores regressionsmodel, vil det medføre endogenitetsbias. Udeladte variable kan undertiden medføre spuriøsitet – at korrelationen mellem to variable forsvinder, når man tager højde for en tredje variabel, som hidtil ikke har været inkluderet. Det er heller

ikke unormalt, at man finder undertrykte sammenhænge, hvor to variable, der oprindeligt ikke var korrelerede, blev det efter, at man tog højde for en tredje variabel. Det er noget, som vi har beskæftiget os med tidligere.

Hvis man synes, at det er rimeligt at antage, at en udeladt variabel kun er korreleret med én af de uafhængige variable, som allerede er specificeret, kan man forudsige, hvilken retning bias vil have. I tabellen neden for er x_1 den allerede specificerede uafhængige variabel, mens x_2 er den udeladte variabel. β_2 er korrelationen mellem den udeladte variabel og den afhængige variabel.

Table 1: Bias ved udeladte variable (Wooldridge, 2013: 86).

	$\text{corr}(x_1, x_2)>0$	$\text{corr}(x_1, x_2)<0$
$\beta_2 > 0$	Bias opad	Bias nedad
$\beta_2 < 0$	Bias nedad	Bias opad

Som det fremgår, vil bias være opad (OLS-koefficienten vil være større end den virkelige sammenhæng), hvis den udeladte variabel er positivt korreleret med den afhængige variabel ($\beta_2 > 0$) samtidig med, at den er positivt korreleret med den uafhængige variabel, som allerede er specificeret ($\text{corr}(x_1, x_2)>0$). Omvendt vil OLS-koefficienten være under den sande sammenhæng, hvis den udeladte variabel er negativt korreleret med den afhængige og positivt korreleret med den uafhængige variabel. Bias vil også være negativ, hvis den udeladte variabel er positivt korreleret med den afhængige og negativt korreleret med den uafhængige variabel. Sidst vil bias i OLS-koefficienten være opad, hvis den udeladte variabel er negativt korreleret med både den afhængige og den uafhængige variabel.

Det betyder, at vi helt overordnet kan sige, at OLS-estimatet vil være lavere end den sande sammenhæng, hvis korrelationerne mellem 1) den udeladte og afhængige variabel og 2) den udeladte og den uafhængige variabel har modsat fortegn. Bias vil derimod være opad, hvis de har samme fortegn. Størrelsen på bias afhænger af, hvor stærkt korreleret den udeladte variabel er med både den afhængige og den uafhængige variabel.

Det er vigtigt at være opmærksom på, at en udeladt variabel kun medfører bias, hvis den både korrelerer med den afhængige og den uafhængige variabel. Det er, hvad der menes med, at den udeladte variabel skal være relevant, før det giver bias. Hvis det udeladte kun korrelerer med de uafhængige variable, vil det medføre en unødvendig multikollinearitet. Hvis det omvendt kun korrelerer med den afhængige variabel, kan det derimod give mening at inddrage, selvom en udeladelse ikke ville give bias, fordi det vil mindske det uforklarede element i vores afhængige variabel og derved give mindre standardfejl.

Er den udeladte variabel korreleret med flere af de uafhængige variable, bliver det hurtigt meget svært at forudsige, hvordan bias vil se ud. Men det er stadig vigtigt, at man gør sig overvejelser om, hvordan bias ville se ud under forskellige antagelser – det er trods alt ofte de bedste bud, vi har.

Omvendt kausalitet

Hvis den afhængige variabel har en selvstændig påvirkning på en eller flere af de uafhængige variable, vil det også medføre bias. Den omvendte kausalitet vil betyde, at vi ikke kun opfanger effekten, af den uafhængige variabel – som faktisk er den, som vi forsøger at estimere. Vores OLS-estimater vil også indeholde den afhængige variabels feedbackeffekt. I nogle metodetekster kalder man omvendt kausalitet for endogenitet. Det er dog standarden i dag, at brud på de tre kriterier, som gennemgås her (dvs. udeladte variable, omvendt kausalitet og målefejl), alle vil medføre endogenitet. Omvendt kausalitet (eller simultanitet som det ofte kaldes) er således kun én af mange kilder til endogenitetsbias.

Målefejl

Den sidste ting, som man normalt siger kan medføre endogenitet, er målefejl. Man skelner mellem systematiske og tilfældige fejl. Afhængig af, hvilken type fejl man har at gøre med, og om målefejlen er i den afhængige eller uafhængige variabel, vil det have forskellige konsekvenser, hvad angår bias og inefficiens i OLS, som bliver behandlet langt mere udtømmende i King, Keohane & Verba (1994)[6]. Sondringen mellem de to typer af målefejl er nært beslægtet med målingsvaliditet og reliabilitet, som vi tidligere har diskuteret.

Det er meget vigtigt, at man holder sig for øje, at antagelsen om eksogenitet ikke kan testes. Derimod bør man gøre sig konkrete teoretiske overvejelser om, 1) hvilke relevante variable, man udelader, 2) om der kan være tale om omvendt kausalitet, og 3) om man kan risikere, at målingen af ens variable er behæftet med fejl.

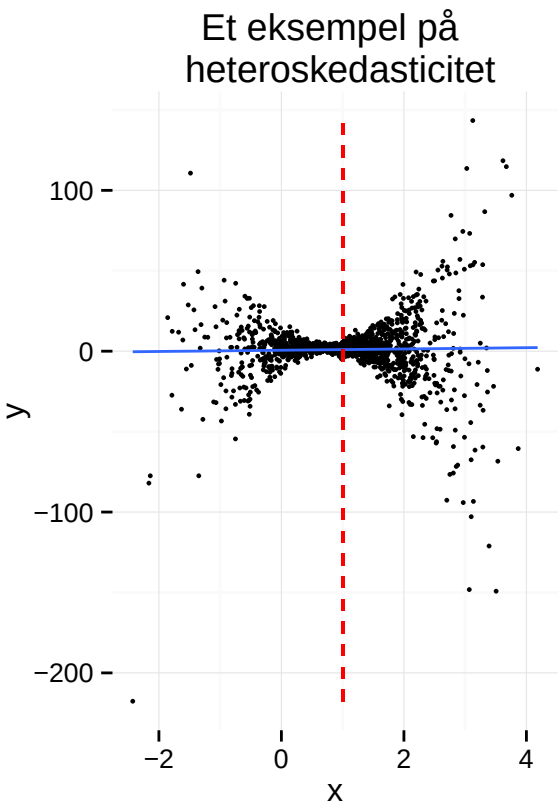
Homoskedasticitet

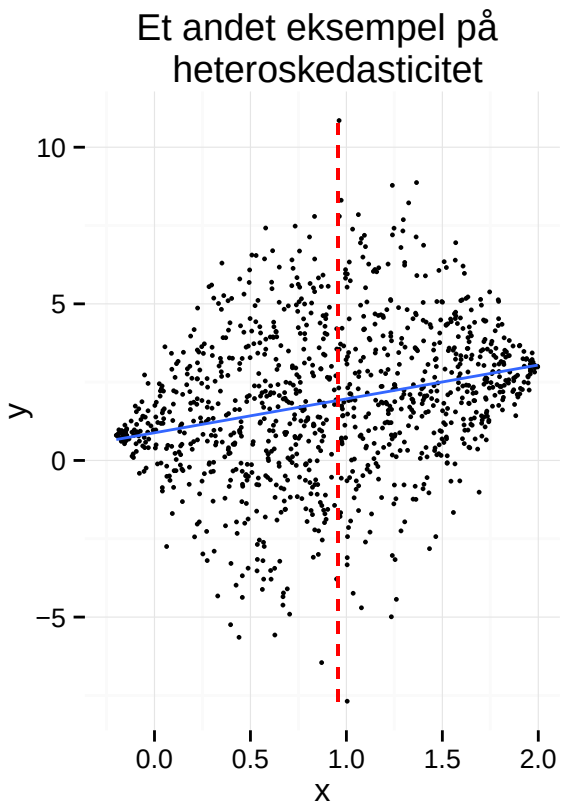
Den sidste af Gauss-Markov-antagelserne handler om ens varians (homoskedasticitet) – uanset den uafhængige variabels værdi, skal OLS-estimaterne have samme præcision. Er det ikke tilfældet, er der uens varians (heteroskedasticitet), hvilket medfører inefficiens. Derfor er OLS ikke længere den bedste estimator, som vi kan finde på, hvis der er heteroskedasticitet. Desuden vil standardfejlene blive forkert estimerede.

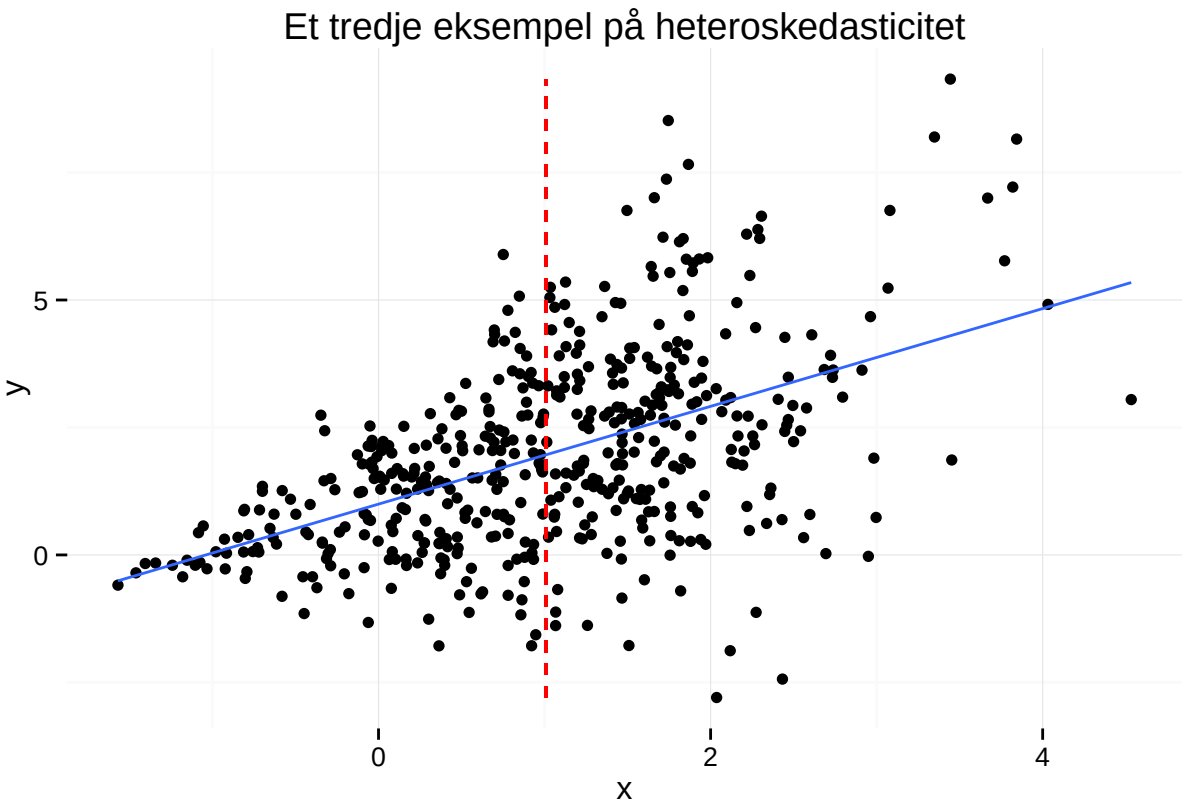
Normalt bruger man residualerne til at diagnosticere heteroskedasticitet. Ligger de lige langt fra regressionslinjen uanset den uafhængige variabels værdi, tyder det på ens varians. Er der omvendt tydelige forskelle på, hvordan residualerne falder afhængig af værdien på de uafhængige variable, tyder det på, at antagelsen er overtrådt. Agresti & Finlay (2014) kommer nærmere ind på, hvordan man tester det. Figuren neden for viser, hvordan ens og uens varians kunne se ud.

På Metode 2 plejer vi normalt at håndtere heteroskedasticitet ved at bruge Whites robuste standardfejl, som korrigerer de konventionelle OLS-standardfejl således, at de kan forventes at estimere den sande standardfejl, selvom antagelsen om homoskedasticitet ikke er opfyldt.

[6] Generelt giver de en enormt god behandling af alle tre kilder til endogenitet og mulige løsninger på problemerne


Et eksempel på heteroskedasticitet
100
0
−100
−200
y
−2
0
2
4
x


Et andet eksempel på heteroskedasticitet
10
5
0
−5
y
0.0
0.5
1.0
1.5
2.0
x


Et tredje eksempel på heteroskedasticitet
5
0
y
0
2
4
x

Alle grafer i figuren oven for er standardeksempler på heteroskedasticitet. I grafen øverst til venstre starter residualet med at ligge langt fra linjen, hvorefter det begynder at ligge nærmere linjen i takt med at x stiger. På et tidspunkt vender mønstret igen, og for de høje værdier af den uafhængige variabel, ligger residualerne igen langt fra regressionslinjen. I grafen øverst til højre ligger residualerne derimod tæt på regressionslinjen for de lave værdier af den uafhængige variabel. I takt med at x stiger, begynder residualet at ligge fjernere fra regressionslinjen igen. Men til sidst vender mønsteret igen, og for de højeste værdier af x, ligger residualerne igen tæt på linjen. I den nederste graf ligger residualerne meget tæt på linjen til at begynde med, men i takt med, at den uafhængige variabel stiger i værdi, så begynder residualerne at falde længere og længere fra regressionslinjen.

Graferne viser en central pointe: uens varians betyder, at vores model har forskellige grader af præcision afhængig af værdien på x. I det nederste plot er vores model f.eks. utroligt præcis for de lave værdier af x. Tilgengæld begynder mange andre ting at få betydning for vores afhængige variabel, når x stiger. Vores uafhængige variabel tilbyder med andre ord ikke en nær så stærk forklaring på den afhængige variabel, når den indtager store værdier. Derfor er heteroskedasticitet også et interessant aspekt ved vores data, og ikke et irritationsmoment, som bare skal overkommes.

Konsekvensen, som heteroskedasticitet har for vores standardfejl, vil afhænge af ens data. Som tommelfingerregel kan man dog sige, at de almindelige OLS-standardfejl *overvurderer* usikkerheden, hvis residualerne ligger tæt på regressionslinjen, når man har at gøre med x- værdier, der ligger langt fra x's gennemsnit. I dette tilfælde er OLS' standardfejl større end de robuste, og det bliver sværere at få statistisk signifikante resultater, selvom der faktisk er en sammenhæng mellem de variable, vi undersøger. Det er tilfældet i grafen øverst til højre. Residualerne falder meget tæt på regressionslinjen for henholdvis høje og lave værdier af x – netop de værdier, der ligger langt fra variablens gennemsnit (den røde stiplede linje).

Omvendt *undervurderer* OLS den faktiske usikkerhed, når residualet ligger tæt på regressionslinjen, når vi befinder os omkring den uafhængige variabels gennemsnit. Dvs. at OLS-standardfejlen i dette tilfælde er mindre end de robuste standardfejl, og vi vil have lettere ved at få statistisk signifikante resultater, selvom der faktisk ikke er en sammenhæng mellem vores variable. Det er tilfældet i den første graf – som det fremgår, er spredningen i residualet her meget lav ved den røde stiplede linje, der er x's gennemsnit.

I den nederste graf er det svært at forudsige heteroskedasticitetens betydning for OLS-standardfejlene – for høje værdier af x, er der høj varians og omvendt for de lave værdier af x. Et godt råd i sådan en situation er at sammenligne de konventionelle og robuste standardfejl.

Som nævnt bruger vi typisk robuste standardfejl som løsning på heteroskedasticitetsproblemer. Der er dog to væsentlige ting at bemærke omkring dem. For det første skal der en vis mængde af data til, før robuste standardfejl i gennemsnit giver det rigtige bud på usikkerheden i vores resultater. Dvs. at ved små stikprøver, kan de robuste standardfejl være mindst lige så forkerte som OLS' konventionelle standardfejl. Det er svært at give en præcis tommelfingerregel for, hvornår man har data nok til at bruge Whites robuste standardfejl, men det har vist sig, at de allerede fra $N \approx 50$ klarer sig godt ved lav til middel heteroskedasticitet – dog kræver det mindst 100-150 observationer, hvis heteroskedasticiteten er meget stærk (Cribari-Neto, 2004).

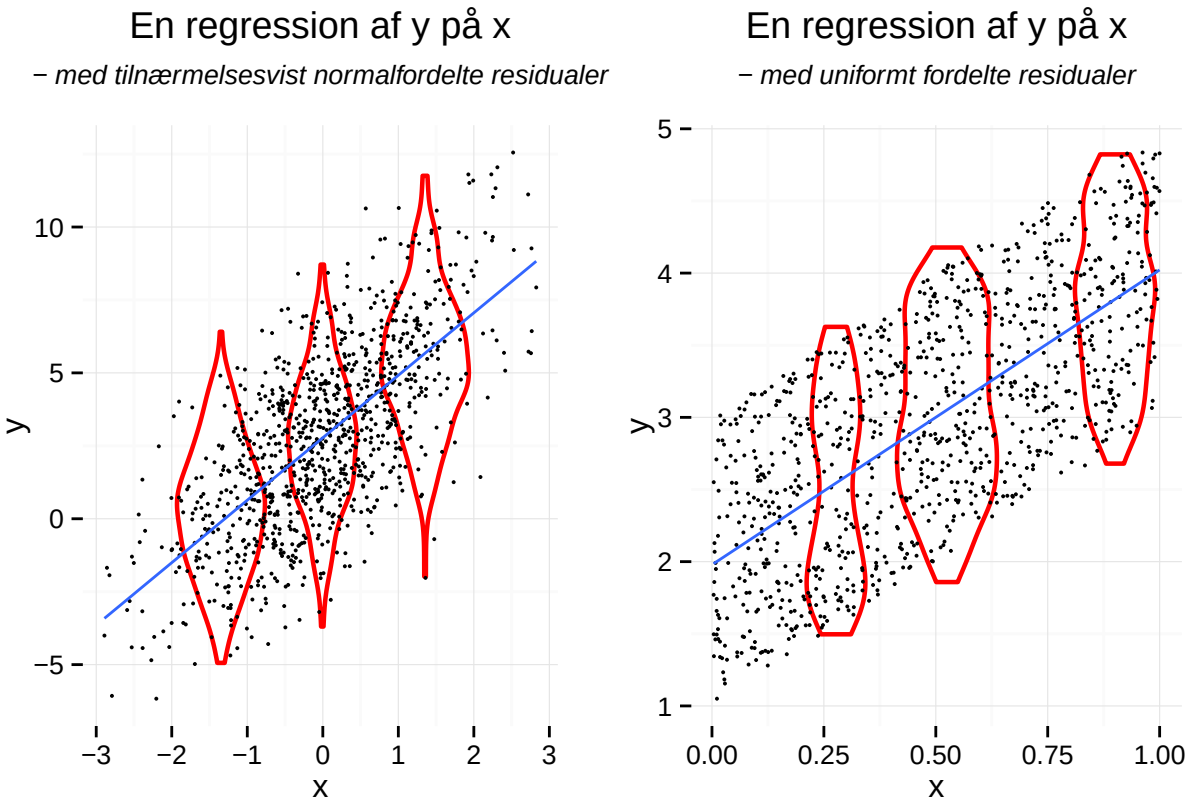
Typisk er det en god ide at sammenligne robuste og konventionelle standardfejl. Hvis forskellige standardfejl ændrer ens resultater markant, kan det indikere, at der er andre af forudsætningerne for OLS, der er brudt (typisk vil det handle om ikke-linearitet eller udeladte variable[7]). For det andet medfører robuste standardfejl kun, at vi kan få

[7] Bl.a. derfor er det blevet foreslået at bruge en sammenligning af konventionelle og robuste standardfejl som en generel test for brud på OLS-forudsætninger (King & Roberts, 2015). Bemærk dog, at dette ikke er Metode 2-pensum.

korrigeret problemerne med vores standardfejl – de gør ikke, at OLS er BLUE. Derfor vil der eksistere mere efficiente estimatorer, når antagelsen om homoskedasticitet er brudt uanset, hvordan vi udregner standardfejlen.

Normalfordelte fejlled

Udover de fem Gauss-Markov-antagelser er det nogle gange nødvendigt at supplere med en yderligere antagelse – om at residualerne følger en normalfordeling. Hvis residualerne er normalfordelte vil det medføre, at OLS' stikprøvemålsfordeling også vil være det, og vi vil derfor kunne bruge t-statistikken og almindelige konfidensintervaller, når vi skal inferere resultaterne fra vores OLS-analyse. Figuren neden for illustrerer antagelsen. De røde linjer viser tyngden og spredningen på data omkring regressionslinjen. Som vi kan se, er residualerne nogenlunde normalfordelte omkring regressionslinjen i grafen til venstre. De er koncentrerede inde omkring selve linjen og tynder ud i takt med, at vi bevæger os væk. De falder således, at kurven er nogenlunde klokkeformet. I grafen til højre er der tale om en såkalt uniform fordeling. Den er flad, og koncentrationen af residualer er den samme uanset, hvor langt man kommer fra regressionslinjen.

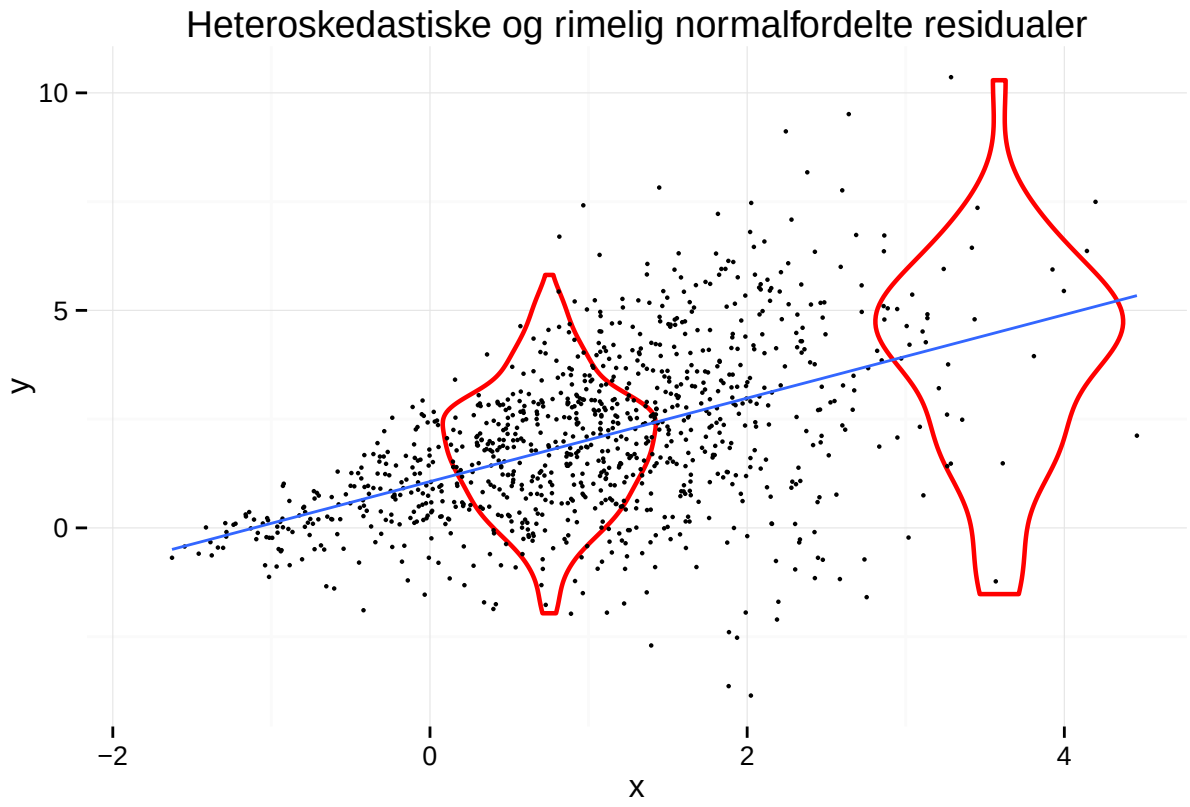


Det er dog vigtigt at holde sig for øje, at den såkaldt udvidede centrale grænseværdisætning gælder for OLS. Det betyder, at hvis man har nok observationer i sit datasæt vil OLS' stikprøvemålsfordeling normalfordele sig uanset residualernes fordeling. Ligesom det var tilfældet for de estimatorer, som vi mødte på Metode 1, er 'nok' observationer et vidt begreb. Afhængig af hvor meget residualernes fordeling adskiller sig fra normalfordelingen kan $N \approx 30$ være nok, men hvis afvigelsen er særligt stærk, kan det kræve væsentligt flere observationer, ligesom man kan klare sig med færre, hvis afvigelsen ikke er voldsom. Hvis residualerne er normalfordelte, siger man, at inferens ved hjælp af OLS er *eksakt*. Når vi er nødt til at forlade os på at have nok observationer til, at stikprøvemålsfordelingen

bliver normalfordelt, bedriver vi *asymptotisk*[8] inferens.

Selvom man har meget data, kan det dog stadig være relevant at skele til residualernes fordeling. Hvis de er normalfordelte, får OLS en endnu stærkere efficiensegenskab – den vil nu ikke bare være den mest præcise af alle lineære estimatorer, men også blandt dem, som ikke antager, at sammenhængen mellem vores variable er lineær (Wooldridge 2013, 111). Man skal dog huske på, at OLS stadig vil være biased, hvis vi overser substantielle ikke-lineariteter.

Det er ikke sjældent, at folk forveksler antagelsen om residualernes fordeling med den om, residualerne har ens varians uanset værdierne på vores uafhængige variable – antagelsen om homoskedasticitet. Den vigtige forskel på de to antagelser er, at den empiriske regel vil gælde, hvis man opfylder antagelsen om, at residualerne er normalfordelte. Dvs. at 68 % af residualerne vil ligge inden for en standardafvigelse af regressionslinjen, mens 95 % vil ligge inden for 2 standardafvigelser og 99,7 % vil ligge inden for tre standardafvigelser. Det siger antagelsen om homoskedasticitet ingenting om – den siger derimod, at der vil være den samme standardafvigelse i residualerne uanset den uafhængige variabels værdi. Figuren neden for illustrerer situationen, hvor residualerne er stærkt heteroskedastiske, men samtidig følger en normalfordeling. Som det fremgår, er residualerne koncentrerede omkring regressionslinjen for lave x-værdier, men langt mere spredt ud i den høje ende af skalaen. I begge tilfælde er residualerne dog også rimelig normalfordelte omkring regressionslinjen.



[8]Når man snakker om, at noget er asymptotisk, betegner det en situation, hvor antallet af observationer går mod uendeligt. På den måde arbejdede vi også med asymptotisk inferens på Metode 1, hvor vi lærte, at stikprøvemålsfordelingerne for de fleste af vores tests normalfordelte sig, når N gik mod uendeligt.

Andre spørgsmål omkring OLS

Den afhængige variabels måleniveau og tilstedeværelsen af indflydelsesrige observationer samt outliers er yderligere forhold, som er væsentlige at forholde sig til, når man gennemfører analyser med OLS. Det er dog vigtigt at holde sig for øje, at selvom begge dele kan indikere brud på Gauss-Markov-antagelserne, så er de ikke som sådan nødvendige for at etablere OLS som BLUE. Herunder behandler jeg begge spørgsmålene kort.

Hvad med måleniveau?

På Metode 1 brugte vi meget tid på at finde den rette estimator givet variablenes måleniveau. Et centralt resultat fra Gauss-Markov-teoremet er dog, at OLS vil være unbiased, så længe man bare kan rangordne sin afhængige variabel[9] – dvs. at så længe ens afhængige variable ikke er nominelt skallerede, vil vi i gennemsnit få gode resultater med OLS. Dette betyder dog ikke, at den afhængige variabels måleniveau er irrelevant. Hvis den afhængige variabel er binær, vil der altid opstå heteroskedasticitet i OLS, hvorfor vi typisk bruger logistisk regression – som vi behandler senere i forløbet – i disse tilfælde. Hvis den afhængige variabel er ordinal, kan det under tiden medføre ikke-normalfordelte fejlled og heteroskedasticitet. Det har dog vist sig, at det yderst sjældent vil være tilfældet, hvis der er mindst fire-fem kategorier i den afhængige variabel. Det er også årsagen til, at vi på Metode 1 blev introduceret til tommelfingerreglen om, at en ordinalt skalleret variabel kan behandles som intervalskalleret, når den har mindst fem kategorier.

Selvom OLS ikke gør nogle direkte antagelser om variablenes måleniveau, kan det derfor alligevel have konsekvenser, fordi variablenes skallering har indflydelse på to af OLS-antagelserne. Men vi kan teste for, hvor voldsomme bruddene på antagelserne er og udbedre dem, hvis de er grelle. Dette gør OLS ekstremt fleksibel og betyder, at estimatoren kan anvendes i mange tilfælde, hvor den ikke i snæver forstand er korrekt at bruge (f.eks. når vi har binære afhængige variable).

Hvad med indflydelsesrige observationer?

Fordi OLS baserer sig på at minimere de kvadrerede residualer i stedet for f.eks. de absolutte residualer, kan outliers og såkaldt indflydelsesrige observationer få indflydelse på resultaterne. Overordnet set kan outliers defineres som enkelte observationer, der ligger langt fra regressionslinjen – og derved har et stort residual. Indflydelsesrige observationer er – som navnet antyder – observationer, der har stor selvstændig indflydelse på regressionslinjen – det kan diagnosticeres ved at kigge på residual, leverage og de såkaldte deletion diagnostics (se Agresti & Finlay, 2014: 452-453). Indflydelsesrige observationer er dem, man bør bekymre sig mest om, eftersom det er dem, der kan ændrer vores resultater markant. Outliers derimod skal nærmere forstås som overraskende observationer, som normalt ikke vil forandre vores resultater meget, medmindre de også er indflydelsesrige.

Betydningen af en indflydelsesrig observation for OLS-resultaterne afhænger i høj grad af, hvorfor den givne observation adskiller sig så markant fra den generelle tendens – og typisk kan man lære meget om sit data ved at inspicere dem. En vigtig sondring i den forbindelse er mellem, hvad observationen betyder for henholdsvis stikprøven og OLS' stikprøvemålsfordeling. Hvis der ikke er nogen særlig årsag til, at den indflydelsesrige observation adskiller sig fra resten af observationerne i ens datasæt udover, at det er en usædvanelig observation, som er anderledes af ren og skær tilfældighed, vil det i udgangspunktet kun betyde noget for selve stikprøven. Den indflydelsesrige observation

[9] Faktisk er independent sample t-test, two sample test of proportion, ANOVA og Pearsons r alle sammen særtilfælde af OLS regression.

vil i det tilfælde gøre stikprøven mindre repræsentativ og medføre, at OLS-estimatet fra den pågældende stikprøve vil give et dårligere bud på den sande værdi i populationen. Derved vil estimatet fra den enkelte stikprøve falde længere fra den sande værdi, end det ellers ville have været tilfældet – men gennemsnittet af stikprøvemålsfordelingen vil være uændret.

Særligt hvis der er tale om mange indflydelsesrige observationer, kan det dog være en indikation på, at man har udeladt en væsentlig variabel eller, at sammenhængen ikke er lineær. I det tilfælde, vil OLS blive biased – ikke som sådan fordi der er indflydelsesrige observationer, men fordi de er til stede pga. brud på Gauss-Markov-antagelserne. Efter man har identificeret sine indflydelsesrige observationer, er det derfor altid en god ide at overveje, om der er noget særligt ved dem. Har de noget til fælles, som kan indikere en udeladt variabel? Når man inspicerer dem grafisk, falder de så i et mønster, der kunne indikere, at en ikke-lineær transformation ville være mere passende? Til forelæsningen vil jeg komme med eksempler på dette. Som generelt råd er det en god ide at køre sine regressioner både med og uden indflydelsesrige observationer for på den måde at se, hvad de betyder for ens resultater. Hvis det ikke ændrer på noget nævneværdigt – eller måske endda forstærker ens resultater – er det et godt tegn.

Konklusion

I dette undervisningsnotat har jeg redegjort for Gauss-Markov-teoremet og antagelserne bag det. Jeg har forklaret, at hvis antagelserne er opfyldt, vil OLS være den blandt alle unbiased estimatorer, der har den laveste varians, og i gennemsnit vil OLS derfor være den mest præcise. Vi har desuden set på, hvilke konsekvenser for OLS-estimaterne brud på antagelserne vil have.

Forelæsningen og den resterende litteratur til undervisningen om OLS-forudsætninger vil behandle nærmere, hvordan vi diagnosticerer brud på de her antagelser. Det kan være kompliceret at stille den rigtige diagnose – men det er normalt et spændende stykke arbejde, der gør én væsentligt klogere på sin data. I løbet af arbejdet med ens diagnose vil det være nødvendigt at komme med gode substantielle forklaringer på, hvorfor ens data ser ud som det gør – hvorfor er residualet højere i den ene ende af skalaen? Hvorfor har få specifikke observationer stor indflydelse på ens resultater? I den forbindelse er det centralt, at man har en god og dyb forståelse for, hvad antagelserne betyder – og det er, hvad jeg forsøger at bidrage med i notatet.

Litteratur

Agresti, Alan & Barbara Finlay (2014): *Statistical Methods for the Social Sciences*, 4^{th} ed. London: Pearson Education.

Cribari-Neto, Francisco (2004): "Asymptotic inference under heteroskedasticity of unknown form". I: *Computational Statistics and Data Analysis*, 45, s. 215-233.

Dunning, Thad (2008): "Improving Quasal Inference: Strengths and Limitations of Natural Experiments". I: *Political Research Quarterly*, vol. 61 (2), s. 282-293.

Gerring, John (2004): "What Is a Case Study and What Is It Good for?". I: *American Political Science Review*, vol. 98(2), s. 341-354.

King, Gary; Robert Keohane & Sidney Verba (1994): *Designing Social Inquiry – Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

King, Gary & Margaret Roberts (2015): "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It". I: *Political Analysis*, 23, s. 159-179.

Thomsen, Søren Risbjerg (1997): *Om anvendelse af signifikanstest i ikke-stikprøve situationer*, Undervisningsotat fra Institut for Statskundskab ved Aarhus Universitet.

Wooldridge, Jeffrey (2013): *Introductory Econometrics – A Modern Approach*. 5^{th} ed. Delhi: Cengage Learning.