# FORUDSÆTNINGER FOR LINEÆR REGRESSION OG VARIANSANALYSE EFTER MINDSTE KVADRATERS METODE

**AF**
**RUNE STUBAGER &**
**KIM MANNEMAR SØNDERSKOV**

5. udgave, januar 2011

DEPARTMENT OF POLITICAL SCIENCE
Aarhus University
Universitetsparken
DK-8000 Aarhus C ● Denmark
Telephone +45 8942 1111
Fax: +45 8913 9839

INSTITUT FOR STATSKUNDSKAB
Aarhus Universitet
Universitetsparken
8000 Århus C
Telefon 8942 1111
Telefax 8613 9839

# FORUDSÆTNINGER FOR LINEÆR REGRESSION OG VARIANSANALYSE EFTER MINDSTE KVADRATERS METODE

Rune Stubager
Tlf. 8942 1308
E-mail: stubager@ps.au.dk

Kim Mannemar Sønderskov
Tlf. 8942 1260
E-mail: ks@ps.au.dk

INSTITUT FOR STATSKUNDSKAB
AARHUS UNIVERSITET
UNIVERSITETSPARKEN
8000 ÅRHUS C

5. UDGAVE JANUAR 2011

# Indholdsfortegnelse

# Indledning[1]

Lineær regression, variansanalyse og kovariansanalyse[2] efter mindste kvadraters metode hviler på en række forudsætninger, der i rimelig grad skal være opfyldte, for at resultaterne kan benyttes til at blive klogere på virkeligheden. Dette notat gennemgår hovedparten af disse forudsætninger med udgangspunkt i følgende multiple regressionsmodel:

$$Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + e$$

hvor $Y$ er en metrisk afhængig variabel, $X_1$ og $X_2$ er metriske uafhængige variabler, $e$ er residualet (dvs. forskellen mellem den af modellen forudsagte værdi og den observerede værdi) og $\hat{\alpha}$, $\hat{\beta}_1$ og $\hat{\beta}_2$ er de estimerede koefficienter for henholdsvis konstantleddet og hældningen på de to uafhængige variabler. Modellen kan udvides med yderligere uafhængige variabler, men til at begynde med holder vi os for overskuelighedens skyld til to metriske uafhængige variabler. I notatets anden del vedrørende forudsætninger i forbindelse med (ko-)variansanalyse tilføjes to kategoriske uafhængige variabler.

Baggrunden for notatet er, at behandlingen af forudsætninger i forhold til lineær regression og variansanalyse er en smule overfladisk i Agresti & Finlay og at udførslen af diverse forudsætningstests i SPPS kan være indviklet uden anvisninger. Denne grundige indføring i disse forudsætninger skal dog ikke tages som udtryk for, at opfyldelse og vurdering af forudsætninger er mere vigtige for disse metoder end for fagets andre metoder.

---

[1] Dette notat bygger i vidt omfang på anbefalinger og overvejelser indeholdt i Fox (1991) og er en lettere revideret version af 3. udgave, forfattet af Rune Stubager. Forfatterne skylder Lotte Bøgh Andersen, Søren Risbjerg Thomsen og Lene Aarøe stor tak for hjælpsomme kommentarer til denne eller tidligere udgaver af notatet. Figurer og tabeller er udarbejdet af Søren Heldgaard Olesen. Denne udgave er udarbejdet med udgangspunkt i SPSS version 18.

[2] I alle tre analysemetoder er den afhængige variabel metrisk. Lineær regression indeholder kun metriske uafhængige variable. Optræder der kun kategoriske uafhængige variabler, taler man om variansanalyse; inddrages også metriske variabler, taler man om kovariansanalyse. Alle metoder hviler på det samme matematiske og statistiske grundlag, og i mange sammenhænge bruges betegnelsen *lineær regression*, uanset måleniveauerne for de uafhængige variable.

Som empirisk eksempel tages der udgangspunkt i en analyse af en landeaggregeret version af Gallups Milleniumsurvey fra 2000, der blev gennemført i 60 lande (tilgængelig på Institut for Statskundskab, AU på N:\M_Data\World\Millenium\Millen01_aggr). Spørgsmålet, der vil blive undersøgt, er, hvorvidt borgernes gennemsnitlige holdning til udsagnet: "On the whole, men make better political leaders than women do" (svarmuligheder: 1: Enig eller 2: Uenig, variabel q15c) i landene ($Y$) påvirkes af landenes bruttonationalprodukt pr. indbygger (målt i 1000$/indbygger, variabel gnpcap99 delt med 1000) ($X_1$) og graden af demokrati i 1999 (målt ved Freedom House-scorer, variabel dem99) ($X_2$).[3]

Forudsætningerne kan indordnes i følgende ni punkter:

1. Modelspecifikation (1): Modellen skal indeholde de relevante uafhængige variabler
2. Modelspecifikation (2): Fravær af endogenitet – den afhængige variabel må ikke påvirke de(n) uafhængige variabel(er)
3. Modelspecifikation (3): Linearitet - sammenhængen mellem de uafhængige og den afhængige variabel skal være lineær i populationen
4. Fravær af indflydelsesrige observationer
5. Fravær af stærk multikollinearitet – de uafhængige variabler må ikke være stærkt indbyrdes forbundne
6. Fravær af autokorrelation – residualerne må ikke være korrelerede
7. Fejlleddene skal være normalfordelte
8. Homoskedasticitet – fejlleddene skal have den samme varians for givne værdier på de uafhængige variabler
9. Datasættets karakter

Forudsætningerne kan kategoriseres på forskellige måder. For det første kan der skelnes mellem forskellige konsekvenser af forudsætningsbrud (der dog alle svækker modellens gyldighed). Brud på forudsætning 1-4 kan

[3] Landeaggregeret betyder, at der for hvert land er udregnet en gennemsnitsværdi for respondenternes svar på spørgsmålene i surveyen. Landenes værdi på q15c afspejler således gennemsnittet af respondenternes svar på dette spørgsmål, hvor høje værdier indikerer udbredt uenighed i påstanden. De to uafhængige variable er ikke baseret på surveyen, men i stedet på landedata fra andre kilder. Det anbefales, at læseren selv udfører de beskrevne test på datasættet parallelt med læsningen.

skabe bias, hvilket vil sige at de estimerede parametre (konstanten og hældningskoefficienterne) er mindre gode bud på parametrenes sande værdier i populationen. Med andre ord bevirker sådanne forudsætningsbrud, at man ikke kan stole på størrelsen af koefficienterne og i grelle tilfælde deres fortegn, hvilket selvsagt er problematisk.[4] Brud på forudsætning 5-8 skaber derimod usikkerhed om, hvorvidt betingelserne er til stede for, på baggrund af den estimerede model, at foretage korrekt inferens, da de beregnede *P*-værdier kan være fejlagtige. Forudsætning 9 vedrører også mulighederne for at foretage inferens, men især til hvilken population dette i givet fald kan gøres. For det andet kan der skelnes mellem, hvilke dele af modellen, forudsætningerne vedrører. Forudsætning 1-5 vedrører de inkluderede variabler (og eventuelt de ikke-inkluderede variabler), mens forudsætning 6-8 vedrører fejlleddenes fordeling. Forudsætning 9 vedrører den måde data er indsamlet på. For det tredje kan der skelnes mellem, hvornår forudsætningerne er relevante at diskutere og teste. Forudsætning 1-2 er kun relevante, når der undersøges kausalpåstande (jf. afsnit 1), mens de øvrige er relevante, uanset om der undersøges kausalpåstande eller blot hypoteser om en statistisk sammenhæng. Derudover er forudsætning 6 kun relevant når data ikke er indsamlet ved hjælp af tilfældig udvælgelse, mens de øvrige er relevante uanset indsamlingsmetode.

Udover disse forudsætninger hviler regression og variansanalyse på visse forudsætninger, der ikke gennemgås grundigt her: den afhængige variabel skal (tilnærmelsesvist) være intervalskaleret og samtlige variabler skal (tilnærmelsesvist) være målt uden fejl (som altid). Dette gælder særligt for de uafhængige variabler. Fravær af målefejl forudsætter valide operationaliseringer samt reliable data. I empiriske analyser bør variablernes måleniveau samt deres validitet og reliabilitet derfor diskuteres – ligesom ved brug af andre metoder.

Nedenfor gennemgås hver enkelt af de ni forudsætninger med fokus på a) årsager til og især konsekvenser af forudsætningsbrud, b) hvordan forud-

[4] Bemærk, at bias betyder, at estimaterne systematisk under- eller overvurderes i gentagne stikprøver. Derfor kan et biased estimat strengt taget godt være korrekt i en given stikprøve; se Agresti & Finlay (2008: 108-109).

sætningsbrud identificeres, og c) hvad, der kan gøres i tilfælde af forudsætningsbrud.

Indledningsvist er det på sin plads at gøre opmærksom på, at mindre brud på nedenstående forudsætninger snarere er reglen end undtagelsen. I mange tilfælde er spørgsmålet således ikke, *om* der er brud på forudsætningerne, men *hvor store* de er, og *hvilke konsekvenser* det kan medføre. Man bør derfor ikke kassere ellers velfungerende modeller på grund af mindre forudsætningsbrud. Omvendt bør man tage større brud alvorligt og eventuelt respecificere modellen (ved for eksempel at inddrage nye variabler eller transformere de eksisterende) eller anlægge en strengere fortolkning af de beregnede *P*-værdier. De anviste test og de oplistede argumenter kan dermed ses som redskaber til at forbedre modellen snarere end som absolutte stopklodser – selvom en model selvsagt kan være så problematisk, at den må kasserers som værende uden informativ værdi. I konkrete analyser (for ikke at nævne eksamensopgaver) vil det således være en dyd, hvis man er i stand til at bruge testene og argumenterne til at forbedre en opstillet model, så den bliver bedre og mere informativ. Subsidiært kan de bruges til at tage de nødvendige forbehold i fortolkningen af resultaterne.

### *1 Modelspecifikation (1): Inklusion af relevante variable*[5]

En af de store fordele ved den multiple regressionsmodel er, at den muliggør undersøgelse af kausalpåstande, altså om en given variabel har en kausaleffekt på en anden variabel (fx om stigende velstand medfører et mere positivt syn på kvinders politiske lederevner). Som bekendt forudsætter kausalpåstande, blandt andet en statistisk korrelation efter kontrol for 3. variabel. Da multipel regressionsanalyse kan inkludere flere uafhængige variabler, kan metoden netop undersøge, om der er en statistisk korrelation mellem to variabler og samtidigt kontrollere for mulige 3. variabler.

Her er det dog væsentligt at indse at kontrol for 3. variabel betyder kontrol for samtlige relevante 3. variabler, og dermed at udeladelse af relevante variabler underminerer gyldigheden af kausalpåstanden. I regressionsanaly-

[5] Dette og det følgende afsnit trækker på King, Keohane og Verba (1994: afsnit 5.2 & 5.4). Se også Lolle og Klemmensen (2010).

se er *konsekvensen* af manglende kontrol for en eller flere relevante variable, at de estimerede parametre er biased, og dermed at man ikke kan stole på størrelsen eller selv fortegnet af hældningskoefficienterne. Bias forårsaget af udeladte relevante variabler kaldes i litteraturen for *omitted variable bias* eller *udeladt variabel bias*.

For at forstå konsekvenserne af omitted variable bias kan der tages udgangspunkt i begreberne *spuriøs* og *undertrykt sammenhæng*. Som bekendt betyder spuriøs sammenhæng, at en sammenhæng mellem to variabler forsvinder (bliver insignifikant), når der kontrolleres for 3. variabel. Undertrykt sammenhæng betyder tilsvarende, at en sammenhæng mellem to variabler først fremkommer, når der kontrolleres for 3. variabel. Spuriøsitet og undertrykt sammenhæng er to tilfælde af omitted variable bias, hvor koefficienterne (estimeret uden kontrol) netop er biased i forhold til de sande parametre (som vi kommer tættere på ved at inkludere kontrolvariabler). Spuriøsitet og undertrykt sammenhæng er altså to særtilfælde af omitted variable bias, hvor effekten forsvinder eller fremkommer ved kontrol. Generelt referer omitted variable bias til situationer, hvor de opnåede koefficienter systematisk er for store eller små pga. udeladte variabler.

*Årsagen* til omitted variable bias er altså udeladte variabler. Mere specifikt opstår bias, fordi effekten af den/de udeladte variabler fejlagtigt tilskrives de inkluderede variabler.[6] Som illustration af dette kan der tages udgangspunkt i modellen med velstand, demokrati og synet på kvinders politiske lederevner. Parameterestimaterne fra denne model er vist i Tabel 1, og de indikerer, at velstand medfører et mere positivt syn på kvinders lederevner, da koefficienten for bruttonationalproduktet er positiv og statistisk signifikant efter kontrol for demokrati.[7]

*Tabel 1* Regressionsoutput med estimaterne for BNP indcirklet

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | 1,629 | ,053 | | 30,616 | ,000 | | |
| | gnp1000 | ,008 | ,002 | ,508 | 4,545 | ,000 | ,716 | 1,397 |
| | dem99 Democracy - sum af Rights and Liberties 1999-2000 | -,019 | ,007 | -,296 | -2,650 | ,010 | ,716 | 1,397 |

a. Dependent Variable: q15c Disagree men better political leaders

Estimatet af velstands effekt kunne dog være misvisende pga. af omitted variable bias. Det er plausibelt at eksempelvis samfundets sekulariseringsgrad påvirker synet på kvinder på den måde at borgere i sekulære samfund har et mere positivt syn på kvinders lederevner. Da en række af verdens religioner indeholder et vist negativt syn på kvinder, er dette ikke urealistisk. Ydermere er det plausibelt, at sekulære samfund er mere velstående da frigørelsen fra religiøse dogmer blandt andet kunne medføre øget imødekommenhed overfor ny, produktiv teknologi. Såfremt sekularisering påvirker både velstand og synet på kvinders evner, og såfremt sekularisering udelades af modellen vil estimatet af velstands effekt på synet på kvinder være biased, da (en del af) effekten af sekularisering fejlagtigt bliver tilskrevet velstand i modellen. Med andre ord vil udeladelsen af sekularisering formentligt betyde, at kausaleffekten af velstand overvurderes. Man kunne sågar forstille sig, at effekten af velstand helt forsvandt ved kontrol for sekularisering, hvilket ville være et eksempel på en spuriøs sammenhæng.

Størrelsen af bias er en funktion af korrelationen mellem henholdsvis den udeladte variabel og den inkluderede uafhængige variabel samt korrelationen mellem den udeladte variabel og den afhængige variabel. Det betyder, at bias er størst, når den udeladte variabel er kraftigt korreleret med både den inkluderede uafhængige og den afhængige. I sådanne tilfælde vil den inkluderede uafhængige variabel fejlagtigt blive tilskrevet en stor del af effekten af den udeladte variabel. På den anden side betyder dette også, at der ikke opstår bias, hvis den udeladte variabel er ukorreleret med den inkluderede variabel og/eller den afhængige variabel. Hvis den udeladte variabel er ukorreleret med den inkluderede variabel, vil den inkluderede variabel ikke kunne overtage effekten fra den udeladte. På samme måde vil der ikke opstå bias, hvis den udeladte variabel ikke har en effekt på den afhængige, da der ikke vil være en effekt at overtage.

Samlet set illustrerer ovenstående, at udeladelsen af visse variabler kan have alvorlige konsekvenser for gyldigheden af vores modeller, da udeladelsen kan skabe systematisk forkerte estimater.

Desværre findes der ikke en simpel test, der kan *identificere* omitted variabel bias. Der findes metoder, der kan teste dette, men de er teknisk komplicerede og lader sig ikke altid implementere i praksis. Disse metoder behandles derfor ikke her (og de ligger uden for pensum i Metode II). I de fleste sammenhænge må man nøjes med at diskutere om (tilnærmelsesvis) alle relevante variabler er inddraget og, hvis dette ikke er tilfældet, diskutere, hvad konsekvenser af at udelade de relevante variabler er. Dette er særligt relevant i opgavesammenhænge, hvor det sjældent er muligt at indsamle data, der ikke allerede er inkluderet i datasættet.

I forbindelse med en sådan diskussion er der flere forhold, det er værd at være opmærksom på, og som kan inddrages i diskussionen. For det første betyder forudsætningen om inklusion af 3. variabel *ikke* at alle tænkelige variabler bør inddrages – kun *relevante* 3. variabler. Som ovenstående eksempler illustrerer, skaber udeladelsen af 3. variabel kun bias, såfremt den udeladte variabel er korreleret med den/de uafhængige variabler og samtidigt påvirker den afhængige variabel i modellen. Det betyder, at variabler, der enten ikke påvirker den/de inkluderede uafhængige variabler eller den afhængige variabel, er irrelevante 3. variabler og derfor kan udelades, uden at dette skaber bias. Da vi ikke har data for den udeladte variabel, må dens relation til den/de uafhængige variable samt den afhængige variabel alene afklares gennem en teoretisk diskussion (jf. nedenfor).

Ydermere er det vigtigt at være opmærksom på, at det kun er variable som *påvirker* den/de uafhængige variabler, der skaber bias. Med andre ord er det kun udeladte variabler, som i kausalkæden ligger bag den uafhængige variabel, der skaber bias. Dette forhold illustrerer vigtigheden af at overveje kausalforholdet mellem potentielle uafhængige variabler. I ovenstående eksempel kunne man argumentere for, at sekularisering primært er en effekt af velstand og ikke en årsag til velstand. Hvis dette er tilfældet, påvirker sekularisering ikke velstand og kan derfor udelades af modellen, uden at dette skaber bias i estimatet af velstands effekt. Ud fra disse betragtninger kunne man overveje om demokrati hører hjemme i ovenstående model, hvor vi er interesseret i effekten af velstand. Hvis man mener at demokratisering er et resultat af velstand og modernisering, så betyder det at demo-

krati ikke bør medtages i modellen. I dette tilfælde er kausalforholdet mellem disse to ret uklart, hvorfor man for en sikkerhedsskyld bør medtage demokrati i modellen.

Størrelsen af bias kan også diskuteres. Selvom der kan identificeres en potentielt udeladt variabel, kan denne muligvis være svagt korreleret med den uafhængige og/eller den afhængige variabel, hvilket vil kunne bruges som et argument for, at udeladelsen ikke skaber alvorlige problemer.

Som det fremgår, beror vurderingen af, hvorvidt ens model er plaget af omitted variable bias på teoretiske overvejelser. Optimalt set trækker disse overvejelser på etablerede teorier og empirisk evidens, men hvis sådanne ikke er tilgængelige, vil man kunne trække på nuancerede common sense betragtninger. Såfremt man står i en situation, hvor det ikke kan afvises at udeladte variable potentielt skaber kraftig bias, må man tage kraftige forbehold for resultaternes gyldighed i konklusionen.

Afslutningsvis bør følgende to forhold bemærkes. For det første kan forudsætningen om kontrol for 3. variabel i visse tilfælde være opfyldt i kraft af analysedesignet. I eksempelvis eksperimentelle designs, hvor observationer har fået tildelt værdier på den uafhængige variabel tilfældigt, og hvor der er et tilstrækkeligt antal observationer, vil der ikke være relevante 3. variabler. Da observationernes værdi på den uafhængige variabel er tildelt tilfældigt, kan der ikke eksistere bagvedliggende variabler, der påvirker den uafhængige variabel. I dette tilfælde kan man trygt udelade 3. variabler – også selvom disse har en kraftig effekt på den afhængige variabel. I eksemplet fra før kan dette argument ikke bruges, da landenes velstandsniveau på ingen måde kan siges at være tildelt eksperimentelt. For det andet kan man forstille sig situationer, hvor man ikke er interesseret i at teste kausalpåstande, men blot at undersøge, om der er en sammenhæng mellem to variabler – her testes ingen kausalpåstand og forudsætningen om inklusion af alle relevante variabler bortfalder. Hvorvidt man er interesseret i at teste kausalpåstande eller ej, afhænger af den konkrete problemstilling. I eksemplet fra før var problemstillingen, hvorvidt velstand (og demokrati) *påvirker* synet på kvinders lederevner. I dette tilfælde er der altså tale om en kausal på-

stand, hvilket betyder, at forudsætningen om fravær af omitted variable bias bør være opfyldt.

### *2 Modelspecifikation (2): Fravær af endogenitet*

Endogenitet refererer til situationer, hvor den afhængige variabel påvirker en eller flere af de uafhængige variabler i modellen, og forudsætningen vedrører dermed det 3. kriterium for kausalitet. Begrebet refererer til det forhold at de uafhængige variabler, der påvirkes af den afhængige variabel, ikke er uafhængige eller eksogene, men endogene. *Årsagen* til endogenitet kan eksempelvis være et feed-back loop mellem *Y* og *X*.

*Konsekvensen* af endogenitet er bias – som ved udeladte variabler – her kaldes bias endogenitetsbias.[8] Ligesom ved omitted variable bias forårsager endogenitet bias – altså at man ikke kan stole på størrelsen og fortegnet af koefficienterne. Mekanismen, der skaber bias, er også meget lig mekanismen bag omitted variable bias: Hvis *Y* påvirker *X*, vil *Y*s effekt på *X* fejlagtigt blive tilskrevet *X*, hvorved *X*s effekt på *Y* fejlestimeres.

Statistisk *identifikation* af endogenitet kræver desværre også tests, der ligger udenfor dette notats område (og udenfor pensum i Metode II). Det betyder, at vurderingen af, hvorvidt en model er plaget af endogenitet, ofte må bero på en teoretisk diskussion. Her skal det ud fra teoretiske overvejelser og eventuelt eksisterende empirisk materiale godtgøres, at kausaleffekten (i langt overvejende grad) går fra *X*'erne til *Y* – og ikke omvendt. Hvis dette ikke kan godtgøres, må modellen opgives eller som minimum behæftes med meget kraftige forbehold.

I relation til eksemplet med velstand og synet på kvinders politiske lederevner kunne man argumentere for, at modellen muligvis er plaget af endogenitetsproblemer. Hvis et samfund systematisk fravælger visse ledere på baggrund af deres køn og ikke deres ledelsesmæssige kompetencer, vil samfundet muligvis have mindre kvalificerede politikere end andre sam-

[8] Endogenitetsbias kaldes også for simultanitetsbias, da *X*'erne og *Y* bestemmes samtidigt. Bemærk også at nogle forfattere anvender begrebet endogenitetsbias for bias forårsaget af både udeladte variable og feed-back effekter, og opsplitter derfor endogenitetsbias i omitted variable bias og simultanitetsbias.

fund. Dette kunne have negative konsekvenser for samfundets økonomiske politikker og muligvis dets velstand. Denne kausalkæde er dog lang og dermed usikker, hvilket kunne tale for, at kausaliteten i overvejende grad går fra velstand til synet på kvinder.

Ligesom i afsnittet om omitted variable bias bør det afslutningsvis bemærkes, at forudsætningen om fravær af endogenitet kun er relevant i forbindelse med test af kausale påstande, og derudover at eksempelvis eksperimentelle designs ikke er plaget af endogenitet, da *Y* ikke kan påvirke *X*, når *X* er tilfældigt tildelt.

### *3 Modelspecifikation (3): Linearitet*

Denne forudsætning vedrører sammenhængen mellem de den/de uafhængige variabler og den afhængige variabel i populationen. Denne forudsætning skal tilnærmelsesvist være opfyldt, uanset om der testes en kausal sammenhæng eller ej. Hvis der ikke er en lineær sammenhæng mellem *X*'erne og *Y* i populationen giver det ikke mening at anvende mindste kvadraters metode. Hvis forudsætningen ikke er rimeligt opfyldt, er analysen uanvendelig, da resultaterne er systematisk misvisende; altså biased. *Årsagen* til fravær af linearitet kan være virkeligheden i den forstand, at en variabel kan have en ikke-lineær sammenhæng med den afhængige variabel. Andre årsager kan dog også være modelspecifikt i form af udeladte variabler (jf. ovenfor).

Mens argumentationen for at alle relevante variabler er inkluderet, og for at kausalitetsretningen er korrekt gennemføres teoretisk, kan lineariteten både vurderes teoretisk og empirisk. Som et første skridt skal det således være muligt teoretisk at argumentere for, at den undersøgte sammenhæng er lineær – altså at der er realistisk at en ændring i *X* medfører den samme ændring i *Y*, uanset niveauet af *X*.

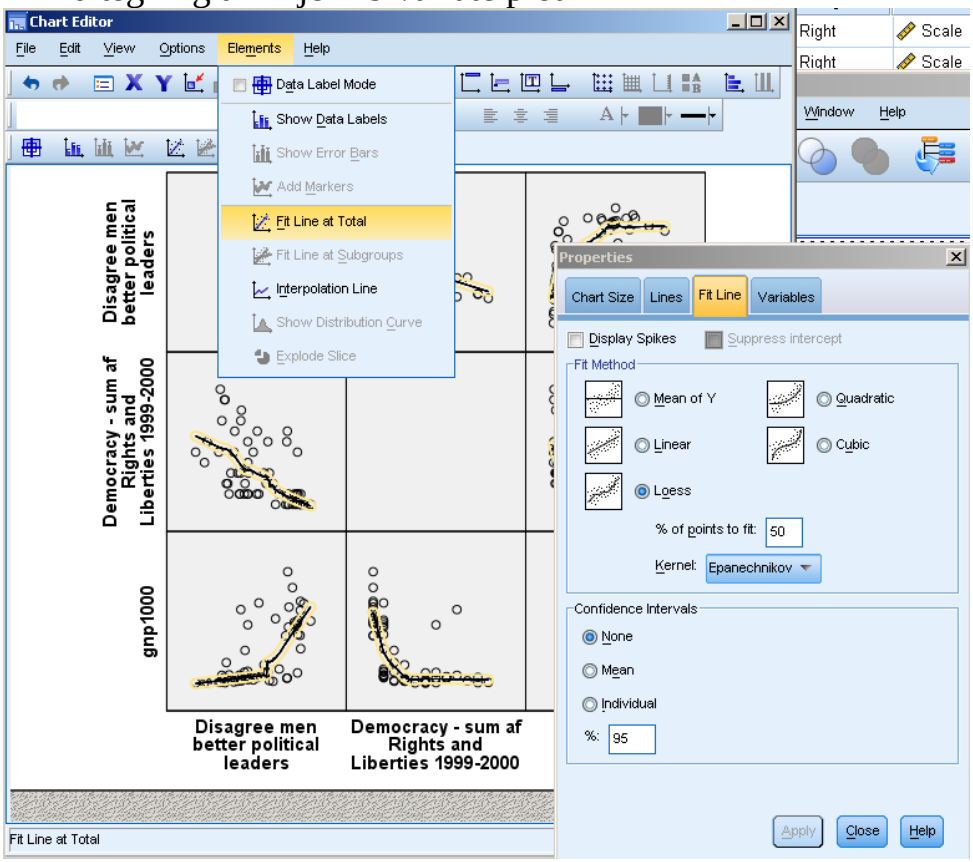
I næste skridt bruges sammenhængen i stikprøven til at vurdere sammenhængen i populationen. Dette kan gøres visuelt ved at plotte den afhængige variabel mod den eller de uafhængige variabler (se Figur 1). Som et første check kan man benytte Graphs | Legacy Dialogs | Scatter/Dot… i SPSS, hvor man vælger Matrix Scatter og overfører den afhængige og de uafhængige variabler til feltet under Matrix Variables. Derved produceres

en matrice, der viser alle variabler plottet mod hinanden. For at afgøre om der er brud på linearitetsforudsætningen, må man vurdere, om det er realistisk, at sammenhængen er lineær i populationen givet sammenhængen i stikprøven. Det betyder, at sammenhængen ikke nødvendigvis skal være fuldstændig lineær i stikprøven; det skal blot være realistisk, at sammenhængen er lineær i populationen.

For at lette vurderingen af, om der optræder en rimelig grad af linearitet i sammenhængene mellem den afhængige og hver af de uafhængige, kan man benytte den såkaldte Lowess-kurve (locally weighted scatterplot smoother). Denne kurve er et deskriptivt redskab, der viser tyngden i sammenhængen ved for hver observation på den variabel, der plottes på *X*-aksen, at estimere en værdi baseret på de omkringliggende observationers værdier. I udgangspunktet medtages i SPSS de omkringliggende 50% af observationerne, der vægtes efter, hvor tæt på den givne observation, de ligger. Dermed kommer de estimerede værdier til at vise de lokale tyngdepunkter i data, og når de plottes som en kurve, kan det afsløres, om der lokalt forekommer afvigelser fra en eventuel overordnet linearitet.

Lowess-kurven produceres ved, at man dobbeltklikker på plottet i outputvinduet. Dermed åbnes Chart Editoren, og man vælger dernæst Elements | Fit Line at Total, hvorved der åbnes et nyt vindue (jf. Figur 1). I dette vindue vælges fanen Fit Line, og der markeres ved Loess og klikkes Apply. Derved indlægges Lowess-kurverne i hvert af de producerede plot. Som det fremgår af nederste venstre hjørne af figuren, er der tegn på markante afvigelser fra lineariteten i sammenhængen mellem den afhængige variabel og BNP-variablen.

*Figur 1* Indlægning af linjer i bivariate plot

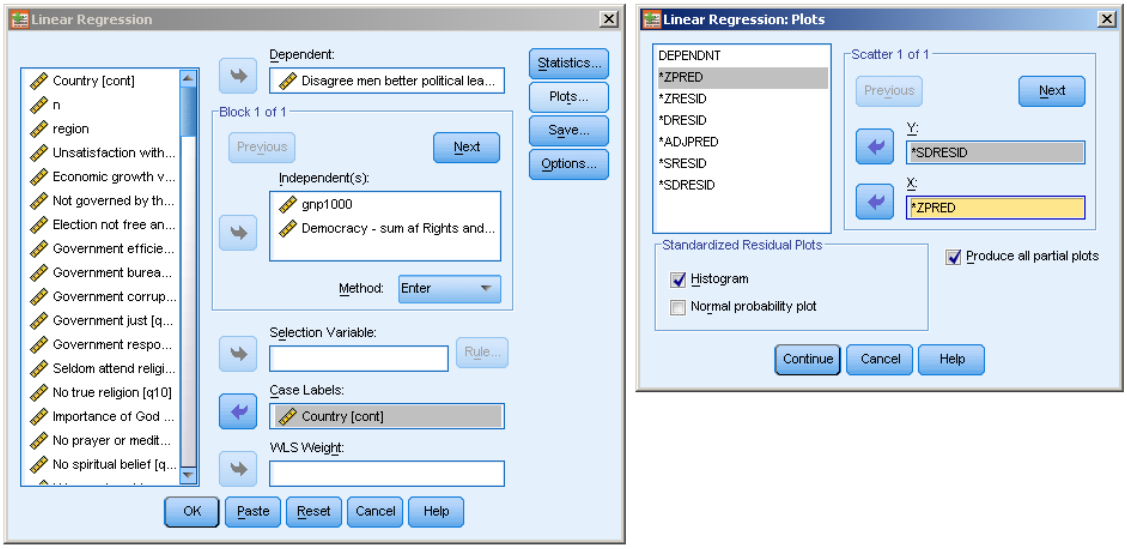



Før der tages stilling til, hvorvidt dette udgør et problem, skal man dog være opmærksom på, at de producerede plot er bivariate. Der er således ikke taget højde for, at effekten af de uafhængige variabler kan ændres når der kontrolleres for andre variable. For at tage dette forhold i betragtning er det i multivariate analyser nødvendigt at benytte sig af såkaldte partielle plot. Disse plot viser den partielle sammenhæng mellem hver af de uafhængige og den afhængige variabel renset for de andre uafhængige variablers indflydelse (se Agresti og Finlay, 2008: 328). I relation til vores model med de to uafhængige variabler $X_1$ og $X_2$, viser det partielle plot for $X_1$ og $Y$ den del af $X_1$, som er uforklaret af $X_2$ (og eventuelle andre inkluderede variabler) mod den del af $Y$, der er uforklaret af $X_2$ (og eventuelle andre).

Partielle plot produceres i forbindelse med selve regressionsanalysen; dvs. der vælges Analyze | Regression | Linear…, og de relevante variabler overføres som henholdsvis afhængige og uafhængige. Når man arbejder med datasæt, hvor de enkelte observationer har meningsfulde navne, kan

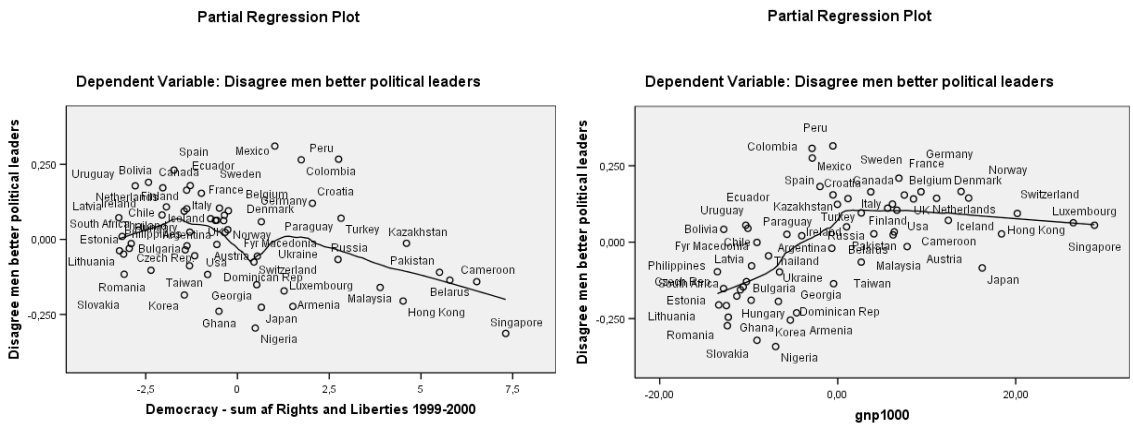
det lette fortolkningen, hvis man desuden overfører navnevariablen (i vores tilfælde landenavnet) til Case Labels-feltet, idet observationerne derved tilknyttes deres navne i de producerede plot. Derefter klikkes på Plots, og der markeres ved Produce all partial plots (jf. Figur 2).

*Figur 2* Fremstilling af partielle plot




I outputtet produceres nu et partielt plot for hver af de uafhængige mod den afhængige variabel. For at lette fortolkningen af disse plot kan man, på samme måde som beskrevet ovenfor, indlægge Lowess-kurver. I vores tilfælde viser det sig, at den overordnede tendens i sammenhængen mellem holdningsvariablen og demokratiscoren er lineær, hvilket indikerer at sammenhængen er lineær i populationen, og dermed at forudsætningen formentligt er opfyldt for denne variabel. Modsat er der er en klar ikke-lineær tendens i sammenhængen med BNP-variablen, og det forekommer derfor urealistisk, at observationerne er udtrukket fra en population, hvor der er en lineær sammenhæng mellem variablene (se Figur 3).

*Figur 3* Partielle plot



Når der optræder problemer af denne type, kan man overveje forskellige strategier. For det første kan det være en mulighed at inddrage yderligere variabler, der kan tage højde for dele af data, der ikke kan indpasses i den eksisterende model. I nærværende eksempel er der umiddelbart ikke noget der indikerer, at fraværet af linearitet skyldes udeladte variable. For det andet kan man vurdere, om en eller flere af variablene skal transformeres, så der opnås en højere grad af linearitet i modellen. En mulighed er her at forsøge, om en ikke-lineær specifikation af modellen har et bedre fit – for eksempel kan en kurvelineær sammenhæng modelleres med en kvadratisk model (hvor den uafhængige variabel kvadreres). En anden mulighed vil være at transformere den problematiske uafhængige variabel for eksempel ved hjælp af logaritme-funktionen. Overvejelser, om hvilken transformation der kan anvendes, bør baseres på teoretiske overvejelser om, hvilke aspekter ved de pågældende variabler, der kan give anledning til de observerede problemer, således at transformationen får basis i teorien snarere end i, hvad der i et givet datasæt mere eller mindre tilfældigt måtte have det bedste fit.

I vort eksempel er den sidste løsning, den mest nærliggende. Det er således velkendt fra litteraturen, at sammenhænge med BNP-variablen ofte er præget af en såkaldt gulveffekt (frembragt af at variablen har et nedre "gulv" lig med 0), som kan fjernes ved at tage logaritmen af BNP-variablen. Foretages en sådan transformation, opnås en høj grad af linearitet også for denne variabel.[9]

## *4 Fravær af indflydelsesrige observationer*

Hovedspørgsmålet i relation til indflydelsesrige observationer er, om enkelte observationer øver en uforholdsmæssig stor indflydelse på de estimerede koefficienter, således at disse er drevet af et fåtal af observationer snarere end hovedsammenhængen i data. Er dette tilfældet, er *konsekvensen* at den estimerede model er misvisende; koefficienterne er biased. Der kan være flere *årsager* til indflydelsesrige observationer. Oftest forekommer indfly-

[9] Såfremt der optræder interaktion mellem to metriske variabler i modellen, giver det ikke mening at fortolke på de partielle plot, hvori disse variabler indgår. Man kan derfor ikke vurdere linearitetsforudsætningen for variabler der indgår i interaktionsled.

delsesrige observationer i situationer med få observationer, da hver observations indflydelse er relativt større i sådanne situationer. En anden årsag kan være målefejl, hvor en eller flere fejlbehæftede observationer påvirker estimaterne kraftigt. Endelig kan manglende linearitet forårsage, at få observationer har en uforholdsmæssig stor indflydelse. Det, der skal undersøges, er derfor, om der forekommer sådanne enkelte observationer, der forrykker resultaterne, og de nedenfor diskuterede mål beregnes altså for hver enkelt observation i datasættet.

Indflydelsesrige observationer er ekstreme observationer, hvor der kan skelnes mellem to former for ekstremitet: Den ene form optræder, når en given observation ligger langt fra gennemsnittet på en eller flere uafhængige variabler. Ekstremitet i denne form giver observationen stor mulighed for at påvirke koefficienten for den eller de pågældende uafhængige variabler – man siger, at observationen har høj leverage. Dette behøver dog ikke være noget problem, med mindre observationen også er ekstrem i den (anden) forstand, at den afviger meget fra den af modellen forudsagte værdi, dvs. at den har et stort residual. Tilsvarende er en observation med et stort residual ikke noget problem, hvis den pågældende observation er placeret tæt på gennemsnittet på den eller de uafhængige variabler. Intuitivt kan man således anskue det sådan, at den indflydelse, som en given observation øver på de estimerede koefficienter, bestemmes som leverage$\times$residual. I analysen bør man altså være opmærksom på observationer, der både har stort residual og høj leverage.

En observations leverage måles med den såkaldte hat-værdi, der er et mål for observationens potentiale for at påvirke den estimerede værdi på den afhængige variabel ($\hat{Y}$ – deraf navnet). Hat-værdier har ingen absolut skala, men varierer fra model til model. Gennemsnittet er dog lig med antallet af uafhængige variabler delt med antallet af observationer,[10] og man vil som tommelfingerregel betragte observationer, der har hat-værdier på over 2 gange gennemsnittet, som ekstreme.

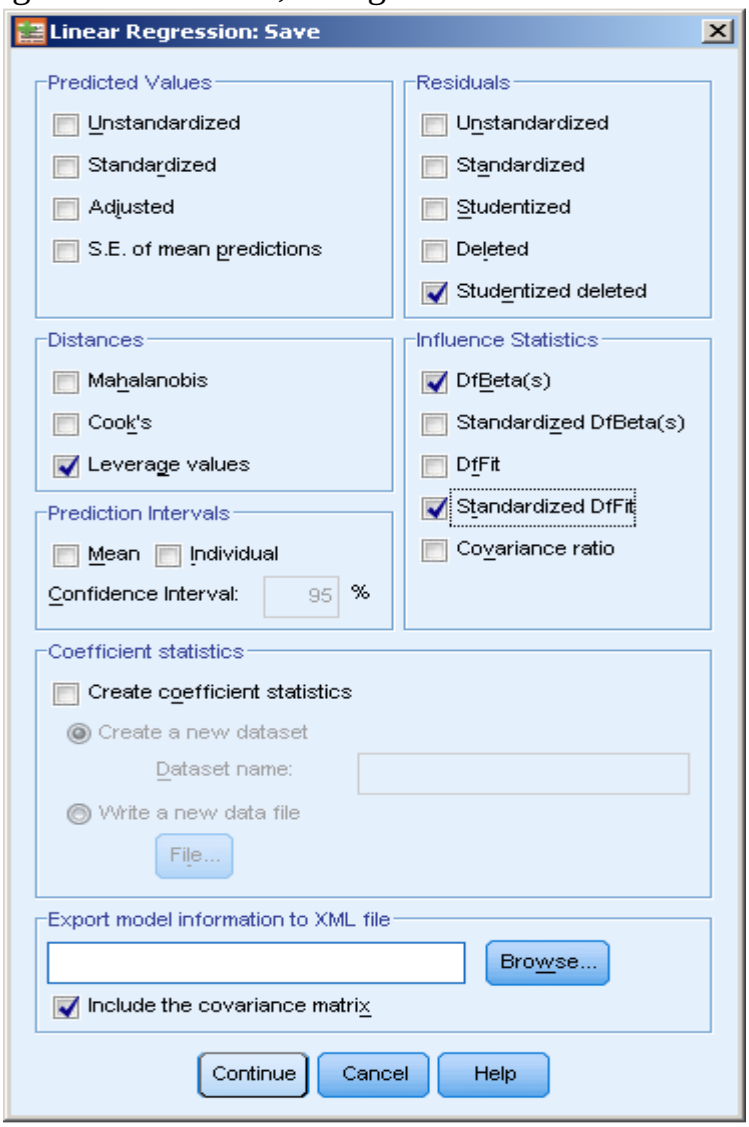
[10] I visse fremstillinger (fx Fox, 1991) angives gennemsnittet som antallet af uafhængige plus 1 delt med antallet af observationer. I SPSS anvendes dog den i teksten angivne beregningsmåde.

Residualet er som anført differencen mellem den observerede og estimerede *Y*-værdi. Denne forskel kan dog ikke umiddelbart anvendes som mål for en given observations grad af afvigelse. For det første varierer residualerne med skalaen for den afhængige variabel; de må derfor standardiseres, før de kan analyseres. Hvis man til dette formål beregner det såkaldte studentiserede residual (jf. Agresti og Finlay, 2008: 449), bliver der i standardiseringen desuden taget højde for, at observationer, der falder langt fra gennemsnittet på de uafhængige variabler (som målt ved hat-værdien), har en større variation end mere gennemsnitlige observationer. For det andet skal der også tages højde for, at observationer med høj leverage jo har muligheden for at påvirke de estimerede koefficienter, hvorved residualet vil formindskes i forhold til den værdi, man ville have fået, hvis den pågældende observation ikke havde været med i analysen og øvet sin indflydelse. For at modgå dette problem anvendes derfor det såkaldte slettede studentiserede residual, som fremkommer netop ved på skift at udelukke hver af observationerne fra estimeringen af modellen, når den pågældende observations residual skal beregnes. Normalt betragtes standardiserede residualer, der numerisk er over 3, som problematiske.

I SPSS genereres hat-værdier og slettede studentiserede residualer ved at klikke på Save-knappen i regressionsvinduet. Derved fremkommer den i Figur 4 viste dialogboks, hvor man markerer ved Leverage values og Studentized deleted. SPSS danner så to nye variabler kaldet LEV_1 og SDR_1,[11] der rummer henholdsvis hat-værdierne og de slettede studentiserede residualer.

[11] Vær opmærksom på, at der produceres nye variabler for residualer mv., hver gang analysen køres igen, med mindre man fjerner markeringerne i Save-boksen. Disse variabler får et nyt tal til sidst i deres navn (for eksempel SDR_2 osv.).

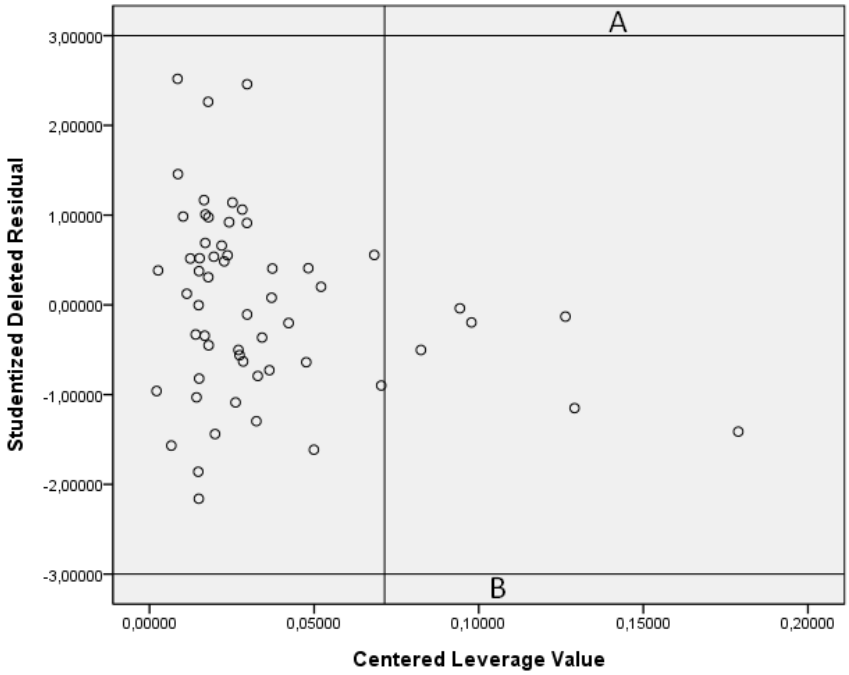
*Figur 4* Dialogboks for værdier, som gemmes i datasættet




Næste skridt er at plotte de to nye variabler mod hinanden i et simpelt scatterplot, hvor residualerne anbringes på *Y*-aksen og hat-værdierne på *X*-aksen (husk også at overføre navnevariablen til Label Cases by-feltet, så eventuelle afvigere kan identificeres). I vores eksempel ser det resulterende plot ud som vist i Figur 5. Linjerne er indlagt ved de kritiske grænser for de to mål, og de observationer, der vil være særligt problematiske, falder i de to felter A og B. Dette gøres ved i Chart Editoren at vælge Options | X-Reference Line og Y-Reference Line og specificere, at linjerne skal ligge ved de angivne grænseværdier (for x-aksen 2*gennemsnittet for Leverage Values, for y-aksen 3 og -3). Navnene på de enkelte observationer fås ved at vælge Elements | Data Lable Mode og derefter klikke på de

relevante observationer. Som det fremgår, er der ingen sådanne observationer for den her estimerede model.

*Figur 5* Slettede studentiserede residualer fra modellen plottet mod leverage-værdier




Såfremt der findes observationer i de to områder, vil disse skulle undersøges yderligere.[12] Til dette formål findes to mål, der begge kan bruges til at afgøre, om enkelte observationer øver en for stor indflydelse på analyseresultaterne. Som det vigtigste mål anvendes DFBETA, der for hver enkelt observation og hver af de estimerede koefficienter angiver, hvordan koefficienten ville ændre sig, hvis observationen blev udelukket fra analysen. DFBETA for en given observation og en given variabel er således defineret som den estimerede koefficient for variablens effekt minus en koefficient estimeret uden den pågældende observation. DFFIT er en art sammenvejning af DFBETA'erne for de forskellige variabler og er dermed et samlet mål for, hvor meget hver enkelt observation påvirker den samlede models fit. Målet med undersøgelsen er at finde ud af, om de problematiske observationer har DFBETA- eller DFFIT-værdier, der ligger markant over eller

[12] Hvis en eller flere observationer kun afviger på det ene af de to mål, kan det under alle omstændigheder være en god ide at undersøge disse nærmere, selvom de ikke ligger i de kritiske områder. Dette gælder særligt i små datasæt.

under de øvrige observationers. Er det tilfældet, udøver observationerne en så stor indflydelse på resultaterne, at det kan være problematisk for fortolkningen af modellen.

De to mål beregnes i SPSS ved at sætte kryds ved DfBeta(s) og Standardized DfFit i Save-boksen (jf. Figur 4). Derved konstruerer SPSS en DFFIT-variabel (SDF_1) og en DFBETA-variabel for hver koefficient (DFB1_1, DFB2_1 osv.) – inklusiv konstantleddet i modellen (DFB0_1). For at undersøge om de problematiske observationer afviger for meget fra de øvrige, kan man dernæst sortere datafilen og se på, hvilke observationer der har de højeste og laveste værdier. Dette gøres ved at vælge Data | Sort Cases… og derefter inspicere fordelingen i Data View-vinduet. Det kan ikke på forhånd specificeres, hvor store eller små værdier, der skal optræde, for at resultaterne bliver problematiske. For DFBETA'erne afhænger det af skaleringen for de enkelte variabler. For DFFIT angiver SPSS, at man betragter observationer med værdier over 2 gange kvadratroden af (antallet af uafhængige variabler +1)/n som problematiske.

Viser det sig ved disse undersøgelser, at observationer, der befinder sig i område A eller B i Figur 5, har for høje eller lave DFFIT- og/eller DFBETA-værdier, bliver spørgsmålet, hvad man så skal gøre ved det. I første omgang kan man skelne mellem problemer, der vedrører primære variabler og kontrolvariabler, hvor primære variabler er den/de variabler, som ud fra problemstillingen regnes for centrale, mens kontrolvariabler er variabler, der blot er medtaget for at sikre et korrekt estimat af den/de primære variablers effekt på *Y*. Hvad der er primære, og hvad der er kontrolvariabler, afhænger derfor af problemstillingen. Hvis eksempelvis problemstillingen er, hvorvidt velstand – alt andet lige – medfører et mere positivt syn på kvinder, vil velstand være den primære variabel, mens demokrati vil være en kontrolvariabel.

Såfremt man har en/flere primære variabler, kan man således se på DFBETA-værdierne for disse variabler, og hvis disse værdier ikke er alarmerende høje, vil man med forsigtighed kunne konkludere, at problemet ikke har et

bekymrende omfang.[13] Selvom DFBETA-værdierne indikerer problemer for primære variabler er det dog ikke nødvendigvis alarmerende. Man kunne således forestille sig, at den/de indflydelsesrige observationer udelukkende trækker estimatet i modsat retning af hypotesens påstand (fx mod en nul-effekt), hvilket vil betyde, at hypotesetesten bliver konservativ. I sådanne tilfælde er indflydelsesrige observationer mindre problematiske. Hvorvidt dette er tilfældet, kan afgøres ved en nærmere inspektion af de indflydelsesrige observationer.

Hvis der er tegn på indflydelsesrige observationer, der hjælper til at bekræfte hypotesen, er problemerne større, og man må forholde sig til det. I de fleste tilfælde vil det være en særdeles dårlig ide uden videre at fjerne de pågældende observationer fra datasættet: Man kan ikke bare slette de observationer, der ikke passer ind i modellen – det tenderer det videnskabeligt uhæderlige. I stedet kan man reestimere modellen uden de indflydelsesrige observationer og undersøge om resultaterne ændres substantielt. Hvis hypotesen holder uden de indflydelsesrige observationer, vil det støtte resultaternes gyldighed. Man bør dog også overveje, om der kan være tale om målefejl, som eventuelt kunne rettes, eller om modellen kan ændres – for eksempel ved at inkludere nye variabler – så den tager højde for de afvigende observationer. Som hjælp til disse overvejelser kan man inspicere observationer med store residualer (men ikke nødvendigvis stor leverage), for at se, om disse har fællestræk, der kan pege på en eller flere variabler, som med fordel kan inkluderes i modellen.

Alternativt kan man vælge at indskrænke gyldighedsområdet for modellen ved at udtage et antal observationer med bestemte kendetegn, hvis man mener, at modellen ikke kan indfange disse korrekt – man kan dog ikke bare lægge problematiske DFBETA- eller DFFIT-værdier til grund for en sådan sortering. Det afgørende er således, at der teoretisk kan argumenteres for, hvorfor bestemte observationer ikke kan dækkes af modellen. Men denne fremgangsmåde er under alle omstændigheder defensiv, idet man

[13] Når vi skriver "med forsigtighed", er det fordi et middelret estimat af den primære variabels effekt fordrer, at de estimerede effekter af kontrolvariablerne er nogenlunde korrekte.

lader sig presse bort fra den oprindelige ambition om en samlende model. Endelig skal man være opmærksom på, at i en stor stikprøve vil der ofte være enkelte observationer, som afviger fra det generelle mønster (i en stikprøve på 1000 vil man som udgangspunkt forvente op til 3 observationer med et residual numerisk større end 3). Man skal altså ikke lade sig skræmme af nogle få outliers til at lave meget omfattende ændringer af en i øvrigt velfungerende model – man skal med andre ord passe på med overfitting.

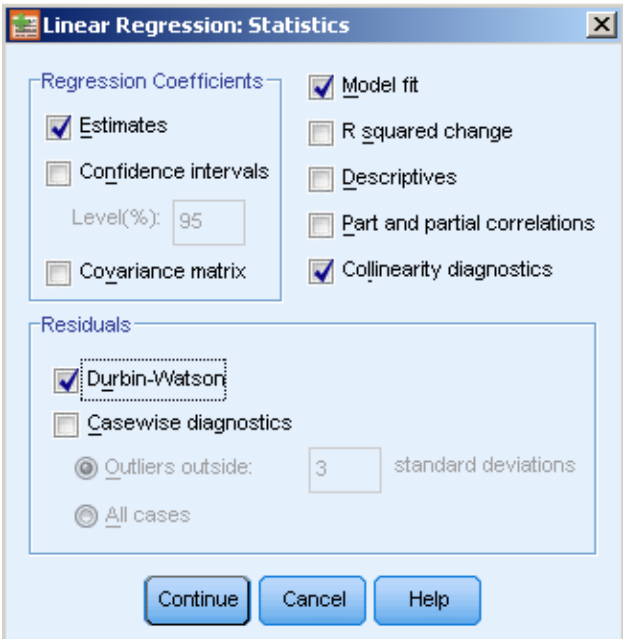
### *5 Fravær af stærk multikollinearitet*

Stærk multikollinearitet refererer til situationer, hvor to eller flere uafhængige variabler er stærkt indbyrdes forbundne. De hidtil behandlede forudsætningsbrud resulterer i biased koefficienter – altså koefficienter, der er systematisk for store eller små. *Konsekvensen* af stærk multikollinearitet er derimod usikre koefficienter – altså koefficienter, der kan være dårlige bud på populationsparametrene. Når to eller flere uafhængige variabler er stærkt forbundne har estimationsmetoden svært ved at skelne deres unikke effekter fra hinanden og dermed med at estimere præcise koefficienter. Denne usikkerhed afspejles i forhøjede standardfejl og dermed brede konfidensintervaller for de involverede koefficienter. Dermed er den direkte konsekvens, at selv stærke effekter fremstår som insignifikante. Med andre ord forhøjer stærk multikollinearitet sandsynligheden for type II-fejl.

Problemer med multikollinearitet kan *identificeres* på flere måder. Typiske indikatorer på multikollinearitet vil være, at de estimerede koefficienter ændrer sig voldsomt, når variable tages ind og ud af modellen, eller at alle de uafhængige variabler er insignifikante, mens *F*-testen for den samlede model er klart signifikant. Mere formelt kan problemet dog afdækkes med det såkaldte tolerancemål, der for en given variabel er defineret som 1 minus $R^2$ fra en regression med den pågældende variabel som afhængig og alle de øvrige oprindelige uafhængige variabler som uafhængige. Tolerance er dermed et mål for, hvor megen selvstændig variation der er i en given variabel i forhold til de øvrige variabler i analysen. Normalt vil der være

grund til bekymring, hvis tolerancemålet falder under 0,1.[14] I SPSS beregnes målet ved at markere Collinearity diagnostics i Statistics-boksen (se Figur 6). Derved produceres tilføjelsen til output-tabellen med koefficienterne vist i Tabel 2.

*Figur 6* Dialogboks for statistiske beregninger ved regressionsanalyse




*Tabel 2* Regressionskoefficienter med indcirklet kollinearitetsdiagnose

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | 1,629 | ,053 | | 30,616 | ,000 | | |
| | gnp1000 | ,008 | ,002 | ,508 | 4,545 | ,000 | ,716 | 1,397 |
| | dem99 Democracy - sum af Rights and Liberties 1999-2000 | -,019 | ,007 | -,296 | -2,650 | ,010 | ,716 | 1,397 |

a. Dependent Variable: q15c Disagree men better political leaders

For vores eksempel kan vi notere, at toleranceværdierne er pænt over det kritiske niveau, og vi skal derfor ikke bekymre os om multikollinearitet.

Havde vi fundet kritiske toleranceværdier, ville der have været forskellige løsningsstrategier. I nogle tilfælde kan man skelne mellem multikollinearitetsproblemer, der vedrører primære variabler, og problemer der vedrører kontrolvariabler. Multikollinearitet er mere problematisk for pri-

[14] Tolerancemålet følges ofte med Variance Inflation Factor (VIF), der blot er defineret som 1/tolerance.

mære variable end for kontrolvariable. Såfremt multikollinearitets-problemet kun vedrører kontrolvariabler, vil man kunne stole på koefficienten og *P*-værdien for den primære variabel og samtidigt være sikker på, at der er kontrolleret for kontrolvariablernes indflydelse. Man vil altså kunne konkludere på problemstillingen på et rimeligt sikkert grundlag på trods af multikollinearitet (se evt. Agresti & Finlay, 2008: 457). Derimod ville man – hvis man finder det nødvendigt at udtale sig om effekten af kontrolvariablene – være nødsaget til at tage forbehold for deres størrelse og *P*-værdier (se nedenfor). I relation til dette bør det bemærkes, at i tilfælde hvor der kun er to uafhængige variabler i modellen (som i eksemplet) vil tolerancеværdierne være identiske for de to variablers koefficienter, hvilket følger af definitionen af tolerancemålet. Ved flere uafhængige variabler vil tolerancemålet (oftest) variere på tværs af variablene.

I de tilfælde hvor tolerancemålet for den/de primære variabler indikerer, at der optræder multikollinearitet, vil problemers omfang afhænge af *P*-værdien for disse variabler. Da multikollinearitet forøger standardfejlene og dermed *P*-værdierne, vil en signifikant koefficient kunne tages som udtryk for, at variablen har en signifikant effekt i den retning, som fortegnet angiver. Størrelsen af effekten vil dog være usikker (hvilket er afspejlet i bredere konfidensintervaller). Hvis effekten derimod er insignifikant, vil man være på mere usikker grund. Da *P*-værdierne er forhøjede, kan man ikke udelukke, at koefficienterne er blevet insignifikante pga. multikollinearitet, og man risikerer altså at begå en Type II-fejl, hvis man konkluderer, at effekten er insignifikant.

Ovenstående viser, at multikollinearitet ikke nødvendigvis er et stort problem, men også at det ville være rart at være foruden. Spørgsmålet, der melder sig, er derfor, om man kan gøre noget for at minimere multikolliniariteten. En umiddelbar mulighed er at udtage en eller flere variabler af modellen og på den måde fjerne den stærke indbyrdes sammenhæng mellem variablerne. Man kunne eksempelvis fjerne visse kontrolvariable. Dette er dog sjældent anbefalelsesværdigt. Som fremhævet i afsnittet om omitted variable bias, vil udeladelse af variabler, der er kraftigt korreleret med inkluderede variable, skabe kraftigt bias. På den anden side er det

værd at huske på (også fra afsnit 1) at man trygt kan udelade kontrolvariabler, der er konsekvenser af en uafhængig variabel, da en sådan udeladelse ikke skaber omitted variable bias. I sådanne tilfælde vil der altså være god grund til at fjerne kontrolvariablen og på den måde minimere multikollineariteten.

Agresti & Finlay (2008: 457f) fremhæver indekskonstruktion som anden løsning – altså at samle de problematiske variabler i et indeks. Dette vil eliminere problemerne med multikollineariteten. I mange tilfælde vil det dog være en pseudo-løsning, da problemet med multikollinearitet er, at man ikke kan skelne effekten af stærkt forbundne variabler fra hinanden. Hvis variablerne er samlet i et indeks, kan man stadig ikke skelne deres effekt fra hinanden.[15]

Endelig er det også den (desværre i mange tilfælde urealistisk) mulighed at indsamle flere observationer, der kan gøre det lettere at skelne mellem betydningen af de uafhængige variabler.

### *6 Fravær af autokorrelation*

Autokorrelation refererer til situationer, hvor en stor del af fejlleddene fra observationer, som ligger tæt på hinanden i tid eller rum, er korrelerede. *Konsekvensen* af autokorrelation er, at standardfejlene er et ringe estimatet af stikprøvemålsfordelingerns standardafvigelser. Det betyder altså, at de opnåede *P*-værdier er fejlagtige. Oftest vil standardfejlene og *P*-værdierne være underdrevne, hvorfor vi kan risikere at tillægge givne koefficienter en signifikans, de ikke fortjener. Det samme gælder i forhold til $R^2$, der også kan blive overestimeret. Det skal dog bemærkes, at autokorrelation ikke påvirker estimaternes middelrethed – dvs. de estimerede koefficienter bliver ikke biased. Men de beskrevne problemer i relation til inferensen kan være næsten lige så ødelæggende.

[15] Bemærk at dette ikke skal opfattes som en advarsel mod at konstruere indeks. Refleksive indeks er oftest en fornuftig måde at øge validiteten og reliabiliteten på, men dette valg bør være drevet af teoretiske og substantielle overvejelser foretaget inden analysen og ikke af tegn på multikollinearitet.

Autokorrelation *opstår*, når observationerne ikke er uafhængige af hinanden, fx fordi de er påvirkede af de samme bagvedliggende faktorer eller fordi én observation påvirker andre observationer. Dette forstås nemmest i forbindelse med tidsseriedata – altså data hvor observationerne er egenskaber ved den samme enhed (fx et individs indkomst, en kommunes andel af flygtninge eller en lands grad af demokrati) på forskellige tidspunkter. Det er oplagt, at et individs indkomster i to på hinanden følgende år ikke er uafhængige af hinanden, hvilket i en regressionsanalyse kan resultere i, at residualerne er positivt korrelerede. Ved positiv korrelation vil et stort positivt(/negativt) residual følges af et stort positivt(/negativt) residual, mens der ved negativ korrelation vil være tale om, at positive residualer følges af negative (og omvendt). Hvis en stor del af residualerne fra to på hinanden følgende observationer er korrelerede, vil der være autokorrelation i modellen.

Autokorrelation i forbindelse med tidsseriedata kaldes også *seriel autokorrelation*. Derudover taler man også om *spatial autokorrelation*, hvilket refererer til situationer, hvor residualerne fra observationer, der ligger tæt på hinanden i rum, er korrelerede. Et eksempel kunne være residualerne fra en regressionsanalyse af et udsnit af verdens lande. Det er nærliggende at antage at fx EU-landene ligner hinanden, fordi de dels er underlagt en række fælles bestemmelser, men også fordi de påvirker hinanden fx igennem diffusion. EU-landene er altså ikke uafhængige af hinanden, hvilket kan resultere i, at residualerne er korrelerede.

Generelt opstår begge former for autokorrelation, når observationerne ikke er tilfældigt udvalgte, mens observationer, der er tilfældigt udvalgt (fra en tilstrækkelig stor population), kan regnes for at være uafhængige, hvorfor autokorrelation ikke opstår. *Det betyder, at man kun bør bekymre sig om autokorrelation, når man ikke har en tilfældigt udvalgt stikprøve*. I fx tidsseriedata udgøres observationerne af egenskaber over tid, og disse er yderst sjældent tilfældigt udvalgt. Derfor vil man forvente (seriel) autokorrelation, når man arbejder med tidsseriedata. Når man arbejder med lande som analyseenheder, vil man sjældent udtrække en tilfældig stikprøve, hvorfor man oftest vil forvente spatial autokorrelation i forbindelse med landedata.

*Identifikation og håndtering af spatial autokorrelation* kræver særlige test og metoder, der falder udenfor pensum. Det betyder, at man (desværre) ofte må nøjes med at antage, at der ikke forefindes spatial autokorrelation ved ikke-tilfældigt udvalgte data. Dog kan man eventuelt diskutere, om der er stærke grunde til at forvente, at observationerne ikke er uafhængige. Hvis dette er tilfældet, må der tages forbehold for de beregnede *P*-værdier.

*Identifikation af seriel autokorrelation* er derimod en del af pensum, og vi vil derfor koncentrere os om seriel autokorrelation. Det betyder, at den nedenfor beskrevne test KUN gennemføres, hvis der er tale om tidsseriedata.

Seriel autokorrelation kan identificeres med Durbin-Watson-testen, der i SPSS produceres ved at markere i Statistics-boksen. Resultatet bringes som en tilføjelse til Model Summary-tabellen i outputtet. Durbin-Watsons *d*, som testen hedder, kan antage værdier fra 0 til 4. Værdier tæt på 0 indikerer en høj positiv autokorrelation, værdier tæt på 4 tyder på en høj negativ autokorrelation, mens værdier omkring 2 tyder på, at der ikke er problemer med autokorrelation. Da data i eksemplet ikke er tidsseriedata, er det ikke relevant at diskutere resultatet her. Afdækker man problemer i en given sammenhæng, findes der en række forskellige tiltag, der – som resten af tidsserieteknikkerne – er emner, der kan studeres på overbygningen.

## *7 Normalfordelte fejlled*

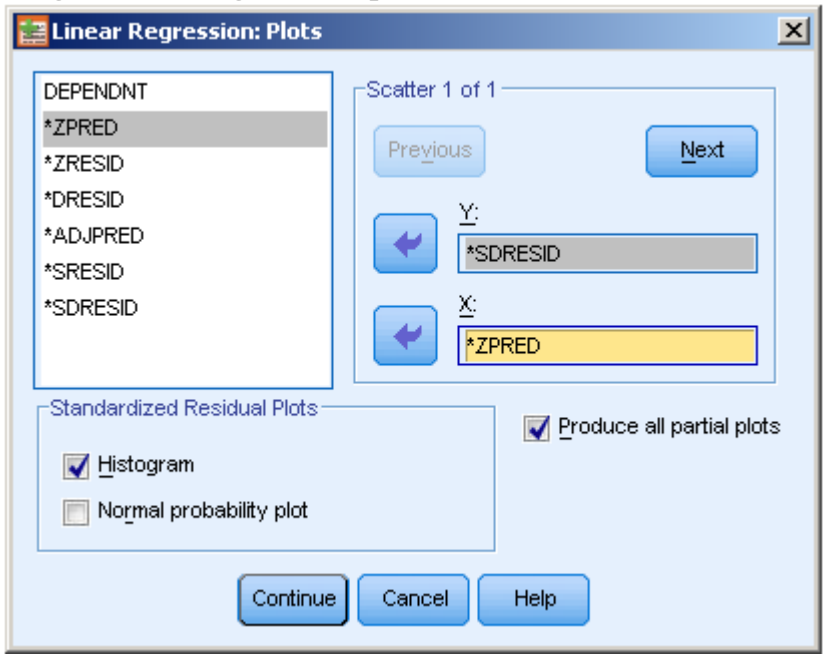
Som det er tilfældet med forudsætningen om fravær af autokorrelation, er forudsætningerne om normalfordelte fejlled og konstant spredning på fejlleddene for alle værdier af de uafhængige variabler (som behandles under punkt 8) som sådan ikke nødvendige for at kunne estimere modellens koefficienter korrekt, sådan som det er tilfældet med de fire første af de forudsætninger, vi har set på hidtil. Hvor brud på disse fire forudsætninger således kan resultere i bias i de estimerede koefficienter og dermed en misvisende model, vil brud på forudsætningerne vedrørende residualernes fordeling ikke berøre de estimerede koefficienter. Derimod vil brud på disse forudsætninger kunne ødelægge muligheden for at foretage inferens. Fejl-

leddenes fordeling er således grundlaget for at kunne udføre signifikansberegninger for koefficienterne – det er den fordeling, der ligger til grund for beregningen af *P*-værdierne – og hvis forudsætningerne på dette punkt ikke er opfyldt, vil det altså være disse beregninger og ikke de estimerede koefficienter, der kan blive problematiske.

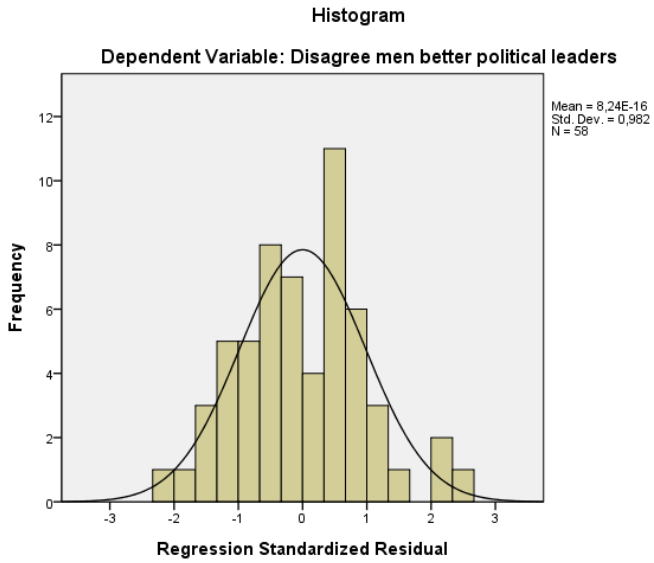
Forudsætningen om, at fejlleddene følger normalfordelingen, er kun strengt nødvendig i små stikprøver, hvor den centrale grænseværdisætning ikke gælder, og det derfor er nødvendigt at sikre, at residualerne følger normalfordelingen for at kunne benytte *t*-test til signifikansberegningerne. Testen udføres dog normalt også, når man arbejder med store stikprøver, idet kraftige afvigelser fra normaliteten kan indikere, at der er udeladt vigtige variabler, kan identificere potentielle outliers eller kan antyde andre specifikationsproblemer. De følgende test bør derfor gennemføres uanset datasættets størrelse.

Ligesom ved vurderingen af linearitetsforudsætningen bruges stikprøven til at vurdere, om forudsætningen om normalfordelte fejlled er opfyldt. Man bruges således residualerne, der er stikprøvens pendant til populationens fejlled, til at vurdere fejlleddenes fordeling. Den overordnede fordeling af residualerne undersøges ved at se på et histogram over deres fordeling, der gerne skal ligge så tæt på normalkurven som muligt, da det sandsynliggør, at fejlleddene er normalfordelte. Histogrammet produceres ved at markere i Plots-boksen som vist i Figur 7. Derved produceres histogrammet vist i Figur 8.

*Figur 7* Dialogboks for regressionsplot



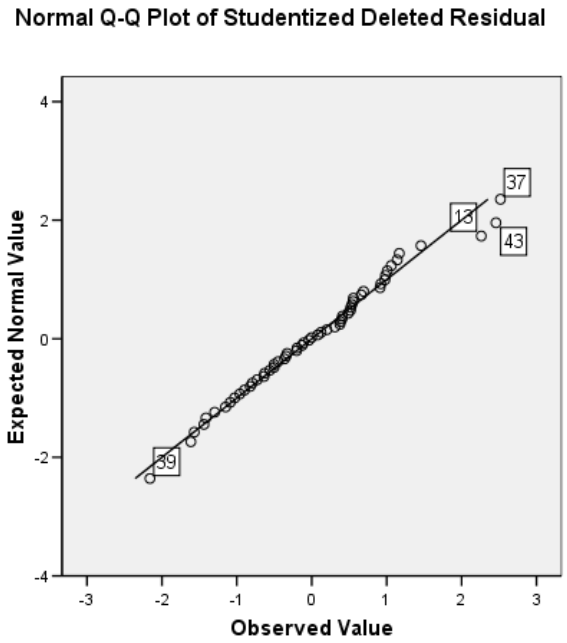
*Figur 8* Histogram over fordelingen af standardiserede residualer




I eksemplet er der, som det fremgår, visse afvigelser fra normalfordelingen. Disse skyldes dog i nogen udstrækning, at histogrammers udseende er følsomme over for antallet af søjler. Desuden er fordelingen – om ikke helt klokkeformet, så – klart unimodal. Der er således ikke tegn på, at der er 'klumper' af data med større afvigelser, hvilket ville være en klar indikation af specifikationsproblemer. Yderligere kan vi observere, at der ikke optræder alvorlige outliers (hvilket vi også tidligere har konstateret).

Næste skridt i undersøgelsen af normalitetsforudsætningen er at konstruere et såkaldt Q-Q-plot af de slettede studentiserede residualer, som vi tidligere har fået produceret i forbindelse med undersøgelsen af ekstreme observationer. Denne type plot kan ligeledes bruges til at identificere afvigelser fra normaliteten, men har den yderligere fordel, at det muliggør en identifikation af de observationer, som afviger derfra. Q-Q-plottet produceres ved at vælge Analyze | Descriptive Statistics | Q-Q Plots… og indsætte variablen med de slettede studentiserede residualer (typisk SDR_1) som plotvariabel. Derved produceres et plot som vist i Figur 9.

*Figur 9* Q-Q plot til undersøgelse af, om de slettede studentiserede residualer er normalfordelte




Q-Q-plottet er konstrueret således, at observationer, der falder på den indtegnede 45°-linje, er normalfordelte. Som det fremgår, følger langt hovedparten af residualerne for vores eksempel pænt normalfordelingen – der er ingen voldsomme afvigelser fra linjen. Dog viser der sig visse afvigelser ved enderne af linjen, hvilket er helt normalt, hvis der optræder outliers. Fordelen ved plottet er, som nævnt, netop at disse outliers kan identificeres. Dette gøres ved at dobbeltklikke på plottet, så Chart Editoren åbnes. Her vælges Elements | Data Label Mode, og man markerer derefter de afvigende observationer. De numre, der vises, henfører til observationernes placering i datafilen (hvordan den end er sorteret), og man kan derefter identificere de konkrete observationer (det drejer sig i vores tilfælde om Colombia, Mexico, Nigeria og Peru). Der er dog ikke tale om voldsomt store afvigelser, og det vil derfor næppe være tilrådeligt at foretage korrektioner (jf. nedenfor) på dette grundlag.

Næste spørgsmål bliver så, hvad der kan gøres ved kraftige afvigelser fra normaliteten. En mulighed er at transformere den afhængige variabel *Y*, så den del, der afviger fra normaliteten, tones ned. Hvis der er problemer med store afvigelser i enderne af 45°-linjen, kan det være en mulighed at bruge logaritmen af *Y*, som vil nedtone ekstremerne og resultere i en fordeling,

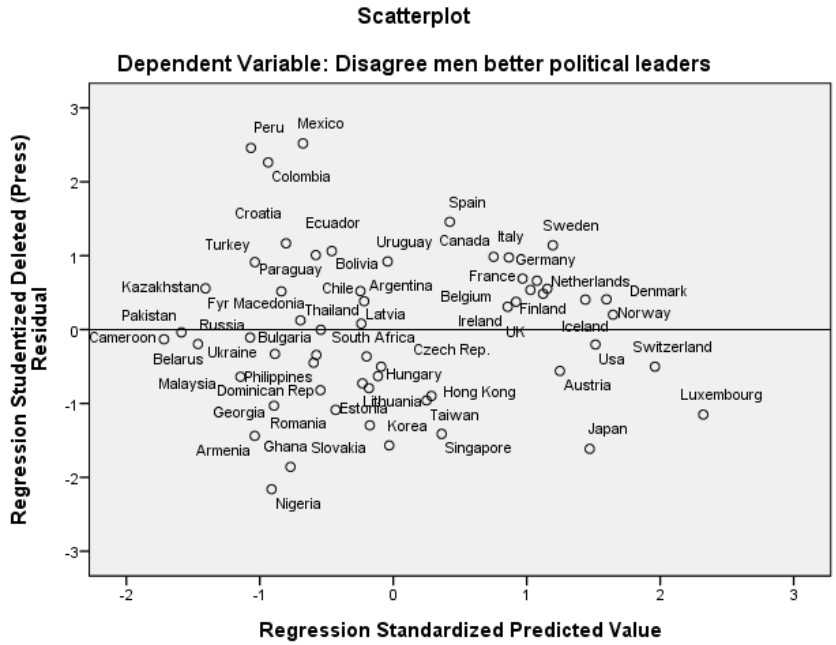
der ligger tættere på normalfordelingen. Hvis der viser sig en bimodal fordeling i histogrammet, bør man overveje at inddrage yderligere variabler, der kan indfange de to 'klumper' i data, således at modellen kan redegøre for den konstaterede opdeling af data. Det er også i denne forbindelse nødvendigt at fundere sine overvejelser over, hvad der kan gøres for at løse problemet, i teorien, således at eventuelle transformationer og/eller ændringer af modellen har en teoretisk basis. Som nævnt ovenfor, skal man dog passe på ikke at lade forholdsvis få ekstreme observationer diktere omfattende ændringer af en i øvrigt velfungerende model. Man skal her desuden huske, at analyser udført på store datasæt er ganske robuste over for afvigelser fra normaliteten, idet den centrale grænseværdisætning sikrer, at stikprøvemålsfordelingerne følger normalfordelingen. Se endvidere overvejelserne om anlæggelsen af en mere konservativ fortolkning af *P*-værdierne nedenfor.

### *8 Homoskedasticitet*

Forudsætningen om homoskedasticitet – eller varianshomogenitet – betyder, at der skal være den samme spredning i fejlleddene for alle værdier af de uafhængige variabler. Hvis forudsætningen ikke er overholdt, og der altså optræder heteroskedasticitet, er *konsekvensen*, at beregningen af *P*-værdierne bliver usikker, idet der jo så vil være forskel på fejlleddenes spredning afhængigt af, hvilket niveau af variablerne man tager i betragtning. Dermed bliver en samlet signifikansberegning umulig.

Til identifikation af heteroskedasticitet bruges igen residualernes fordeling, som inspiceres for at undersøge to spørgsmål: 1) Hvorvidt residualernes varians vokser med den forudsagte værdi på den afhængige variabel, 2) Hvorvidt residualernes varians ændrer sig med værdierne på de uafhængige variabler. I SPSS kan det første spørgsmål besvares ved at plotte den standardiserede forudsagte værdi (*ZPRED på *X*-aksen) mod de slettede studentiserede residualer (*SDRESID på *Y*-aksen), hvilket kan gøres i Plots-boksen som vist i Figur 7 ovenfor. Derved produceres plottet vist i Figur 10 (linjen indlægges som beskrevet ovenfor).

*Figur 10* Slettede studentiserede residualer plottet mod standardiserede forventede værdier




Hvis observationerne ligger i et pænt bånd omkring 0-linjen, indikerer det, at forudsætningen er opfyldt. Dette synes at være tilfældet i eksemplet her (dog kan det bemærkes, at de tidligere identificerede afvigere også viser sig her). Det vil derimod være et klart tegn på problemer, hvis der optræder en trompetform i fordelingen, således at variationen altså er meget større i den ene ende end i den anden.

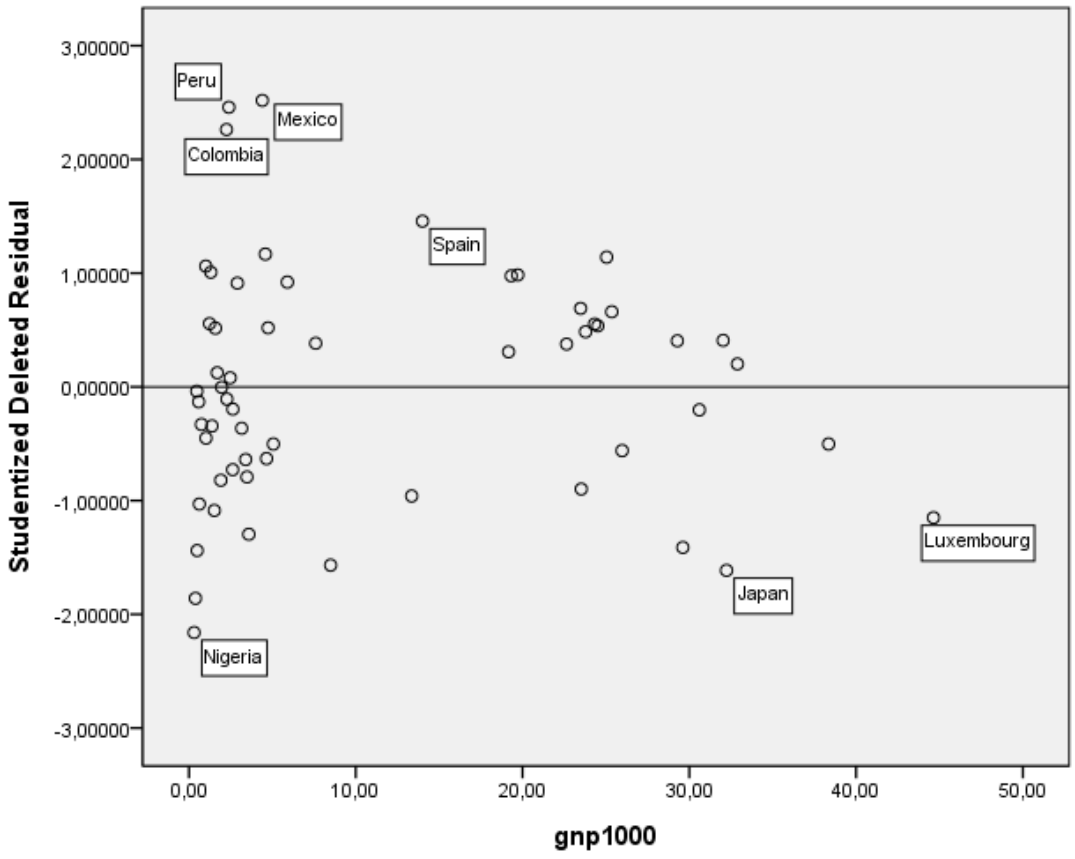
Uanset svaret på det første spørgsmål bør man også undersøge det andet. Dette gøres ved at plotte de slettede studentiserede residualer mod hver af de uafhængige i et almindeligt scatterplot.[16] Mens fordelingen hen over demokrativariablen er rimelig, viser der sig klare tegn på problemer i relation til BNP-variablen. Som det fremgår af Figur 11, optræder der her en trompetform, hvilket er et klart brud på forudsætningen om varianshomogenitet.

Forklaringen på det konstaterede problem er den samme, som blev noteret i relation til undersøgelsen af linearitetsforudsætningen: BNP-variablen (el-

[16] Der findes også en mere formel test for denne type heteroskedasticitet – den såkaldte White-test. Denne test kan enkelt udføres med SPSS. For beskrivelse af testen, se for eksempel Studenmund (2001: 360-362).

ler variabler tæt beslægtede hermed) har ofte en logaritmisk sammenhæng med andre variabler, og hvis de inddrages utransformerede i analyser, resulterer det i en række problemer som for eksempel heteroskedasticitet.

*Figur 11* Slettede studentiserede residualer plottet mod en uafhængig variabel




Løsningen på problemet i nærværende analyse ligger således også lige for: BNP-variablen skal transformeres ved at tage logaritmen af den. Dermed opnås en højere grad af linearitet i sammenhængen, og det konstaterede problem mindskes. I andre tilfælde kan det være mindre åbenlyst, hvad der kan gøres mod problemer med heteroskedasticitet. Hvis det er en bestemt variabel, der skaber problemerne (som i vort eksempel) vil man ligesom med multikollinearitet kunne skelne mellem primære og kontrolvariable; hvis problemerne vedrører en kontrolvariabel, kan man stadig stole på *P*-værdien for den/de primære variabler.

Der findes også andre løsninger, der desværre ligger uden for pensum: Weighted Least Squares (WLS) estimation eller anvendelse af heteroskedasticitetskonsistente (eller robuste) standardfejl. I fald det ikke er muligt at benytte sig af disse tiltag, vil man i stedet kunne anlægge en mere konservativ fortolkningsstrategi i relation til de beregnede *P*-værdier. Når disse således må antages at være tvivlsomme som følge af forudsætningsbruddet, må man som en art kompensation kræve, at de er endnu lavere end ellers, for at en given koefficient accepteres som signifikant. Så hvis man for eksempel arbejder på et $\alpha$-niveau på 0,05, vil man kun godtage *P*-værdier under for eksempel 0,01. Hvor strenge krav, man vil stille, er i princippet arbitrært, men som tommelfingerregel kan det angives, at jo større problem, jo lavere *P*-værdi må der kræves for at acceptere en effekt som signifikant. Disse overvejelser er også relevante i relation til problemer med normalitetsforudsætningen.

## *9 Datasættets karakter*

For at kunne foretage valid inferens til en eventuel bagvedliggende population er det vigtigt, at datasættet er sammensat på en måde, der opfylder sandsynlighedsteoriens krav om uafhængighed mellem observationerne. Dette betyder i udgangspunktet, at stikprøven skal være udtrukket simpelt tilfældigt. Simpel tilfældig udtrækning sikrer fravær af autokorrelation (jf. afsnit 6) og en entydig fortolkning af *P*-værdierne (se nedenfor).

Der er ikke noget mindstekrav for stikprøvestørrelsen for at kunne udføre en regressionsanalyse – matematisk kan den gennemføres med blot to observationer. Selvsagt påvirkes standardfejlene på de estimerede koefficienter dog af stikprøvens størrelse, og i små stikprøver vil standardfejlene være så store, at det kan være vanskeligt at opnå signifikante resultater, hvilket jo begrænser nytten af analysen afgørende.

I forhold til stikprøven skal man desuden holde sig for øje, at den opstillede model i udgangspunktet kun er gældende inden for det variationsområde på de inddragede variabler, der er afspejlet i stikprøven. Det er således ikke muligt uden grundig teoretisk baseret diskussion at ekstrapolere modellens resultater til observationer, der falder uden for dette område. Med andre ord

er modellens forudsigelsesrum afgrænset til de værdier af de analyserede variabler, der er repræsenteret i stikprøven.

I vores eksempel anvendes som bekendt et datasæt bestående af observationer for 60 lande, der ikke kan antages at være udtrukket tilfældigt fra populationen af alle verdens lande. Dermed er det ikke muligt at foretage inferens til denne population på baggrund af analyserne her. I stedet kan man betragte datasættet som populationsdata for de 60 lande. Dermed kunne der argumenteres for, at det ikke er relevant at diskutere inferensproblemer overhovedet, men som anført af Søren Risbjerg Thomsen (1997) i noten om signifikanstest i ikke stikprøvesituationer, kan sandsynlighedsteoretisk baserede test betragtes som konservative test af de fundne sammenhænge, og det er ud fra denne vinkel, resultaterne her skal vurderes.

## Særlige forhold vedrørende variansanalyse/regression med dummyvariabler

Inddragelsen af kategoriske uafhængige variabler enten alene eller sammen med metriske variabler i en regressionsanalyse giver anledning til en række særlige forhold i relation til forudsætningerne, som herunder skal omtales.

Til illustration inddrages to kategoriske variabler i eksemplet. Det drejer sig for det første om en variabel, som angiver, hvorvidt der er en stor ikke-religiøs befolkningsgruppe i landet (variabel q9_9 er omkodet, så lande med en ikke-religiøs andel over 0,2 får værdien 1, mens alle andre får værdien 0). Dernæst inddrages også en variabel, der angiver, hvilken af de seks regioner, landene tilhører (variablen region med kategorierne Vesteuropa, Østeuropa, Vest- og Sydafrika, Nordamerika, Latinamerika og Sydøstasien; sidstnævnte fungerer som referencegruppe). Den estimerede model kommer dermed til at se ud som følger:

$$Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6 + \hat{\beta}_7 X_7 + \hat{\beta}_8 X_8 + e,$$

hvor $X_1$ og $X_2$ er de to metriske variabler analyseret tidligere, $X_3$ er dummyen for religionsvariablen, $X_4$ til $X_8$ er dummyerne for regionsvariablen og $e$ som før residualet. I det følgende gennemgås punkterne 1-7 med fokus på de forhold, der er anderledes for de kategoriske variabler i analysen.

### *1 Modelspecifikation (1-3)*

Forudsætningerne og fravær af omitted variable bias og endogenitet ændrer sig ikke. Derimod er det ikke relevant at undersøge, om der er en lineær sammenhæng mellem de kategoriske variabler og den afhængige. Det giver således ikke mening at forvente en lineær sammenhæng, og modellen er baseret på, at man undersøger, om der er forskelle på gennemsnittene på den afhængige variabel i kategorierne på de uafhængige kategoriske variabler. Forefindes der metriske variabler i modellen, vil man selvsagt stadig undersøge, om der er en lineær sammenhæng med disse og den afhængige variabel. Dette udføres på samme måde som ovenfor – dog med inklusion af de kategoriske variabler i modellen.

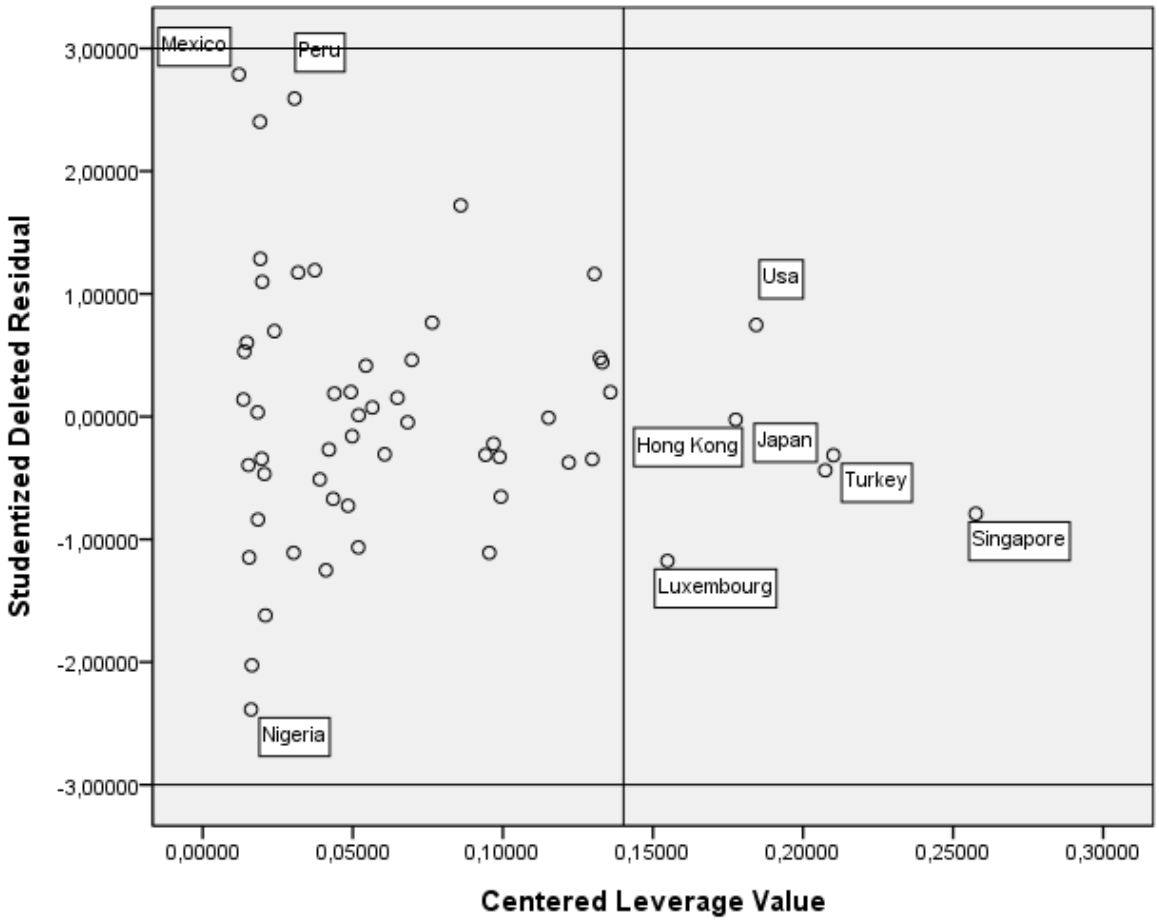
### *2 Fravær af indflydelsesrige observationer*

Undersøgelsen af indflydelsesrige observationer har normalt ikke fyldt så meget i relation til kategoriske variabler (Agresti og Finlay nævner det således slet ikke i gennemgangen af forudsætninger for variansanalyse). Dette skyldes formentlig, at problemerne i relation hertil er knap så alvorlige, når man arbejder med kategoriske variabler, idet der ikke estimeres en linje, som kan 'drejes' af ekstreme observationer. Det er dog klart, at ekstreme observationer også kan skabe bias i analyser af kategoriske variabler, hvor fokus, som anført, er på forskelle i gennemsnittene mellem grupperne. Selve målet gennemsnit er således følsomt over for outliers, der vil kunne forrykke estimaterne i forhold til hovedtendensen i data. Desuden vil tilstedeværelsen af outliers betyde, at signifikanstestene får vanskeligere ved at identificere sammenhænge, idet variationen i data øges. Det bør altså også undersøges, om der er indflydelsesrige observationer i forbindelse med variansanalyse.

Gennemføres analysen uden inddragelsen af metriske variabler, er inspektionen af leveragemålet ikke relevant, idet en observations leverage i dette tilfælde udelukkende afgøres af størrelsen på den gruppe, hvortil den hører. I sådanne analyser vil man derfor nøjes med at se på det slettede studentiserede residual, hvor værdier, der numerisk er over 3 betragtes som problematiske og udsættes for særlig analyse med DFBETA'er og DFFIT som beskrevet ovenfor.

Inddrages der derimod også metriske variabler, foretages analysen helt som beskrevet ovenfor. I eksemplet produceres et residual/leverage-plot (Figur 13), hvoraf det fremgår, at der ikke er observationer placeret i de problematiske regioner. I det tilfælde at der identificeres problemer, vil overvejelserne være de samme som de ovenfor beskrevne.

*Figur 13* Slettede studentiserede residualer plottet mod leverage-værdier




### *3 Fravær af stærk multikollinearitet*

Multikollinearitet kan også være et problem, når man analyserer kategoriske variabler. Kategoriske variabler kan således også være så stærkt forbundne med andre variabler i analysen, at det bliver vanskeligt at identificere deres unikke effekter. Særligt skal det bemærkes, at man vil skabe perfekt multikollinearitet imellem dummyvariablerne for en given kategorisk variabel, hvis man inkluderer en dummyvariabel for hver kategori i stedet for at udelade en af dem som referencekategori. Hvor dette kan forhindres ved almindelig omtanke, er det desværre forbundet med større vanskeligheder at undersøge eksistensen af multikollinearitet i relation til samlede kategoriske variabler, der i analysen repræsenteres ved dummyvariabler.

Således kan tolerancemålet ikke anvendes, når en kategorisk variabel er splittet op i *flere* dummyvariabler, der jo samlet repræsenterer variablen. Der er udviklet en metode, som generaliserer tolerancemålet (eller mere præcist VIF) til også at omfatte denne situation (se Fox og Monette, 1992), men denne er ganske kompliceret og lader sig ikke umiddelbart implementere i SPSS. Vi må derfor benytte en mere lav-teknisk fremgangsmåde til at vurdere, om der er for stort sammenfald mellem observationernes værdier på de kategoriske variabler i analysen. Et sådant sammenfald kan nemt afdækkes i en simpel krydstabel, hvor man skal være opmærksom på, hvorvidt der er en stærk sammenhæng mellem placeringerne på de givne variabler. Dette vil afsløres ved, at der er mange tomme celler i tabellen, således at alle observationerne i en given kategori på den ene variabel er placeret i den samme kategori på den anden variabel. Har man flere end to kategoriske variabler i analysen, kan multikollinearitetsundersøgelsen foretages for variablerne parvist. Arbejder man med ordinalskalerede variabler, kan det også være en mulighed at benytte $\gamma$ for at afgøre, om sammenhængen mellem to variabler er for stærk. Dette kan siges at være tilfældet, hvis målene ligger over 0,9.

Multikollinearitet for metriske variabler og variabler, der indgår i analysen med en *enkelt* dummyvariabel (for eksempel køn eller religionsvariablen i eksemplet), undersøges som normalt (dvs. ved at undersøge tolerancemålene fra analyser hvor alle de uafhængige variabler er inkluderet). I eksemplet viste der sig ingen problemer med hverken den ene eller den anden type variabler.[17]

### *4 Fravær af autokorrelation*

Som i analyser med metriske variabler udføres denne test kun i forbindelse med tidsseriedata.

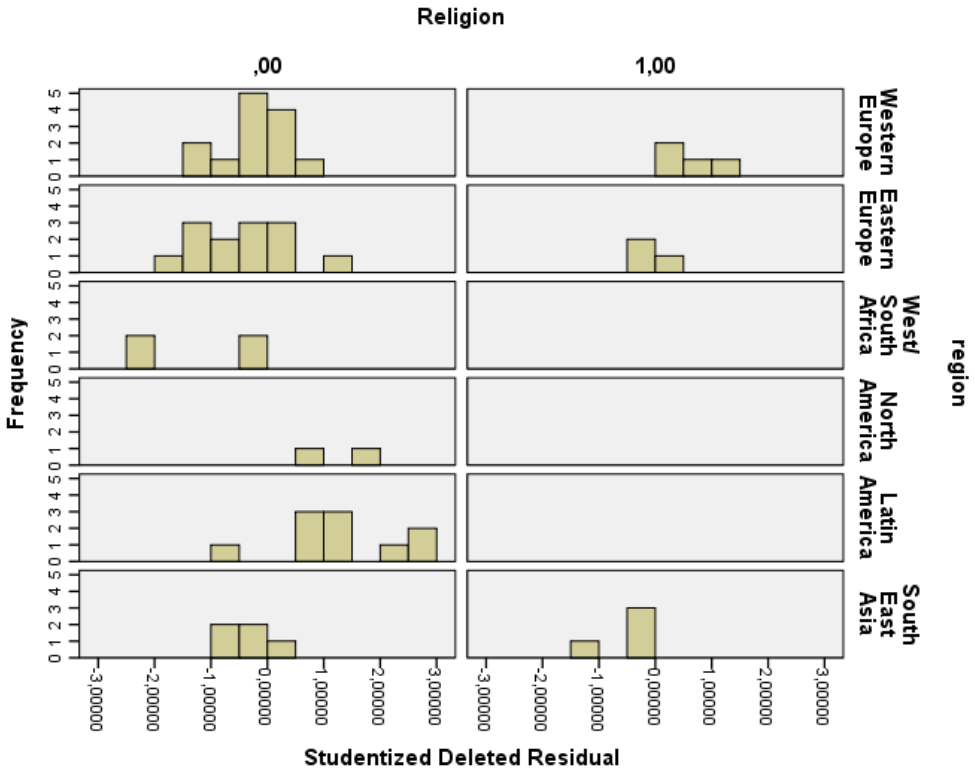
### *5 Normalfordelte residualer*

Forudsætningen om normalfordelte fejlled gælder også i relation til kategoriske variabler og det af de samme grunde som for metriske variabler. De

[17] Inklusionen af interaktionsled i analysen kan give anledning til særlige forhold i relation til spørgsmålet om multikollinearitet – se Sønderskov (2011).

overordnede test foretages som normalt ved hjælp af histogrammet og Q-Q-plot, men det er desuden nødvendigt at undersøge, om residualerne nogenlunde følger normalfordelingen inden for hver kombination af de kategoriske variabler, der inddrages i analysen. I relation til eksemplet betyder det for eksempel, at residualerne samtidigt skal være nogenlunde normalfordelte inden for alle regionerne og de to niveauer på religionsvariablen.

*Figur 14* Fordeling af slettede studentiserede residualer i forskellige kategorier




Testen udføres ved at lave histogrammer over fordelingen af de slettede studentiserede residualer[18] for alle disse kombinationer af de kategoriske variabler. Dette gøres ved at vælge Graph | Legacy Dialogs | Histogram… og anbringe de kategoriske variabler under Rows og Columns under Panel by. Derved produceres i eksemplet det i Figur 14 viste plot.

[18] Arbejder man med SPSS GLM-rutine i stedet for regressionsrutinen, kan man her bruge de slettede residualer, der kan produceres i GLMs Save-funktion, hvor de studentiserede slettede residualer desværre ikke er tilgængelige.

De tomme celler i figuren reflekterer, at der ikke findes observationer i datasættet med de givne variabelkombinationer. Figuren viser desuden, at der forekommer visse afvigelser fra normaliteten (om end fordelingerne i betragtning af det begrænsede antal observationer er ganske pæne). Konsekvenserne af disse problemer er en forøget usikkerhed om de estimerede koefficienter: Mens de stadig forventes at være middelrette, vil de beregnede *P*-værdier være tvivlsomme.

I de tilfælde hvor der ikke kan foretages yderligere observationer, der kunne forstørre datasættet, vil 'løsningen' på problemet med den forøgede usikkerhed være at anlægge en mere konservativ vurdering af *P*-værdierne for de forskellige koefficienter og/eller variabler, således som det er diskuteret ovenfor. I eksemplet ville et sådant skærpet krav til *P*-værdierne ikke ændre afgørende på fortolkningen (koefficienterne ikke vist). En mulig anden fremgangsmåde ville være at kollapse kategorier på variablerne, således at der ville blive færre sammenligninger at foretage og flere observationer i hver celle. Denne fremgangsmåde garanterer dog ikke varianshomogenitet, og man risikerer desuden at miste muligheden for at undersøge sin problemstilling, hvis de interessante kategorier kollapses.

### *6 Homoskedasticitet*

Forudsætningen om varianshomogenitet gælder ligeledes i relation til kategoriske variabler. Spørgsmålet i denne forbindelse er, hvorvidt variansen er den samme for alle kombinationer af de kategoriske variabler. Hvis dette ikke er tilfældet, opstår der problemer i relation til inferensen, idet usikkerheden på estimaterne så vil variere på tværs af de forskellige grupper og dermed undergrave signifikanstestene, der er baseret på, at der er den samme variation i hele datasættet (jf. også diskussionen ovenfor).

Testen gennemføres lettest med Levenes test, som desværre ikke er indbygget i SPSS' regressionsrutine. Det er derfor nødvendigt at benytte GLM-rutinen for at gennemføre denne test. Her fås den til gengæld nemt ved at markere ved Homogeneity tests i Options…-boksen, hvor man også bør bede om Descriptive statistics. I eksemplet produceres en Levenes test, der ser ud som vist i Tabel 3. Da *P*-værdien ikke falder under kon-

ventionelle $\alpha$-niveauer, kan det konstateres, at der ikke er grund til at antage, at fejlleddenes varians afviger for meget fra hinanden. Dog kan man i descriptives-tabellen (ikke vist) notere, at den største standardafvigelse i en gruppe er mere end dobbelt så stor som den mindste, hvilket normalt regnes som tegn på heteroskedasticitet.

*Tabel 3* Levenes test for varianshomogenitet

**Levene's Test of Equality of Error Variances[a]**

Dependent Variable: q15c Disagree men better political leaders

| F | df1 | df2 | Sig. |
|---|---|---|---|
| 1,503 | 8 | 48 | ,181 |

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+religion+region+dem99+gnp1000

Konsekvensen af brud på forudsætningen er, som ved normalitetsforudsætningen, at der opstår tvivl om *P*-værdierne, der potentielt varierer på tværs af grupperne; og problemet forværres, når der er stor forskel på antallet af observationer i de enkelte grupper (hvilket er tilfældet i eksemplet). Der er således også i denne sammenhæng grund til at være konservativ i sin fortolkning af signifikansen for de estimerede parametre og kræve lavere *P*-værdier end ellers. Yderligere skal det bemærkes, at sammenfaldet af problemer i relation til normalitets- og homoskedasticitetsforudsætningerne betyder en endnu større usikkerhed om *P*-værdierne, så hvor sådanne problemer optræder sammen, og/eller hvor stikprøven er lille, må der altså udvises særlig skærpet opmærksomhed i fortolkningen. I eksemplet er de signifikante *P*-værdier dog i forvejen så lave, at de konstaterede problemer ikke kan formodes at spille nogen større rolle.

## *7 Datasættets karakter*

Når der foretages variansanalyse er det et krav, at datasættene fra de forskellige grupper er statistisk uafhængige. Som udgangspunkt vil uafhængighedskravet være opfyldt, når der er tale om forskellige grupper i en given population, idet sandsynligheden for at udtrække en given observation fra en bestemt gruppe normalt ikke påvirkes af, at en given anden observation er blevet udtrukket fra en anden gruppe (jf. diskussionen i forbindelse med

sammenligning af to grupper i uge 5). Og er den samlede stikprøve udtrukket tilfældigt, er dette også gældende for de enkelte grupper, der indgår deri. Er data derimod fremkommet ved gentagne observationer af de samme respondenter (for eksempel i et før og efter-design) vil der være tale om statistisk afhængighed, og de særlige metoder herfor skal så benyttes (jf. Agresti og Finlay, 1997: 462-473).

## Referencer


Agresti, A. og Finlay, B. (2008) *Statistical Methods for the Social Sciences* 4. udgave, Upper Saddle River: Prentice Hall.

Fox, J. (1991) *Regression Diagnostics*, Thousand Oaks, CA: Sage Publications.

Fox, J. og Monette, G. (1992) 'Generalized Collinearity Diagnostics', *Journal of the American Statistical Association*, 87 (417), 178-183.

Lolle, H. og Klemmensen, R. (2010) 'Multivariat analyse', pp. 364-392 i Andersen L.B., Hansen, K.A. og Klemmensen, R. *Metoder i statskundskab*, København: Hans Reitzels Forlag.

King, G., Keohane, R.O. og Verba, S. (1994) *Designing Social Inquiry*. Princeton: Princeton University Press.

Studenmund, A.H. (2001). *Using Econometrics. A Practical Guide*. 4. udgave, Boston: Addison Wesley.

Sønderskov, K.M. (2011) *Fortolkning, illustration mm. af interaktion i lineære regressionsmodeller ved hjælp af MS Excel og SPSS*, 2. udgave. Note. Aarhus: Institut for Statskundskab, Aarhus Universitet.

Thomsen, S.R. (1997) *Signifikanstest i ikke-stikprøvesituationer*. Note. Aarhus: Institut for Statskundskab, Aarhus Universitet.

ISBN 87-91610-05-2