Multipel linær regression i samfundsfag A

Guide til undervisningsforløb på 4x90 minutter
med data fra Eurobarometer

Produceret af DEO, november 2019

Pædagogisk konsulent: Christian Aarøe, Favrskov Gymnasium

Introduktion

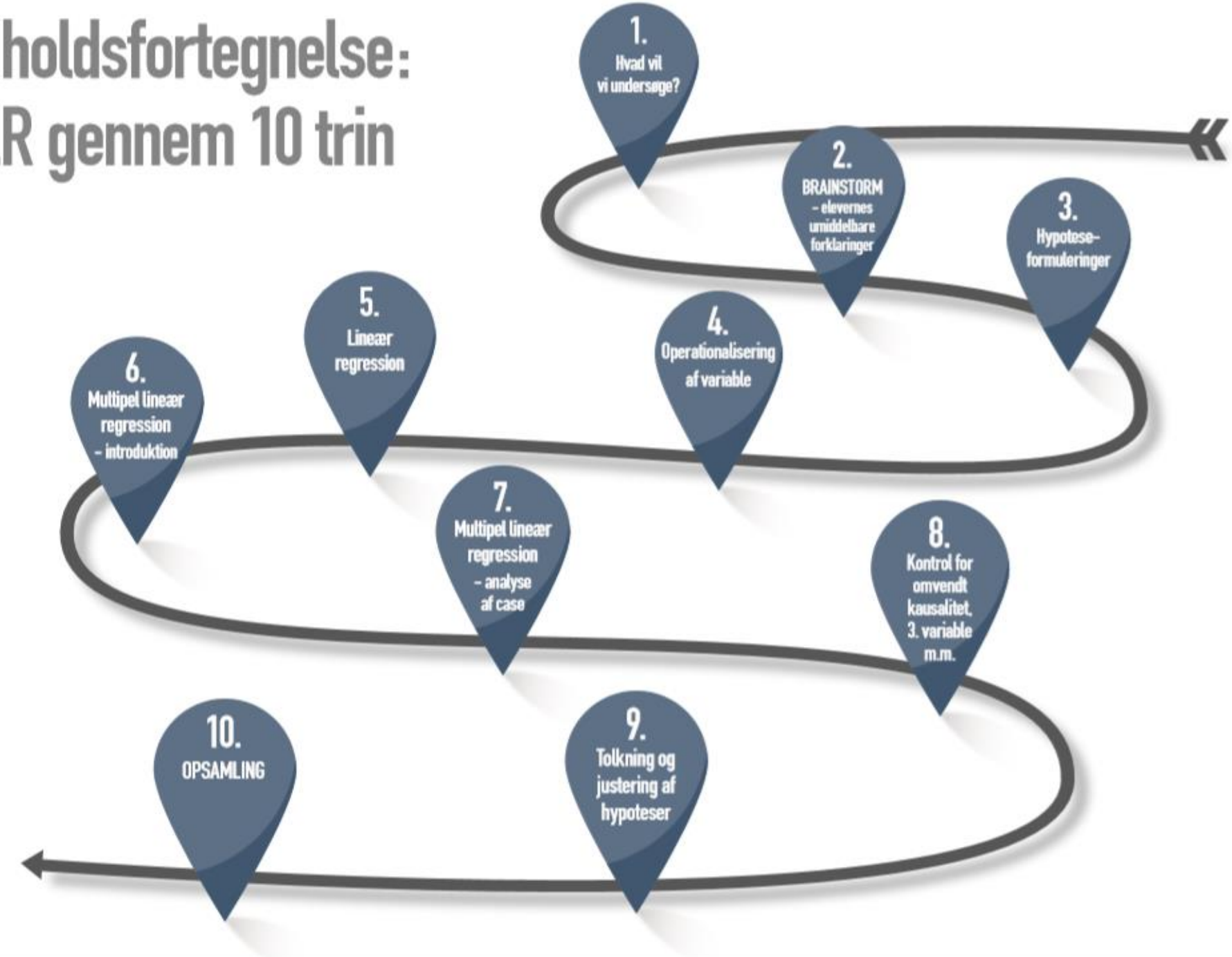
Projektets første moduler fokuserede på, hvordan multipel regression kan bruges i matematikfaget og hvilke statistiske faldgruber man skal være særligt opmærksom på. I dette modul vil vi præsentere en guide for, hvordan eleverne kan gennemføre et kort forløb om multipel regression i samfundsfaget.

En af fordelene ved at bruge multipel regression i samfundsfag er at det ansporer eleverne til at arbejde eksplorativt med en tilgang med fokus på at komme igennem væsentlige metodiske overvejelser forbundet med at opstille et ordentligt undersøgelsesdesign. Et tilgang, hvor de gradvist finder frem til de uafhængige variable med den største forklaringskraft. I nedenstående guide bruger vi primært data fra Eurobarometer som case. Man kan vælge at lade eleverne gentage den beskrevne analyse eller bruge det vedhæftede datasæt til at opstille deres egne modeller. Guiden kan naturligvis også bruges som forlæg til analyser med helt andre datasæt og tematikker.

Denne guide henvender sig til undervisere, men kan downloades og omskrives til en elevopgave. Det vil tage cirka fire blokke, 4x90 minutter, at gennemføre analysen af Eurobarometers data men metoden er også velegnet til en SOP eller SRP-opgave.

Har du spørgsmål i forbindelse med guiden, skal du være velkommen til at skrive til undervisning@deo.dk

Indholdsfortegnelse: MLR gennem 10 trin



1. Hvad vil vi undersøge?

En multipel regressionsanalyse begynder altid med et spørgsmål eller en undren. I dette tilfælde er vores case:

Hvad forklarer EU-opbakningen i medlemslandene?

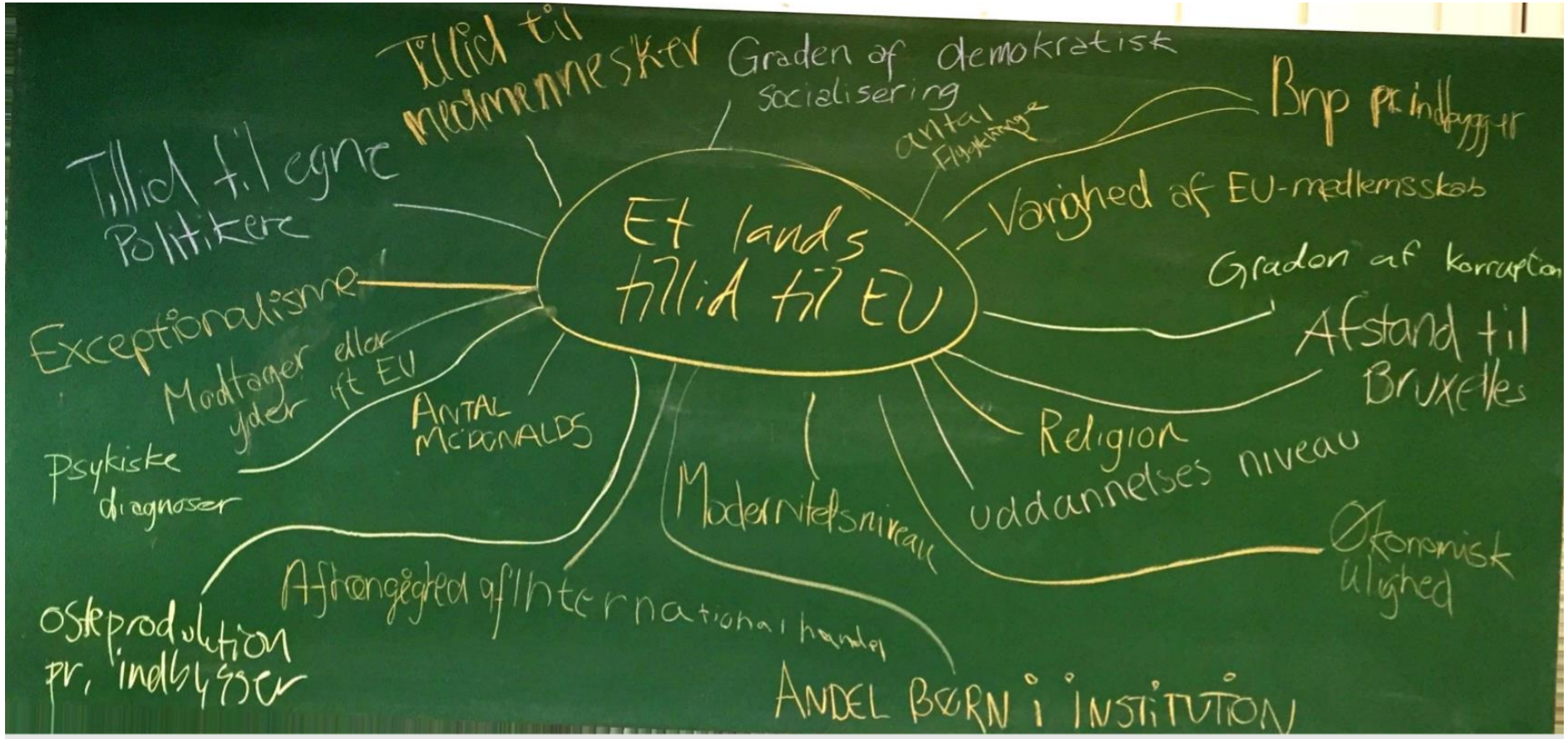
Det første trin er det vigtigste ud fra et idémæssigt og kreativt perspektiv. Der findes ingen nemme svar på dette komplekse spørgsmål, og eleverne skal derfor anvende deres logiske sans, baggrundsviden og kendskab til samfundsfaglige teorier for at kunne opstille plausible hypoteser. Her begynder vi med en tabel fra Eurobarometer (dataindsamling i foråret 2019), der viser andelen af EU-landenes borgere, der mener at 'EU-medlemsskabet er godt for vores land'.

Topscoreren er Luxembourg med 72% og landet, hvor færrest mener at EU-medlemsskabet er er en god ting er Tjekkiet med 33%. Umiddelbart er det svært at gennemskue sammenhænge, men med udgangspunkt i elevernes eksisterende viden og kendskab til samfundsfaglige teorier, vil det være muligt at udpege en række variable, der formodes at kunne forklare graden af EU-opbakning.

Præsenter tabellen for eleverne og giv dem 10 minutter til at brainstorme i grupper. Hvilke variable kan forklare det nationale niveau af EU-opbakning?

Land	Andel der mener at "EU-medlemsskabet er godt for vores land"
BEL	72
BUL	53
TJE	33
DK	76
TYSK	76
EST	74
IRL	83
GRÆ	46
SPA	69
FRA	54
KRO	40
ITA	36
CYP	52
LET	54
LIT	71
LUX	86
UNG	61
MAL	67
NED	84
ØST	46
POL	68
POR	69
RUM	57
SLOVEN	61
SLOVAK	51
FIN	66
SVE	79
EU27 – gns.	62,3703704

2. Brainstorm:
Eleverne skriver deres umiddelbare forklaringer på tavlen



2. Brainstorm

Elevernes brainstorm kunne resultere i følgende forslag:

En befolknings EU-opbakning kan forklares med:

- Tillid til egne politikere
- Tillid til medmennesker generelt (putnam)
- Religion (katolicisme, protestantisme). Weber og rationalitetens jernbur
- Økonomisk ulighed (gini)
- Uddannelsesniveau (bordieu, kulturel kapital)
- BNP per indbygger
- Modernitetsniveau: Religion (Durkheim), traditioner – nationalkonservativt - antiglobalister (Giddens), Holdninger til homoseksualitet, kvinder etc.
- Varighed af EU-medlemskab
- Graden af politisk socialisering (Almond & Verba)
- Exceptionalisme: Vores land er større og vigtigere end EU (nationalismeteorier, socialkonstruktivisme, diskursanalyse)
- Afstand til Bruxelles i km
- Afhængighed af international samhandel
- Antal sprog per indbygger (sociologisk liberalisme, transaktionsteori, kontaktteori)
- Er landet nettomodtager eller bidragyder ift. EU?
- Graden af korruption (Almond & Verba om politisk kultur, demokratisering)
- Antal flygtninge (nationalismeteorier)
- Andel børn i institution (senmoderne samfund, socialisering)

3. Hypoteseformulering

Nu bevæger vi os fra brainstorm til hypoteseformulering. Her kan det være en god idé at dele eleverne ind i par eller grupper, som hver får tildelt 3-4 påstande fra brainstormen.
Og her kommer en helt central intellektuel øvelse: De løse påstande skal omformuleres til skarpe hypoteser, der legitimeres af en forklaring og samfundsfaglige teorier. Giv eleverne god tid til at debattere, researche og skrive dette. Skemaet herunder et et eksempel, der kan bruges som forlæg. Opret evt. en fælles excel-fil, så eleverne kan se og give feedback til hinandens hypoteser og faglige begrundelser. Til sidst kan excel-filen og alle hypoteserne gennemgås og 'stres-testes' på klassen.

Afhængig variabel: Andel borgere der mener, at 'EU medlemsskabet er godt for vores land'

Uafhængig variabel: Tillid til nationale parlamenter

Hypotese: **Indbyggere med en høj grad af tillid til egne politikere, vil også have en høj grad af tillid til EU fordi der vil forekomme en spill-over effekt fra det nationale til det internationale niveau.**

Forklaring: Tillid, også kendt som social kapital, opbygges over mange generationer i samfund, hvor borgerne har vænnet sig til, at myndighederne tjener dem i et system med lav korruption. Tilliden er generel og kan forventes at have en spill-over effekt i forhold til internationale politiske systemer, der ligeledes anses for at være troværdige og bidrage til landets velstand. Omvendt vil en national mistillid til politikere også give sig udslag i en lige så lav grad af tillid til at EU har en positiv indvirkning på landets udvikling.

Faglig begrundelse/teori: Robert D. Putnam har beskrevet, hvorledes tilliden til demokratisk valgte politikere svigter, når samfundet plages af korruption. National tillid vil typisk have en spill-over effekt internationalt.

4. Operationalisering

Eleverne skal nu operationalisere deres hypoteser. Operationalisering er processen omkring at oversætte, hvad man vil undersøge i verden til målbare enheder i ens undersøgelse. Der er f.eks. ikke én bestemt måde, man kan måle tillid på. Der er mange forskellige måder at gøre dette på, og derfor er operationaliseringen ofte et punkt i ens design, der kan møde kritik. Måler vi det vi ønsker at måle?
Et andet eksempel kunne være et lands grad af modernitet som forklarende variabel, hvor hypotesen lyder at mere moderne befolkninger har en tendens til at anerkende værdien af internationalt samarbejde og handel. Hvordan undersøger man, hvor moderne et land er? Her kan vi bruge Giddens forståelse af senmodernitet og hans begreb om udlejring af sociale relationer til at se på, hvor stor en andel af de forskellige landes børn, der passes i vuggestuer og børnehaver og befolkningens accept af minoriteter. Hvis flere variable bør indgå i operationaliseringen af eksempelvis 'modernitet', kan man overveje at lave et indeks der bruges som en samlet forklarende variabel.
Når operationaliseringen er på plads, kan eleverne åbne et excel-ark, hvor de udfylder den første kolonner med den afhængige variabel og herefter kolonner med de forklarende variable. Dataindsamlingen kan hjælpes på vej med links til officielle bureauer som Danmarks Statistik, Eurobarometer og Eurostat.

Operationalisering af den afhængige variabel

Kausalmodellens afhængige variabel er allerede operationaliseret: Vi anvender Eurobarometers data, der beskriver, hvor stor en andel af de nationale befolkninger, der mener at EU-medlemsskabet er godt for deres land (se 2019-undersøgelsen - i data annex).

Operationalisering af den uafhængige variabel

Vores første hypotese er at graden af national tillid kan forklare opbakningen til EU. Den samme Eurobarometer-undersøgelse indeholder en tabel over borgernes generelle tillid til det nationale parlament. De øvrige uafhængige variable har vi fundet data på i Eurostat.

5. Lineær regression

Nu har eleverne opstillet hypoteser, legitimeret dem fagligt og operationaliseret dem. Alt er klart til regressionsanalyserne i excel.
Hypoteserne er stadig blot påstande, som nu skal bevises eller forkastes på baggrund af en dataanalyse. Før vi kaster os ud i multipel lineær regression, skal eleverne undersøge om der er samvarians mellem den afhængige og de uafhængige variable. Og hvis der er samvarians: Er sammenhængen statistisk signifikant og hvor meget af variansen kan forklares af de individuelle x-variable?

Instrukser til eleverne

1. Se videoen 'lineær regression med excel'. Eleverne bør være opmærksomme på at de sætter kryds ud for r2.

2. Gennemfør den lineære regression med den afhængige variabel og hver af de uafhængige variable.

Se eksempel på DEO's analyser af 26 uafhængige variable i excel-filen **EU er godt for mit land – LR og MR**.

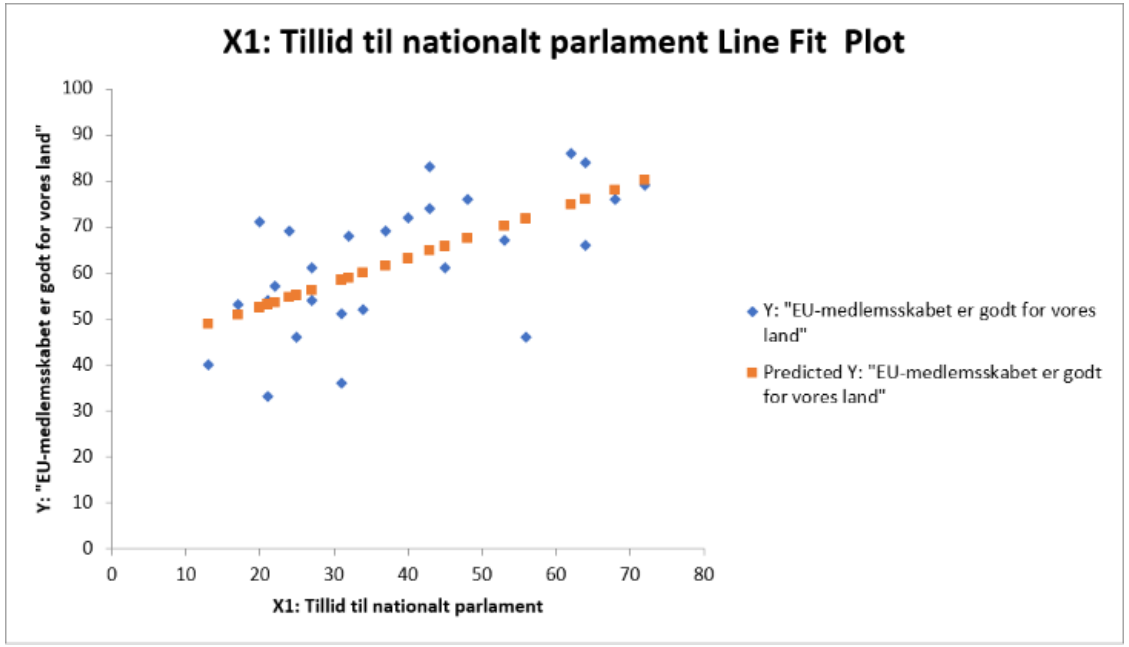
3. Tolkning på resultater: Scatterplots, P-værdier, forklaringskraft (R2). Hvilke variable går vi videre med og hvilke bør vi tage ud?

4. Overvej, hvorvidt der kunne være relevante 3.variable, der i virkeligheden kunne forklare variationen i henholdsvis Y. Når vi ser at 'national tillid' korellerer positivt, er det så i virkeligheden 'generel tillid' vi måler på? Og kunne der være tale om omvendt kausalitet, hvor Y i virkeligheden forklarer X? Det kan man undersøge ved at køre den lineære regression, hvor X og Y byttes rundt.

5. Præsentation af den lineære regression for klassen.

OBS: Dette punkt lever op til fagets nye krav om, at eleverne ikke længere blot skal kunne fortolke en lineær regression, men at de også skal kunne skabe den.

5. Lineær regression:
Sammenhæng mellem tillid til nationalt parlament og EU-opbakningen



Tolkning: Analysen viser at der er en lineær samvarians, hvor den uafhængige variabel forklarer 39% af variationen i den afhængige variabel. Modellen er statistisk signifikant da p-værdien holder sig indenfor 5% usikkerhed (0,0005).
Man kan dog ikke ud fra dette konkludere noget omkring kausalitet, da tilliden til det nationale parlament kan være udtryk for en bagvedliggende variabel.

Vores eksempel: Regressionsanalyser af 26 variable

Hensigten er naturligvis, at eleverne selv vælger en relevant problemstilling og selv finder uafhængige variable, hvorefter de kører analyserne. I forbindelse med dette projekt har vi foretaget en række analyser til inspiration for læreren. Vores case fokuserer på EU-opbakningen og i den forbindelse har vi udvalgt 26 uafhængige variable.

I excel-filen (se oversigten for modul 3) fremgår det, at den lineære regression viser, at 17 ud af de 26 variable har en statistisk signifikant sammenhæng med vores afhængige variabel. Forklaringskraften varierer fra 14-45%.

Umiddelbart tyder det på, at der er en sammenhæng mellem EU-opbakningen og:

- en stærk økonomi (X2, X6, X9, X20 og X23)
- tillid (X1, X19, X21)
- demokratiets tilstand (X3, X23)
- antal asylansøgere (X24)
- livsglæde (X26)
- modernitet/progressivitet (X15, X16, X17, X18)

Kort tolkning af lineær regression

En stærk økonomi øger EU-opbakningen
Fem variable indikerer, at en sund økonomi med høj beskæftigelse øger opbakningen til EU. Denne tendens bekræftes af forskningen på området, der viser at den makroøkonomiske situation påvirker borgernes perception af deres politikere.

Tillid til medborgere skaber tillid til EU
Tre variable indikerer, at der en sammenhæng mellem en høj grad af tillid til egne politikere, og til fremmede generelt, og EU-opbakningen. EU's maskinrum er fysisk og mentalt langt væk for mange mennesker og opbakningen til et relativt abstrakt og fjernt politisk system er afhængigt af om man stoler på, at politikerne vil forvalte deres magt samvittighedsfuldt og til fordel for alle europæere.

Demokratiets tilstand
Der er en positiv sammenhæng mellem variablene X3 og X23 og EU-opbakningen. Velfungerende nationale demokratier kan forventes at socialisere indbyggerne til at vurdere overnationale demokratier i et positivt lys.

Antal asylansøgere
X16 og X24 indikerer at man i lande med mange asylansøgere har en tendens til at tænke, at EU-medlemsskabet ikke har en positiv indflydelse på ens land. En tolkning af dette kunne være at mange borgere ser Schengen-samarbejdet og EU's svage ydre grænser som en direkte åsag til den øgede tilstrømning af asylansøgere gennem det seneste årti. Krisen har mange årsager, men EU er en oplagt syndebuk.

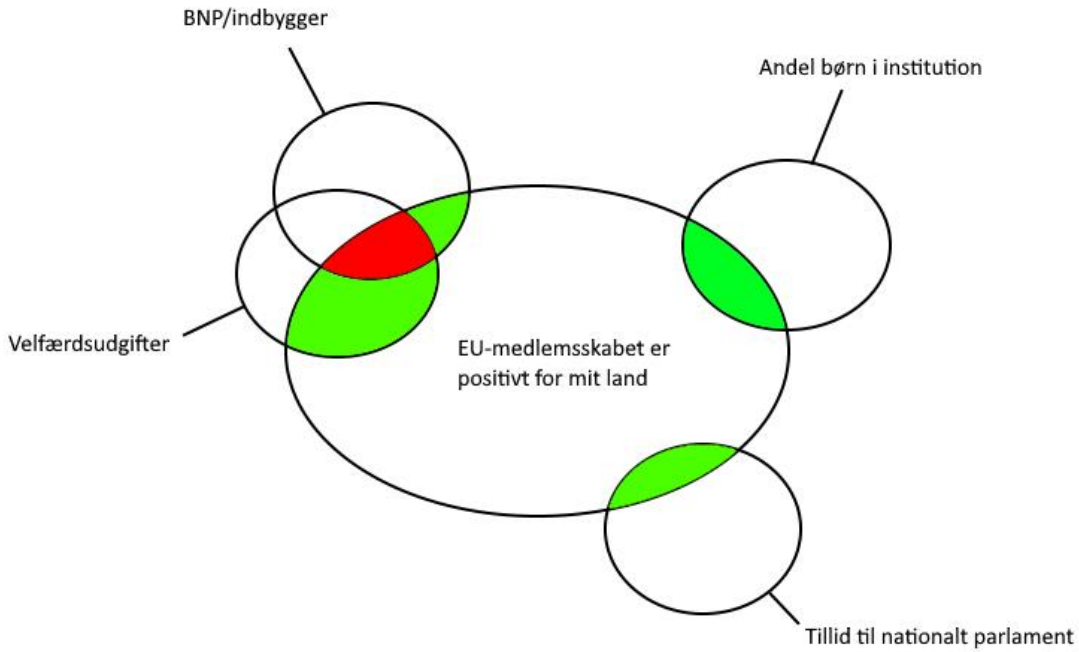
Velfærdsydelser styrker EU-opbakningen
Analyserne med X7 og X22 viser, at der i velfærdsstater er en relativt større andel af borgere der mener, at EU-medlemsskabet er positivt for deres land. En mulig tolkning er at en øget interaktion med myndighedspersoner gennem institutioner osv skaber et mere tillidsfuldt og mindre formelt forhold mellem stat og borgere. Borgerne føler ydermere at deres skattekroner bruges direkte på borgerne og forbedrer deres livsvilkår. Denne forventning projiceres muligvis til EU.

6. Multipel lineær regression: Introduktion

Nu er vi klar til at kaste os ud i multipel lineær regression. Hvor lineær regression kun har én forklarende variabel, inddrager multipel regression flere variable i den samme model. Derved kan vi øge forklaringskraften og opnå en mere dækkende forståelse for, hvad der ligger bag EU-opbakningen. En statistisk forsvarlig multipel regression er dog ikke helt simpel at gennemføre. Vi får fx ingen tendenslinje i et koordinatsystem og vi kan ikke aflæse placeringen af datapunkterne og herudfra tolke noget omkring sammenhængen.

Et andet opmærksomhedspunkt er risikoen for multikollinaritet. To uafhængige variable der korrelerer indbyrdes kan næppe bruges i en meningsfuld multipel regression, da de begge er udtryk for det samme bagvedliggende fænomen. Dette kan illustreres grafisk som her, hvor det røde område repræsenterer multikollineariteten og vil resultere i p-værdier der er højere end de fastlagte 5%.

To variable der er indbyrdes uafhængige, som børn i institution og tillid til nationalt parlament, kan derimod begge inddrages i en multipel regression med en højere forklaringskraft end de individuelle variable.



6. Multipel lineær regression: Analyse i excel

1) Eleverne skal først tilføje funktionen *dataanalyse* i excel. Find en installationsguide til både Mac og PC her.

2) Eleverne skal forberedes på, hvordan man rent teknisk gennemfører MLR-analysen i excel. Dette forklares i videoen, der hører til dette modul.

3) Gennemgå de vigtigste indikatorer i *resumeoutput/summary output* (som i videoen) :

- Koefficienterne for de enkelte forklarende variable (positiv/negativ relation)
- Modellens samlede p-værdi (skal være under 5%)
- De enkelte variables p-værdier (skal være under 5%)
- Justeret R2-værdi: Hvor meget af variationen i den afhængige variable kan forklares af de valgte uafhængige variable?
- Nu er eleverne klar til at teste deres hypoteser i en MLR-analyse.

7. Multipel regression – caseanalyse (X1 og X22)

- Vi opstiller hypotesen: Borgere i lande, hvor der både er en høj tillid til nationale politikere og en høj andel af børn i institution, vil udvikle et positivt syn på 'det offentlige' som en integreret del af 'det gode liv'. Disse borgere vil derfor have en tendens til at vurdere EU i et positivt lys, hvorimod borgere i lande med høj korruption og ringe velfærd vil se politiske institutioner som dysfunktionelle.
- Vi tester begge variable og får følgende tal ud af analysen i excel (se resumeoutput herunder).
- Modellen er statistisk signifikant, da F-værdien er 0,00024
- P-værdierne for de enkelte uafhængige variable holder sig indenfor 5% usikkerhed (3% og 5%)
- Vores justerede R2 lyder på 46%, hvilket er en moderat til stærk forklaringskraft.
- Vi kan nu rimeligvis argumentere for, at vores hypotese har hold i virkeligheden, men før vi kan opstille en kausalmodel må vi foretage yderligere analyser med flere variable. For kan der være andre årsager til, at lige netop disse to variable skaber en god MLR-analyse i kombination?

RESUMEOUTPUT

Regressionsstatistik	
Multipel R	0,707747
R-kvadreret	0,500905
Justeret R-kvadreret	0,459314
Standardfejl	10,83789
Observationer	27

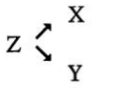
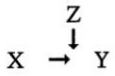
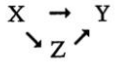
ANAVA

	fg	*SK*	*MK*	*F*	*Signifikans F*
Regression	2	2829,261	1414,63	12,04353	0,000239
Residual	24	2819,035	117,4598		
I alt	26	5648,296			

	Koefficiente	*Standardfej*	*t-stat*	*P-værdi*	*Nedre 95%*	*Øvre 95%*	*Nedre 95,0%*	*Øvre 95,0%*
Skæring	39,37634	5,300172	7,429258	1,14E-07	28,43733	50,31536	28,43733	50,31536
X22: Andel børn under 3 år i insti	0,33078	0,140796	2,34936	0,02737	0,040192	0,621368	0,040192	0,621368
X1: Tillid til nationalt parlament	0,321014	0,152596	2,103684	0,046069	0,006071	0,635956	0,006071	0,635956

8. Kontrol for omvendt kausalitet, 3. variable m.m.

Multikollinearitet er kun en ud af flere relationer, der ikke kan indgå i en multipel regressionsanalyse, og det er værd at overveje, hvilke relationer der er skyld i høje p-værdier, når man ellers benytter to variable, der hver for sig havde en statistisk signifikant forklaringskraft. Variable der i sig selv har en signifikant og positiv forklaringskraft kan skabe dårlige resultater i kombination. Se eksempler på dette i grafikken til højre.

Graf	Relations navn	Hvad sker der ved kontrol for tredjevariable?
	Spuriøs relation mellem X og Y	Relationen mellem X og Y forsvinder ved kontrol for Z
	Kæderelation	Forholdet mellem X og Y forsvinder ved inddragelse af Z
	Interaktionsrelation	Størrelsen af X1's effekt på Y varierer med størrelsen af Z
	Multipel relation	Forhold mellem X og Y forsvinder ikke og ændres ikke ved inddragelse af Z
	Direkte og indirekte relation	Forholdet mellem X og Y ændres, men forsvinder ikke ved inddragelse af Z

Kilde: Metoder i Statskundskab af Lotte Bøgh Andersen, Kasper Møller Hansen og Robert Klemmensen (red.), 2010.

9. Tolkning og justering af hypoteser

- Når analyserne er blevet gennemført, er det tid til at genbesøge hypoteserne. Hvilke kan vi blankt afvise og – ikke mindst – hvorfor?
- Hvilke kan vi statistisk set bekræfte? Og er sammenhængen så stærk at vi kan begynde at gøre os tanker om kausalrelationen? Er det virkelig X1 og X22, der påvirker Y direkte, eller er der sandsynligvis en bagvedliggende eller mellemkommende variabel?
- Kan vi justere på nogle hypoteser og forsøge en ny MLR med andre variable?
- Vær opmærksom på risikoen for 'overfitting'; At modellen får så mange variable, at man ikke kan gennemføre en meningsfuld analyse. En kompleks MLR-analyse med så relativt få observationer som her bør ikke indeholde mere end 3-4 uafhængige variable.

10. Afrunding: Grav dybere med MLR

Vi håber at denne guide har givet dig mod på at gennemføre et MLR-forløb med dine elever i Samfundsfag A. Dette projekt er tænkt som et tværfagligt forløb med Mat-A/Samf-A, så det er oplagt at indlede et samarbejde med matematiklærere (hvilket også er intentionen med den seneste gymnasiereform). Afslutningsvist vil vi fremhæve, hvad vi mener er merværdien ved at introducere MLR for eleverne:

- En (monokausal) lineær regression med en høj R2 er ikke nødvendigvis tilfredsstillende for den ambitiøse samfundsanalytiker. Den sociale virkelighed er næsten altid mere kompleks.
- MLR giver bedre mulighed for at formulere og 'bygge' dækkende modeller.
- MLR er intellektuelt udfordrende, da mange variable skal tænkes sammen. Eleverne tvinges til at overveje alternative kausalrelationer og være metodisk påpasselige.

- Feedback og kommentarer: Skriv til undervisning@deo.dk