2.4 Hvad er lineær regression, og hvordan udføres den?

Meget ofte i samfundsfag optræder der data, hvor du kan have en formodning om en sammenhæng. Det kan være indkomst og forbrug, hvor forbruget antages at afhænge af indkomsten. Forbruget er afhængig variabel, mens indkomsten er uafhængig variabel (om afhængige og uafhængige variabel, se s. 27). Fra mikroøkonomien er det også velkendt, at den efterspurgte mængde afhænger af prisen. Disse sammenhænge kan i mange tilfælde mere præcist beskrives ved hjælp af formlen for en ret linje. For eksempel kan sammenhængen mellem den efterspurgte mængde (E) og prisen (p) beskrives således:

$$E = -ap + b$$

Den efterspurgte mængde antages at falde, jo højere prisen er. Derfor det negative fortegn på linjens hældning (a).

I tabel 2.7 nedenfor er vist ulighed og grad af tillid for en række lande.

Tabel 2.7. Ulighed og tillid i udvalgte lande. 2010

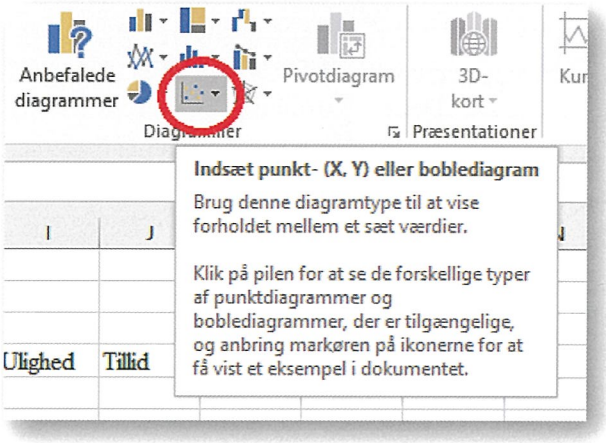
Land	Ulighed	Tillid	Land	Ulighed	Tillid
Australien	7,00	39,90	**Japan**	3,40	43,10
Belgien	4,60	30,70	**Holland**	5,30	59,80
Canada	5,63	38,80	**New Zealand**	6,80	49,10
Danmark	4,30	66,50	**Norge**	3,85	65,30
Finland	3,72	58,00	**Portugal**	8,00	10,00
Frankrig	5,60	22,20	**Singapore**	9,70	16,90
Tyskland	5,20	34,80	**Spanien**	5,55	36,20
Grækenland	6,20	23,70	**Sverige**	3,95	66,30
Irland	6,05	35,20	**Schweitz**	5,73	41,00
Israel	6,78	23,50	**Storbritannien**	7,17	29,80
Italien	6,65	32,60	**USA**	8,55	35,80

Note: Graden af ulighed måles som hvor meget den rigeste femtedel tjener i forhold til den fattigste femtedel. Tillid måles som den andel af befolkningen, der erklærer sig enige i udsagnet: Generelt har jeg tillid til mine medmennesker. Graden af tillid kan variere fra 0 til 100.

Lineær regression: Beregning til at finde den lineære sammenhæng mellem to variable.

For at finde ud af, om der er en statistisk sammenhæng, er vi nødt til at undersøge det nærmere ved hjælp af **lineær regression**. Allerførst konstrueres et punktdiagram (xy) i regneark. I Excel mærkes tabellen af, og i menuen vælges Indsæt ➞ Diagram. Klik på ikonet med den røde cirkel omkring, som vist i figur 2.14.

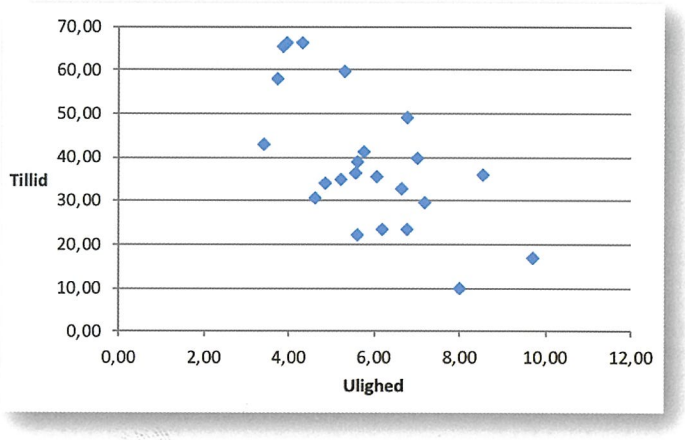
Figur 2.14. Valg af diagramtype i Excel



Du får nu et diagram som vist i figur 2.15. Umiddelbart kan du konstatere, at der er en samvariation (også kaldet korrelation, se s. 28). Jo større ulighed, jo lavere grad af tillid. Men hvad er samvariationen, og hvor stærk er den? For at besvare dette skal der konstrueres en **tendenslinje** (bedste rette linje).

I Excel klikkes på et punkterne, og punkterne forvandles til 'blomster'. Højreklik og vælg Tilføj Tendenslinje. Hak af i Vis ligning og Vis R-kvadreret. Du skal nu have et diagram som vist i figur 2.16 næste side.

Figur 2.15. Ulighed og tillid. Udvalgte lande. 2010



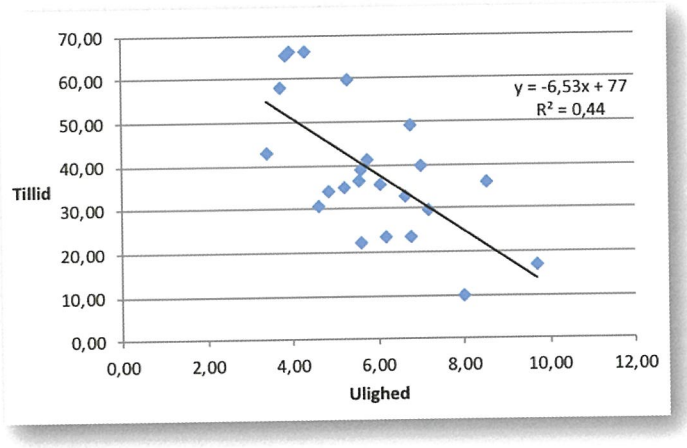
For det første kan du se, at linjen har en negativ hældning – den falder fra venstre mod højre. Dette bekræftes af fortegnet på hældningen (-6,53). For det andet betyder fortegnet, at hver gang uligheden (x-vædien) vokser med 1, vil tilliden (y-værdien) falde med 6,53. For det tredje udtrykker R^2, hvor stærk samvariationen er. R^2 benævnes **forklaringsgraden** og udtrykker i et eneste tal, hvor stor en del af variationen i y der kan forklares ved hjælp af x. I dette tilfælde kan 44 % af variationen i tillid forklares ved hjælp af ulighed. Med andre ord må vi lede efter andre faktorer, der kan bidrage til at forklare resttallet på 56 %. Det kunne for eksempel være, hvor korrupt landet er, hvor udbredt kriminalitet er, eller ... For det fjerde skal vi være opmærksomme på eventuelle **outliers**, altså punkter, som af en eller anden speciel grund stikker ud og kan bidrage til at mindske forklaringsgraden. I figur 2.17 på s. 54 er der sat landenavne (ikke alle er med) på punkterne, og lande med selektiv velfærdsmodel (Tyskland, Frankrig, Portugal) scorer lavere på tillid, end modellen tilsiger, mens lande med den universelle model har lav ulighed og høj grad af tillid.

Tendenslinje: Den rette linje som ligger tættest på punkterne i et diagram med to variable

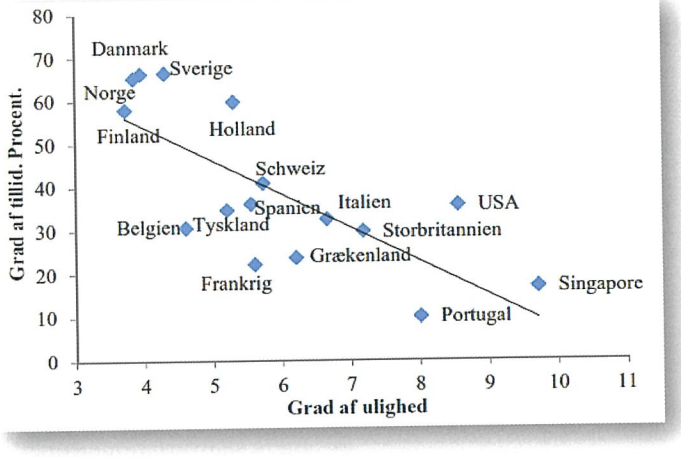
Forklaringsgrad: Den del af variationen i den ene variabel (den afhængige) som kan forklares ved variationen i den anden variabel (den uafhængige).

Outliers: En værdi for en observation som ligger længere væk fra tendenslinjen og dermed gør forklaringsgraden mindre.

Figur 2.16. Ulighed og tillid. Udvalgte lande. 2010



Figur 2.17. Ulighed og tillid. Udvalgte lande. 2010



For at vende tilbage til udgangspunktet: Er der en sammenhæng mellem graden af ulighed og graden af tillid? Vi kan konstatere ved at kigge på figur 2.16 og fortolke regressionsligningen, at de to variable samvarierer på et nogenlunde niveau: Når uligheden vokser, så falder tilliden. Mere firkantet: Når du kender værdien af den uafhængige variabel (gradaf ulighed) kan du bedre ved hjælp af regressionsmodellen gætte værdien af den afhængige variabel. Der er dog alene tale om en statistisk sammenhæng (korrelation) og *ikke* en årsagssammenhæng (kausalitet). Hvorvidt uligheden (uafhængig variabel) er bestemmende for graden af tillid (afhængig variabel), kan ikke afgøres på baggrund af analysen her. Eventuelt kan en sammenhæng begrundes teoretisk – for eksempel med henvisning til velfærdsmodeller eller lignende. Desuden kan der være en tredje (intervenerende) variabel på spil, som er bestemmende for både ulighed og tillid. Man kunne godt

Hvor god skal R^2 være?

Hvis R^2 er større end en forklaringsgrad på 0,75, er vi normalt godt tilfredse i samfundsfag, da der også meget tit er knyttet en eller anden form for stikprøveusikkerhed til variablene.

Hvis R^2 ligger mellem 0,25 og 0,75, er der en nogenlunde samvariation, men vi skal på jagt efter andre faktorer, der kan bidrage til at forklare variationen i den afhængige variabel.

Hvis R^2 er under 0,25, vil vi normalt være skeptiske og overveje, om der kan findes andre variable, der bedre kan forklare variationen i den afhængige variabel.

tilbage til ud-
t: Er der en
g mellem gra-
d og graden af
konstatere ved
gur 2.16 og for-
ionsligningen,
ble samvarierer
ande niveau:
'n vokser, så
n. Mere firkan-
ender værdien
'ngige variabel
ied) kan du bed-
af regressions-
'tte værdien af
ge variabel. Der
tale om en sta-
ienhæng (kor-
ikke en årsags-
ng (kausalitet).
igheden (uaf-
abel) er bestem-
graden af tillid
'ariabel), kan
s på baggrund af
'r. Eventuelt kan
hæng begrundes
iodeller eller lig-
de) variabel på
Ian kunne godt

samfundsfag,
til variablene.

vi skal på jagt
ge variabel.

indes andre va-

Hvad er årsag til hvad? Nogle gange er det oplagt, hvad der er årsag og virkning, men andre gange kræver det grundige overvejelser – og måske kan årsagssammenhængen (kausaliteten) gå begge veje.

have en hypotese om, at korruption er bestemmende for både ulighed og mistillid, altså:

Voksende korruption ➞ voksende ulighed og

Voksende korruption ➞ faldende tillid

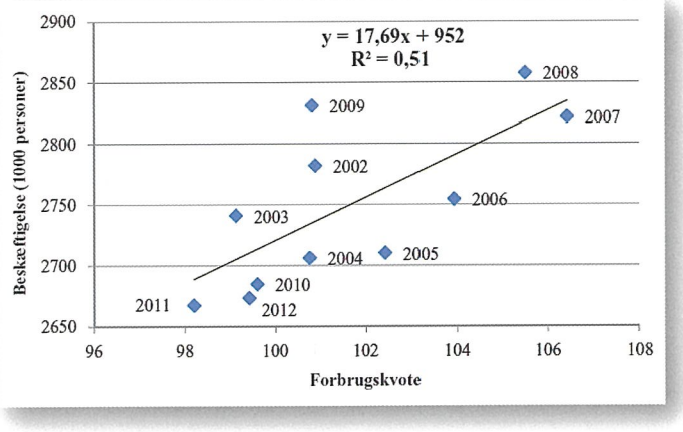
I kapitel 1 så vi, at den afhængige variabel normalt betegnes y, og den uafhængige variabel betegnes x. Normalt angives den afhængige variabel også på y-aksen og den uafhængige variabel på x-aksen i et koordinatsystem. Dette er helt som i matematik, hvor y afhænger af x. I nogle tilfælde kan der dog argumenteres for, at hvad der er afhængig og hvad der er uafhængig variabel, ikke er helt entydig. I figur 2.18 på næste side er vist en regression over sammenhængen mellem beskæftigelse og forbrugskvote.

Der kan med udgangspunkt i figur 2.18 argumenteres for både at:

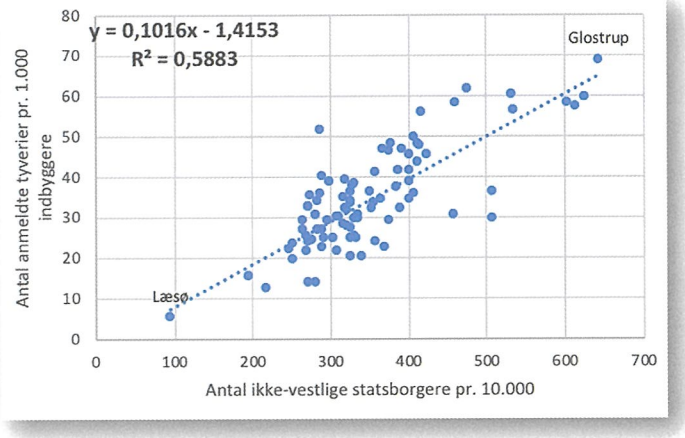
a) Forbrugskvote → Beskæftigelse, idet det må være korrekt, at når forbrugskvoten vokser, så vokser forbruget og der skal ansættes flere arbejdere i industrien til at producere de pågældende varer og

b) Beskæftigelse → Forbrug, idet en voksende beskæftigelse er udtryk for opgang i økonomien, som giver øget tillid til fremtiden, og dermed kan en større andel af indkomsten anvendes til forbrug.

Figur 2.18. Forbrugskvote og beskæftigelse



Figur 2.19. Antal ikke-vestlige statsborgere pr. 10.000 indbyggere og anmeldte tyverier pr. 1.000 indbyggere. Udvalgte kommuner. 2017.



Kilde:
www.noegletal.dk

Det er værd at bemærke, at R^2 bliver den samme, uanset hvilken af de to variable der vælges som afhængig og uafhængig variabel.

Den økologiske fejlslutning

I figur 2.19 er vist sammenhængen mellem antal 10.000 indbyggere og antal anmeldte tyverier pr. 1.000 indbyggere for udvalgte kommuner i Danmark.

Den statistiske sammenhæng er nogenlunde med en R^2 på ca. 0,59. Men alligevel skal der trædes meget varsomt omkring hvad der kan konkluderes:

Du kan ikke konkludere, at Ahmed er tyvagtig, fordi han bor i Glostrup og er ikke-vestlig statsborger. Man kan ikke slutte fra aggregerede data (antal ikke vestlige statsborgere) til et enkelt-individ (Ahmed). Du kan ej heller konkludere, at ikke-vestlige statsborgere er mere tyvagtige end etniske danskere. Det kan være – vi ved det ikke – at alle tyverier begås af danskere i de pågældende kommuner.

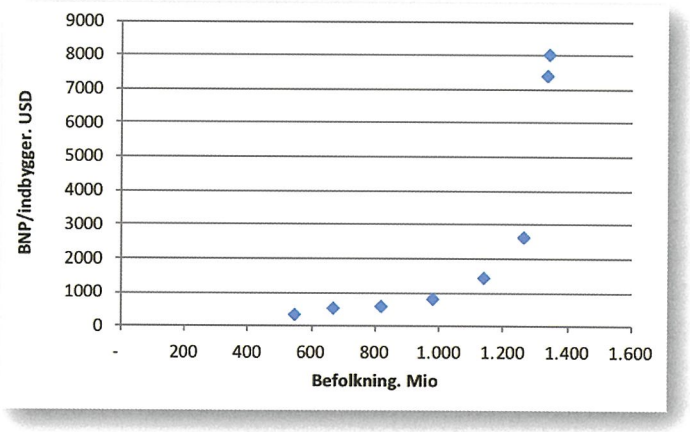
Disse to fejlslutninger: 1) at slutte fra aggregerede data til indvid-adfærd og 2) at slutte fra andele til den pågældende gruppes adfærd benævnes økologiske fejlslutninger.

Til den skriftlige prøve i samfundsfag skal du selv kunne gennemføre en lineær regression og tolke resultatet på baggrund af et givet datasæt.

Andre former for regression

Ved lineær regression findes den bedste rette linje, som minimerer afstanden mellem punkterne og den rette linje. I nogle tilfælde kan det være relevant at afprøve andre former for regression, som måske vil give en højere forklaringsgrad. Hvis du for eksempel ser på udviklingen i Kinas BNP sammen med befolkningsudviklingen (figur 2.19), kan der konstrueres nedenstående diagram. Umiddelbart kan du se, at der ikke er tale om en pæn lineær sammenhæng.

Figur 2.20. Befolkningsudvikling og BNP/indbygger i Kina. 1950-2014



En lineær regression, hvor BNP vokser i takt med befolkningen, giver en forklaringsgrad på 0,63, hvorimod en **eksponentiel regression** giver en bedre forklaringsgrad på 0,89, og den eksponentielle sammenhæng udtrykker samtidig, at BNP vokser hurtigere end befolkningen (arbejdsstyrken). Det vil sige, at produktiviteten er voksende.

Også hvis du vil finde en regression for en udvikling over tid kan det være relevant at afprøve andre former for regression end den lineære.

Eksponentiel regression: Beregning til at finde den eksponentielle sammenhæng mellem to variable.

2.5 Hvordan måles økonomisk ulighed?

I den politiske debat spiller indkomster og dermed indkomststatistik en central rolle. Hvem skal have hvad? Skal Robin Hood på banen og omfordele fra de rigeste til de fattigste? I det følgende vil de mest centrale mål for fordelingen af indkomster blive gennemgået.

I tabel 2.7 er vist udviklingen i de gennemsnitlige personindkomster i Danmark fra 1990 til 2014.

Gennemsnit: Udregnes som summen af alle observationer divideret med antal observationer.

Gennemsnit udregnes ved at tage den totale indkomst og dividere med antal personer. Eksempelvis var der i Danmark i 2017 4.910.000 skattepligtige personer, hvis totale indkomst var 1368 milliarder kroner. Gennemsnittet bliver 278.615 kroner, nemlig: