Befolkning og velferd ECON 1730, H2016

Regresjonsanalyse

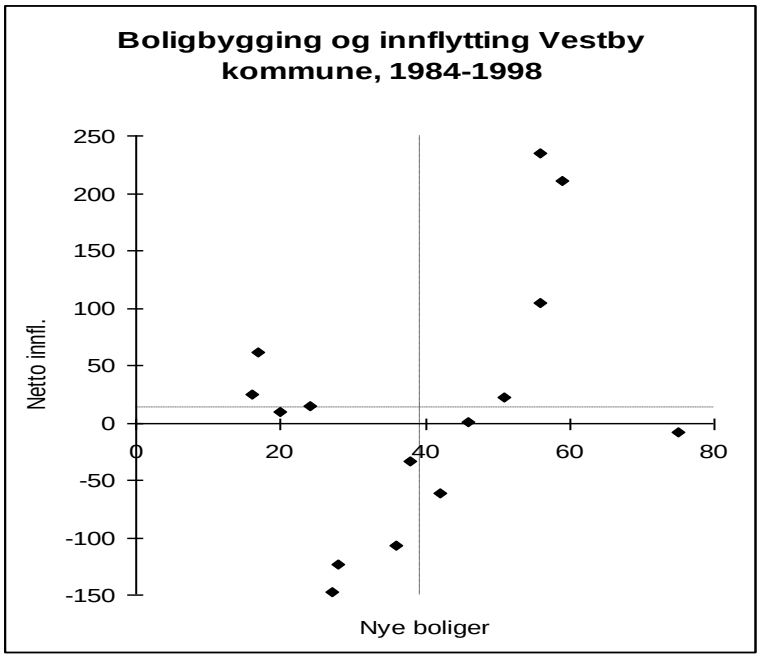
Problem: Gitt planer for 60 nye boliger i kommunen neste år, hvor mange innflyttere kan vi forvente?

Tabell

Vestby kommune

	Nye boliger	Netto innfl.
1984	56	104
1985	56	235
1986	59	211
1987	42	-61
1988	51	22
1989	28	-123
1990	27	-147
1991	75	-8
1992	36	-107
1993	38	-34
1994	20	9
1995	46	1
1996	24	14
1997	17	61
1998	16	25

Spredningsdiagram



I spredningsdiagrammet ser vi en svak sammenheng mellom boligbygging og innflytting: det er en tendens til at år med forholdsvis mange nye boliger (dvs. flere enn gjennomsnittet – den loddrette stiplede linjen) også har forholdsvis mye innflytting (over gjennomsnittet – den vannrette stiplede linjen). Også omvendt: år med få nye boliger har en tendens til å ha lav innflytting – til og med utflytting!

Men: hvor mye netto innflytting gir 60 nye boliger? Eller generelt et visst antall X nye boliger?

Regresjonsanalyse er en teknikk som finner en sammenheng mellom en variabel X og en variabel Y når vi antar at X påvirker Y: X $\longrightarrow$ Y.

Den enkleste forutsetningen for en slik sammenheng er at X og Y henger lineært sammen. Dette blir en rett linje i spredningsdiagrammet.

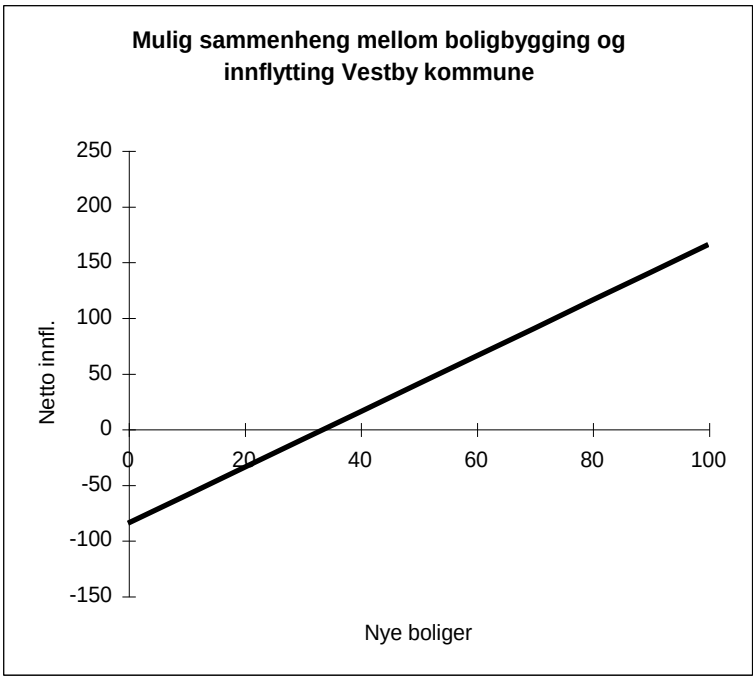
Algebraisk: $Y = a + b*X$, eller netto innflytting = $a + b*$boliger.

Variabelen X (her boligbygging) kalles for den uavhengige variabelen, variabel Y (netto innflytting) er den avhengige variabelen.

Når vi antar en slik lineær sammenheng kaller vi metoden for lineær regresjon.

Størrelsene *a* og *b* kalles for parametrene. De bestemmer hvordan den lineære sammenhengen ser ut.

a: netto innflytting når ingen boliger bygges (X = 0). Kan være negativ!
b: ekstra innflytting når 1 ekstra bolig bygges.



a kalles også for skjæringspunkt eller konstant, *b* for stigningstall eller koeffisient.

Gitt en rekke tall med observerte verdier for X og Y (f.eks. som i tabellen) finnes det statistiske metoder som beregner ("estimerer") optimale verdier for *a* og *b*.
Hvordan er ikke pensum (jf. "minstekvadratersmetode" i håndbøker).

Men viktig er tolkningen av resultatene.

I dette tilfellet regnet regresjonsfunksjon i Excel ut følgende resultater for Vestby:

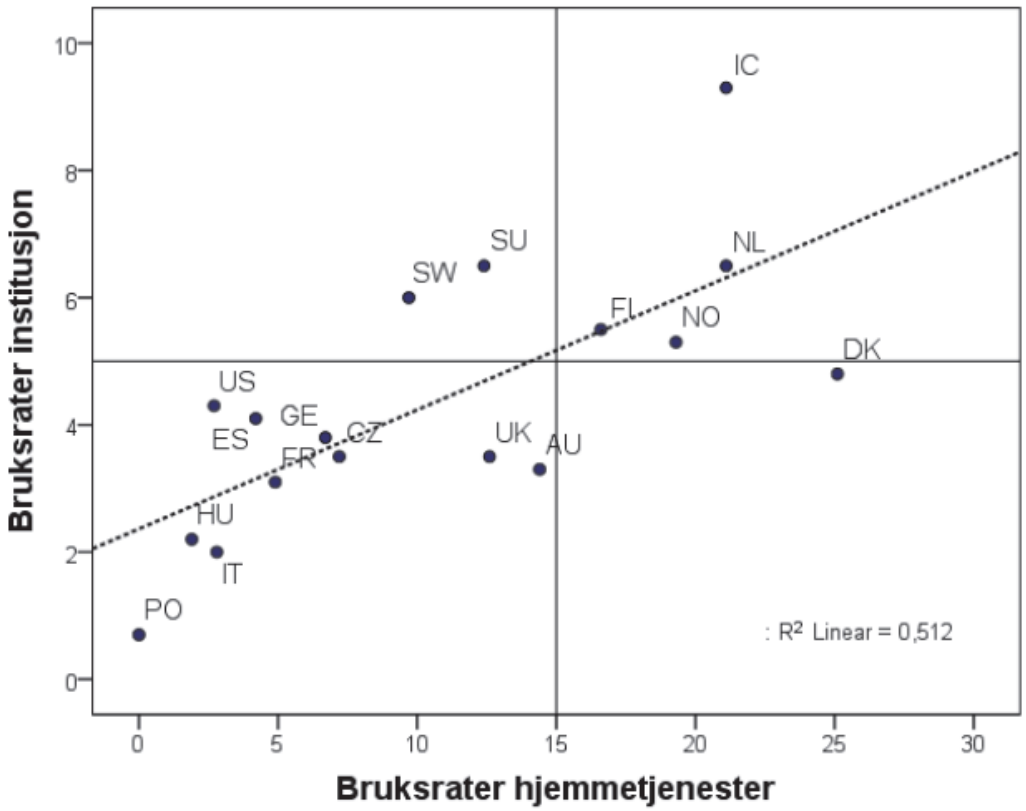
a = -85,1 og *b* = 2,5.

slik at sammenhengen blir: netto innflytting = -85,1 +2,5*boliger.

Tolkning: uten boligbygging (dvs. X = 0) blir det en netto utflytting på 85 personer (avrundet);
1 ekstra bolig medfører i snitt 2,5 personer ekstra som flytter inn i kommunen.
60 nye boliger ville medføre en netto innflytting som er lik
(-85,1 + 2,5*60) = 65 personer (avrundet).

Spørsmål: var verdien 2,5 et rimelig estimat for b?

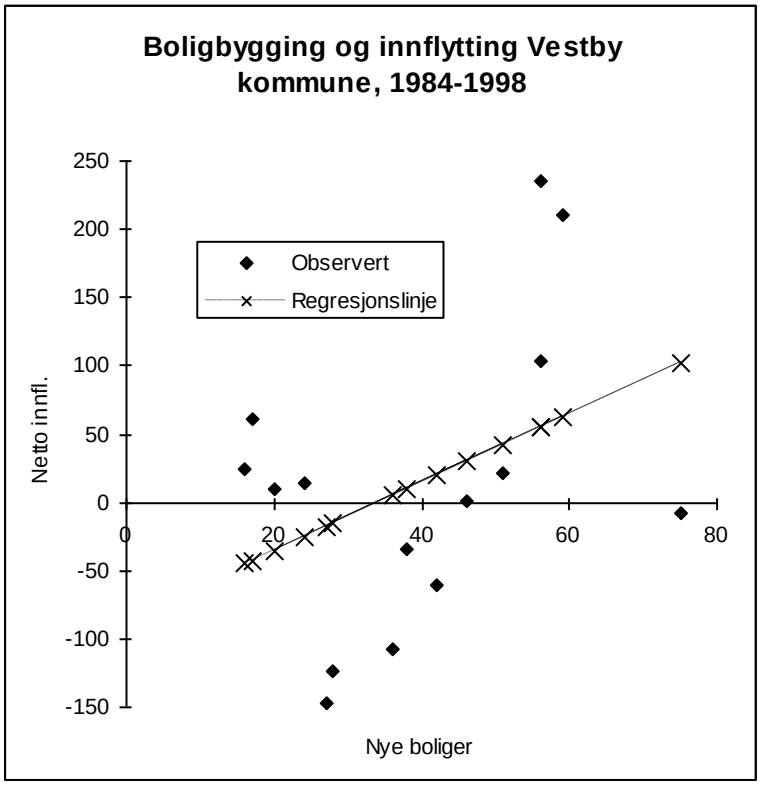
Et annet eksempel



Kilde: Huber et al. 2009. Forkortelse land: IC (Island), DK (Danmark), NL (Nederland), NO (Norge), FI (Finland), SU (Sveits), SW (Sverige), AU (Østerrike), (UK) Storbritannia, (CZ) Tsjekkia, (GE) Tyskland, (FR) Frankrike, (ES) Spania, (US) USA, (IT) Italia, (HU) Ungarn, (PO) Polen.

Figur 10.2. Tilgang til pleie- og omsorgstjenester etter land (andel 65år + med hjemmetjenester og institusjonstjenester)

Fra Daatlands pensumartikkel «Komparative perspektiver på omsorgstjenestene – Norge i en internasjonal sammenheng»



Statistisk usikkerhet

Mange av de observerte datapunkter ligger på stor avstand fra den estimerte linjen. Sagt på en annen måte: den estimerte sammenhengen (rett linje; netto innflytting = -85,1 +2,5*boliger) er langt fra perfekt, sammenlignet med de observerte tallene. Det må ha vært andre faktorer enn boligbygging som har påvirket innflytting, for eksempel arbeidsmarkedsforhold i Vestby eller i fraflyttingskommunene.

Linjen representerer en tenkt sammenheng, m.a.o. en modell. Avvikene er modellfeil. Modellfeilene gjør at estimatene for *a* og *b* er usikre. Med litt andre observerte verdier ville også *a* og *b* fått andre estimerte verdier. Med andre ord: de estimerte verdiene for *a* og *b* (skrives som hhv $\hat{a}$ og $\hat{b}$) har en statistisk fordeling. Hvis vi antar en viss fordeling for modellfeilene, kan vi regne ut fordelingen til estimatene $\hat{a}$ og $\hat{b}$. Det er vanlig å anta at modellfeilene er normalfordelte ("klokkekurve") med forventning ("gjennomsnitt") lik null.

Med denne forutsetningen har estimatene $\hat{a}$ og $\hat{b}$ en bestemt fordeling (NB ikke nødvendigvis en normalfordeling!), med en bestemt forventning/gjennomsnitt og et bestemt standardavvik.

Forventningen til $\hat{b}$ er lik den ukjente verdien *b*. Standardavviket kan beregnes (bl.a. i Excel). Dette standardavviket kalles for estimatets standardfeil.

Stor standardfeil medfører at sjansen er stor for at estimatet $\hat{b}$ blir nokså forskjellig fra den virkelige (men ukjente) verdien for *b*.

Standardfeilen til $\hat{b}$ må sees i sammenheng med selve $\hat{b}$-verdien. En stor standardfeil er mindre alvorlig når $\hat{b}$ er stor, sammenlignet med en liten $\hat{b}$. Det samme gjelder estimatet $\hat{a}$, men det er særlig viktig for $\hat{b}$: hvis vi ikke kan stole på verdien til $\hat{b}$, vet vi ikke om den sanne *b*'en er positiv, negativ, eller kanskje null (ingen lineær sammenheng mellom X og Y!).

Sannsynligheten på at den virkelige *b* har motsatt fortegn, øker når standardfeilen til $\hat{b}$ øker i forhold til $\hat{b}$. Derfor er det vanlig å se på brøken $\frac{\hat{b}}{\text{standardfeil for } \hat{b}}$.

Denne brøken kalles for t-verdien til $\hat{b}$, og Excel regner den ut sammen med $\hat{b}$.

Tommelfingerregel: når t er større enn +2, eller mindre enn –2, er det bare maks. 2½ % sannsynlig at den virkelige *b* har motsatt fortegn sammenlignet med $\hat{b}$. Det er akseptabelt.

I vårt eksempel er t-verdien for $\hat{b}$ lik 1,60. Det betyr at sannsynligheten for en negativ verdi for b er større enn 2½%. (Den er faktisk rundt 6-7 prosent.)

R^2

En viktig størrelse som brukes for å karakterisere hvor godt modellen passer til dataene er R^2 ("coefficient of determination"). Den forteller oss hvor mye usikkerheten er blitt redusert etter at vi estimerte modellen. R^2 ligger mellom 0 og 1.

- $R^2 = 1$: perfekt lineær sammenheng mellom X og Y. Alle observerte datapunkter ligger på regresjonslinjen $Y = \hat{a} + \hat{b} * X$.
- $R^2 = 0$: ingen lineær sammenheng mellom X og Y.

Når den estimerte helningen $\hat{b}$ er positiv, er R, d.v.s. kvadratroten av R^2 også positiv. Når $\hat{b}$ er negativ, har regresjonslinjen $Y = \hat{a} + \hat{b} * X$ en negativ helning. I dette tilfellet er også R negativ. Både $R = +1$ og $R = -1$ avspeiler en perfekt lineær sammenheng. I begge tilfeller blir R^2 lik +1.

Jo nærmere R^2 ligger 1, desto bedre er modellens forklaringskraft. I vårt eksempel for Vestby er R^2 lik 0,16. Det betyr at mye av variasjonen i innflytting henger sammen med andre faktorer enn boligbygging.

NB1: Å finne en modell med størst mulig verdi for R^2 er ikke mål i seg selv i analysen. Det nytter ikke å finne to variabler X og Y som har veldig høy R^2, med mindre vi kan forklare hvorfor. F.eks. modellen $Y = a + b*X$ har like stor R^2 som $X = c + d*Y$!

NB2. En kan bevise at R^2 har samme verdi som kvadraten av utvalgets korrelasjonskoeffisient r_{XY}: $R^2 = r_{XY}^2$.

Generalisering
Multivariat regresjonsmodell: flere uavhengige variabler samtidig

Formel: $Y = a + b*X + c*Z$

Uavhengige variabler: X og Z. Konstant *a*, stigningstall *b* og stigningstall *c* estimeres fra data.

Z kan f.eks. være kommunens befolkningsstørrelse: jo større kommunen, desto mer innflytting.

Framgangsmåte i multivariat regresjon er den samme som før:

1) estimer verdiene av *a*, *b* og *c*;
2) sjekk t-verdiene for $\hat{a}$, $\hat{b}$ og $\hat{c}$;
3) beregn R^2.

Jfr. PC-øvelse i uke 42.

Anbefalt literatur: M.S. Lewis-Beck: "Applied Regression: An Introduction." Series Quantitative Applications in the Social Sciences nr. 22. Beverly Hills: Sage Publications 1980.