Multippel lineær regresion: Interaksjonseffekter

Ingeborg Hem Sørmoen, med hjelp fra Kenneth Aase

Dette er et kort notat om interaksjoner i multippel lienær regresjon (MLR). Se også læringsmaterialet om regresjon.

Superkort oppsummering

En interaksjonseffekt/samspillseffekt kan tenkes på som en ny forklaringsvariabel, bestående av produktet av to andre forklaringsvariabler. Legg til denne i modellen din, og du har en interaksjonseffekt. Merk at det ikke er alle interaksjonseffekter som gir mening, så det er viktig å undersøke betydningen av modellen din før du setter i gang!

I Python kodes dette som: `y ~ x + cat + x:cat` der y er respons og x og cat er forklaringsvariabler. `x:cat` lager interaksjonseffekten. Dette kan også skrives slik: `y ~ x*cat`, det blir det samme.

Endre referansekategori: `+ x*C(cat, Treatment("nyrefcat"))`.

Introduksjon

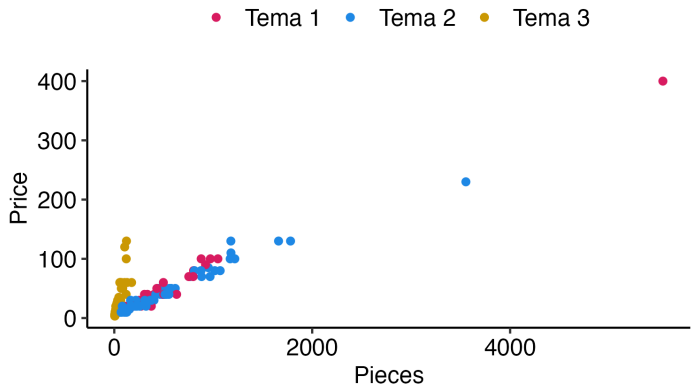
I dette eksempelet bruker vi et subsett av datasettet som benyttes i Oppgave 1 i prosjektet, nemlig LEGO-datasettet (hentet fra Peterson and Ziegler (2021)). Dette er et datasett bestående av en kontinuerlig responsvariabel (pris), en kontinuerlig forklaringsvariabel (antall brikker), og en kategorisk forklaringsvariabel (tema). Variablene er som følger:

`Theme` Tema settet hører til
`Pieces` Antall brikker i settet
`Price` Pris

Vi vil se hvordan pris kan beskrives av antall brikker i LEGO-settet. De øverste 6 radene i datasettet ser slik ut:

	Theme	Price	Pieces
1	Tema 2	9.99	47
2	Tema 3	129.99	123
3	Tema 3	19.99	15
4	Tema 2	99.99	1218
5	Tema 3	19.99	20
6	Tema 1	19.99	121

Hvis vi plotter dette datasettet får vi følgende figur:

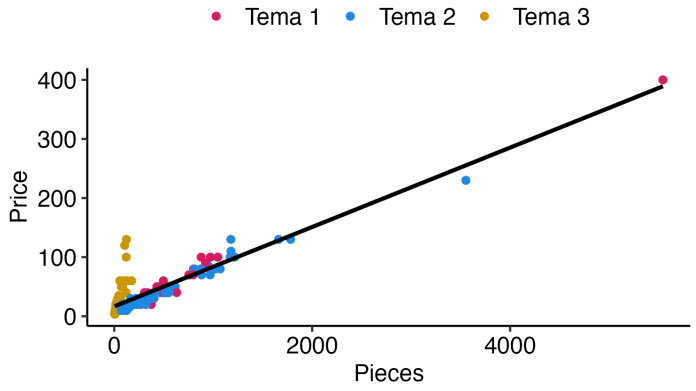


Her har vi antall brikker i settet på x-aksen, prisen på settet på y-aksen, og fargen forteller hvilket tema settet hører til.

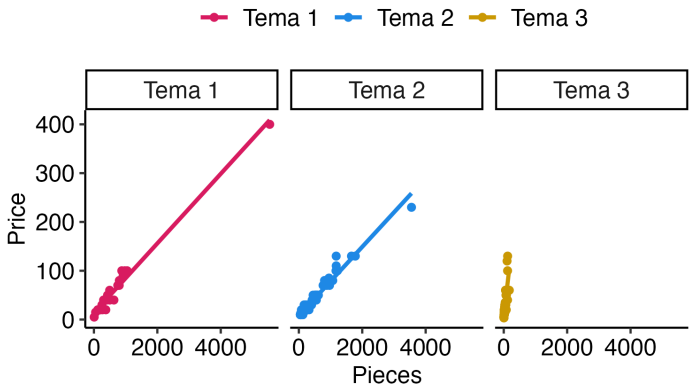
La oss nå tilpasse en enkel lineær regresjonsmodell til disse dataene, der pris er responsen og antall brikker er forklaringsvariabelen. Modellen vår er da:

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

der Y_i er respons, β_0 er skjæringspunkt, β_1 er stigningstall og e_i er feilleddet (se læringsmaterialet om MLR for detaljer). Ved å tilpasse denne modellen får vi følgende estimater av skjæringspunkt og stigningstall $\hat{\beta}_0 = 16.65376$ og $\hat{\beta}_1 = 0.06718$. Regresjonslinja ser slik ut:



Denne linja ser ut til å passe greit til Tema 1 og Tema 2, men dårlig til Tema 3. Derfor prøver vi nå og tilpasse tre ulike regresjonsmodeller (fremdeles enkel lineær regresjon) til dataene fra de tre ulike settene:



der modellene har følgende estimater:

- Tema 1: $\hat{\beta}_0 = 13.47586$ og $\hat{\beta}_1 = 0.07129$
- Tema 2: $\hat{\beta}_0 = 7.85741$ og $\hat{\beta}_1 = 0.07057$
- Tema 3: $\hat{\beta}_0 = 6.0666$ og $\hat{\beta}_1 = 0.5636$

Det er tydelig av både figuren og estimatene at linjene er ganske ulike (men vi har ikke sett på usikkerhet i estimatene ennå, så vi kan ikke vite at de er signifikant forskjellige!). Disse tre modellene gir oss totalt 9 parametere vi må estimere: skjæringspunkt, stigningstall og standardavviket til feilleddene, ganger tre.

Nå kunne det vært interessant å sammenligne de tre linjene for å se om de er signifikant forskjellige eller ikke. Vi forventer også at feilleddene i de tre modellene kommer fra den samme fordelingen (med det samme standardavviket), så her ser vi potensiale for å kunne låne informasjon mellom modellene. Vi får til begge deler ved å innføre **interaksjonseffekter** i modellen.

Interaksjoner

En interaksjonseffekt, eller samspillseffekt om du vil, er i en multippel lineær regresjonsmodell en forklaringsvariabel som består av mer enn én type informasjon. En interaksjonseffekt lar i praksis effekten av en av forklaringsvariabelene våre avhenge av verdien til en av de andre forklaringsvariablene. For eksempel, i vårt tilfelle kan vi la hvordan responsen (pris) varierer med forklaringsvariabelen (antall brikker i LEGO-settet) avhenge av verdien til en annen forklaringsvariabel (temaet til settet).

Vi starter med å kode om temasettene med dummy-variabelkoding (det skjer automatisk i Python, se læringsmaterialet om MLR for detaljer): La y_i være pris, $x_{i,pris}$ være antall brikker og $x_{i,tema}$ være tema. $x_{i,tema}$ er kategorisk, og vi må gjøre den om til en eller flere numeriske variabler. Siden den har tre nivåer, trenger vi to variabler for å representere den:

- $x_{i,t2}$ er 0 for Tema 1 og 3, og 1 for Tema 2
- $x_{i,t3}$ er 0 for Tema 1 og 2, og 1 for Tema 3

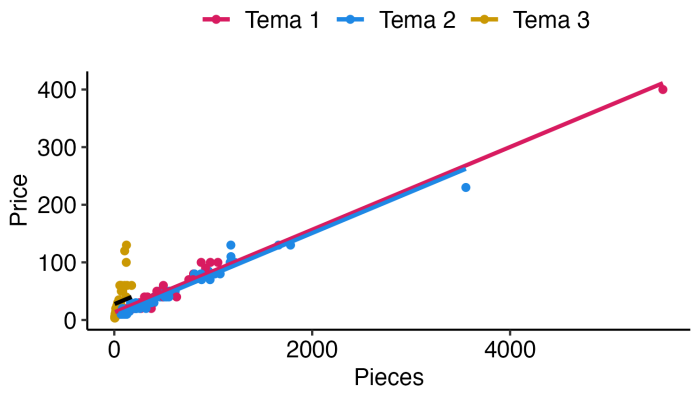
Vi kan utvide datasettet vårt med de nye forklaringsvariablene (men i Python skjer dette automatisk og vi trenger ikke gjøre dette), som vil se sik ut for de første 6 observasjonene:

	Theme	Price	Pieces	t2	t3
1	Tema 2	9.99	47	1	0
2	Tema 3	129.99	123	0	1
3	Tema 3	19.99	15	0	1
4	Tema 2	99.99	1218	1	0
5	Tema 3	19.99	20	0	1
6	Tema 1	19.99	121	0	0

Ved å tilpasse en modell der Theme er en forklaringsvariabel lar vi modellen gi hvert tema et eget skjæringspunkt. Stigningstallet vil fremdeles være felles. Modellen kan da skrives slik:

$$Y_i = \beta_0 + \beta_1 x_{i,Price} + \beta_2 x_{i,t2} + \beta_3 x_{i,t3}$$

og under ser vi denne modellen i en graf:



Merk at jeg har brukt svart for å markere regresjonslinja for Tema 3, fordi ellers går den i ett med observasjonene og blir umulig å se. Estimatene er som følger:

- $\hat{\beta}_0 = 13.12844$
- $\hat{\beta}_1 = 0.07179$
- $\hat{\beta}_2 = -5.80309$
- $\hat{\beta}_3 = 14.34460$

og gir oss tre ulike regresjonslinjer, én for hvert tema. Dette var ikke en god nok løsning: vi har i praksis bare flyttet regresjonslinja opp/ned for de forskjellige temaene. Forholdet mellom pris og antall brikker er helt klart annerledes for Tema 3 enn de to andre, og å anta at det samme stigningstallet kan brukes virker som en dårlig antagelse.

Derfor innfører vi en *interaksjonseffekt* mellom antall brikker og tema. Vi kan tenke at vi lager oss *nye* forklaringsvariabler, som er lik $x_{i,Price} \cdot x_{i,t2}$ og $x_{i,Pieces} \cdot x_{i,t3}$. Resultatet blir det samme. Med de nye variablene vil starten av dataene se slik ut:

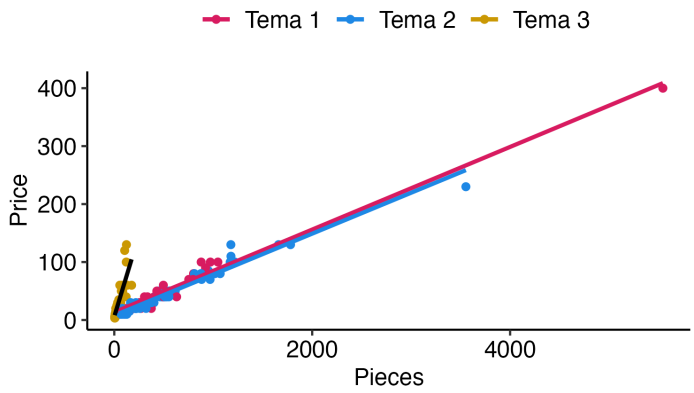
	Theme	Price	Pieces	t2	t3	Pieces, t2	Pieces, t3
1	Tema 2	9.99	47	1	0	47	0
2	Tema 3	129.99	123	0	1	0	123
3	Tema 3	19.99	15	0	1	0	15
4	Tema 2	99.99	1218	1	0	1218	0
5	Tema 3	19.99	20	0	1	0	20
6	Tema 1	19.99	121	0	0	0	0

De "nye" variablene $x_{i,Pieces} \cdot x_{i,t2}$ og $x_{i,Pieces} \cdot x_{i,t3}$ er verdien til Pieces når temaet er hhv Tema 2 og Tema 3, og 0 ellers! Dermed får vi et ekstra stigningstall for Tema 2 (β_4) og Tema 3 (β_5), som vi skal legge oppå det vi allerede har for hele modellen (β_1). Da har vi én modell som tillater tre ulike regresjonslinjer, avhengig av tema.

Nå kan vi skrive ned modellformelen for en modell med interaksjoner:

$$Y_i = \beta_0 + \beta_1 x_{i,Pieces} + \beta_2 x_{i,t2} + \beta_3 x_{i,t3} + \beta_4 x_{i,Pieces} \cdot x_{i,t2} + \beta_5 x_{i,Pieces} \cdot x_{i,t3} + e_i$$

Nå har vi fått hele seks β-er vi må estimere, og så har vi i tillegg standardavviket til feilleddene. Men, det er bare 7 parametere totalt, altså to mindre enn da vi tilpasset tre ulike modeller. Vi tilpasser modellen, og grafisk ser den nå slik ut (igjen med en sort strek for Tema 3 for synlighetens del):



Merk at dette ligner veldig mye på de tre ulike modellene vi plottet tidligere, men denne modellen er forskjellig fra dem fordi vi har én istedenfor tre modeller, og kun ett estimert standardavvik (felles for alle temaene). Estimatene er:

- $\hat{\beta}_0 = 13.4758574$
- $\hat{\beta}_1 = 0.0712949$
- $\hat{\beta}_2 = -5.6184461$
- $\hat{\beta}_3 = -7.4092521$
- $\hat{\beta}_4 = -0.0007267$
- $\hat{\beta}_5 = 0.4922606$

Dette tilsvarer følgende regresjonslinjer:

- $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,Pieces} = 13.4758574 + 0.0712949 x_{i,Pieces}$ for Tema 1
- $y_i = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_4) x_{i,Pieces} = 7.857411 + 0.0705682 x_{i,Pieces}$ for Tema 2
- $y_i = (\hat{\beta}_0 + \hat{\beta}_3) + (\hat{\beta}_1 + \hat{\beta}_5) x_{i,Pieces} = 6.066605 + 0.5635555 x_{i,Pieces}$ for Tema 3

Vi får så og si de samme tallene for regresjonslinjene som for de tre individuelle modellene. Merk at vi alltid kan regne ut disse linjene fordi forventningsverdien til $X + Y$ alltid er $E(X) + E(Y)$, men usikkerheten i linja er ikke like enkel å regne ut, da koeffisientene ikke kan antas å være uavhengige. Dette er allikevel to ganske ulike situasjoner (én vs. tre modeller), selv om linjene blir like, blant annet fordi vi kan finne ut mer om forskjellen på de tre linjene i den ene store modellen.

Tolkning og usikkerhet

Hittil i dette notatet har vi ikke tatt for oss usikkerheten i koeffisient-estimatene, eller gjort noen form for testing på verdiene av dem. Det dekkes i annet læringsmateriale, så hvis du ikke er helt

stø på usikkerhet kan det være lurt å repetere det før du leser videre. Vi kan vurdere usikkerheten til koeffisientene på akkurat samme måte som i MLR, også for interaksjonskoeffisientene. Et godt argument for hvorfor er at vi i praksis bare har laget oss flere variabler, og dermed får flere koeffisienter. Det er altså ikke noe ekstra vi må kunne for å se på signifikansen til ett og ett estimat, som før.

Men, det vi derimot må ha kontroll på, er hva betydningen av estimatene er, og hvordan vi kan bruke usikkerheten i dem til å avgjøre om det er forskjeller mellom de ulike LEGO-temaene. Under er en tabell med estimatene med tilhørende standardfeil, t-verdi og p-verdi:

Parameter	Estimat	SE	t-verdi	p-verdi
β_0	13.4759	2.9603	4.55	0.0000
β_1	0.0713	0.0023	30.98	0.0000
β_2	-5.6184	3.4424	-1.63	0.1048
β_3	-7.4093	3.8523	-1.92	0.0564
β_4	-0.0007	0.0034	-0.21	0.8326
β_5	0.4923	0.0419	11.74	0.0000

Vi antar at vi er interesserte i å se på resultatene på et 5% signifikansnivå.

Her ser vi at skjæringspunktene for Tema 2 (β_2) og Tema 3 (β_3) som legges til det felles skjæringspunktet (β_0) ikke er signifikant forskjellige fra 0, da den tilhørende p-verdien er større enn 0.05 for begge to. Tema har altså lite å si for skjæringspunktet (merk at skjæringspunktet i teorien beskriver pris for et LEGO-sett med 0 brikker, og den burde jo være 0, men vi inkluderer skjæringspunkt i modellene våre og legger ikke vekt på videre tolkning av dette).

Den ekstra delen av stigningstallet for Tema 2, β_4, er heller ikke signifikant forskjellig fra 0, og vi har dermed ikke grunnlag for å påstå at Tema 2 har et stigningstall forskjellig fra det Tema 1 har. Det høres riktig ut basert på hva vi har sett tidligere. Tema 3 burde derimot ha et eget stigningstall, for den ekstra delen av stigningstallet til Tema 3, β_5, er absolutt ulikt 0 på signifikansnivået vi har valgt oss.

Merk: Det går an å teste forskjellen mellom den ekstra delen av stigningstallet for Tema 2 og for Tema 3 (altså, β_4 vs. β_5). Dette kunne vært relevant for eksempel hvis Tema 3 var referansekategorien. Men, dette er ikke like rett frem som å teste om koeffisientene er signifikant forskjellige fra 0 (vi må blant annet huske at β-ene ikke nødvendigvis er uavhengige!). Vi kan unngå dette ved å velge en god referansekategori, i dette tilfellet var det Tema 1 (eller 2). I Python velger og endrer du referansekategorien slik:

- `+ Theme` i formelen gir den alfabetisk første kategorien (Tema 1) som referansekategori
  `+ C(Theme, Treatment("Tema 1"))` i formelen det samme som `+ Theme`
- `+ C(Theme, Treatment("Tema 2"))` i formelen gir Tema 2 som referansekategori
- `+ C(Theme, Treatment("Tema 3"))` i formelen gir Tema 3 som referansekategori

Andre steder du kan lese om interaksjoner: Store Norske leksikon (for et annet eksempel enn LEGO), og `https://en.wikipedia.org/wiki/Interaction_(statistics)#In_regression`Wikipedia (på engelsk og ganske detaljert).

Referanser

Peterson, A. D. and Ziegler, L. (2021). Building a Multiple Linear Regression Model With LEGO Brick Data. *Journal of Statistics and Data Science Education*, 29(3):297–303.