NTNU

Kunnskap for en bedre verden

ISTG1003 - Statistikk

Prosjekt - Oppgave 1

Betaler vi ekstra for LEGO som assosieres med et varemerke?

21.09.2023

1 Introduksjon

Koden brukt i denne oppgaven er tilgjengelig her: https://github.com/olemorud/istg1003-prosjekt

I denne rapporten analyseres et datasett fra Peterson og Ziegler[5], som omfatter 1304 LEGO-sett fra perioden 1. januar 2018 til 11. september 2020. Hvert sett i datasettet er beskrevet med variabler som navn, tema, antall brikker, pris, antall sider i instruksjonsmanualen, antall LEGO-figurer, og antall unike brikker. Gjennom multippel lineær regresjon, utforskes sammenhengene mellom disse variablene, med spesielt fokus på prisdannelse og faktorer som potensielt påvirker denne.

2 Problemstilling

Hovedfokuset i rapporten er å undersøke prisforskjeller i LEGO-sett basert på deres assosiasjon med lisensierte varemerker. Spesifikt adresseres problemstillingen: **"Betaler vi ekstra for LEGO som assosieres med et varemerke?"**. Analysen inkluderer en undersøkelse av hvordan lisensiering påvirker LEGOs prissetting, og utforsker forskjellene i prising mellom LEGO-sett med egne varemerker og de med eksternt lisensierte varemerker.

3 Gruppering av data

Før observasjonene i datasettet ble gruppert, ble det foretatt en gjennomgående rensking av datasettet. Dette inkluderte å omtolke pris (i dollar) til flyttall og fjerning av:

- Overflødige forklaringsvariabler
- Observasjoner med manglende datapunkter
- Uønskede tegn

Etter renskingen ble de resterende LEGO-settene delt i tre tema: *lisensiert*, *ulisensiert* og *usikker*. Grupperingen er mye basert på informasjon om lisenser i Wikipedia-artikkelen *List of Lego Themes* [6], men er manuelt bestemte. Lisensierte LEGO-sett er assosierte med varemerker som f.eks. *Spider-Man* og *Star Wars*, hvorav LEGO-sett av typen *Ninjago* og *City* er eksempler på sett som ikke er assosierte med varemerker utenfor LEGO selskapet. Vi har plassert tema som inneholder blandinger av lisensiert og ikke-lisensiert i kategorien *usikker*.

4 Modell og hypotese

4.1 Begreper

- *Enkel lineær regresjon*: I enkel lineær regresjon finner man et utrykk for en linje

$$y = \beta_0 + \beta_1 x \tag{1}$$

  som passer datapunktene best etter et kriterium. Et vanlig kriterium er *least squares*, der det forsøkes å minimere arealet av kvadratene som kan tegnes mellom regresjonslinjen og datapunktene.

- *Multippel lineær regresjon*: I multippel lineær regresjon uten interaksjon finnes det flere forklaringsvariabler

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + e \tag{2}$$

- *Interaksjon*: Multippel lineær regresjon med interaksjonsledd brukes når koeffisienten av én variabel er avhengig av en annen variabel. Dette uttrykkes som[3]

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2 + e \tag{3}$$

  En modell der timer søvn og timer med lesing er brukt til å predikere karakteren, er et eksempel der multippel lineær regresjon med interaksjon passer bra. Ettersom en elev trenger å både studere og sove for å gjøre det bra på en prøve, er det en interaksjon. Å sove 10 timer er ikke ekvivalent med 3 timer lesing, de er avhengige av hverandre.

- *Responsvariabel*: En variabel som skal predikeres. I en enkel lineær regresjonsmodell kalles denne variabelen y (Eq.1) [1]

- *Forklaringsvariabel*: En variabel som blir brukt for å predikere responsvariabelen. I en regresjonsmodell får forklaringsvariabel n koeffisienten β_n (Eq.1) [1]

4.2 Modeller og Hypotese

For å svare på problemstillingen blir det først funnet en modell for pris av LEGO-sett som ikke tar lisens i betraktning. Det bør da finnes en lineær sammenheng mellom en forklaringsvariabel og pris. Ett eksempel på dette er sammenhengen mellom antall brikker og pris. En formell definisjon av dette, som er et mulig premiss for at problemstillingen kan undersøkes, er da:

$$H_1 : \beta_{\text{Brikker}} > 0 \tag{4}$$

og nullhypotesen som skal avkreftes er

$$H_0 : \beta_{\text{Brikker}} = 0 \tag{5}$$

Deretter må det vises at multippel lineær regresjon, med lisens-status som en kategorisk variabel i interaksjon med forklaringsvariabel, viser en økning i pris. Det brukes interaksjon da det er ønskelig å se om prisen per brikke er høyere avhengig av LEGO-settet sin lisens-status. Dersom antall brikker er en god forklaringsvariabel for pris av LEGO-sett, blir den formelle definisjonen av hypotesen

$$H_1 : \beta_{\text{Brikker,Lisensiert}} > 0 \tag{6}$$

og nullhypotesen blir

$$H_0 : \beta_{\text{Brikker,Lisensiert}} = 0 \tag{7}$$

Med andre ord er fremgangsmåten å teste enkel lineær regresjon med alle kontinuerlige forklaringsvariabler med pris som responsvariabel, og deretter utvide modellen til å bruke lisensstatus som en forklaringsvariabel i interaksjon med den best egnede forklaringsvariabelen. Dersom det ikke finnes en enkel lineær regresjon som modellerer prisen godt må mer komplekse modeller undersøkes. Enklere modeller foretrekkes fremfor mer komplekse modeller, fordi de er vanskelig å forstå og de kan overtilpasses datasettet.

5 Tilpasning og evaluasjon av modellen

I utforskningen av datasettet ble det dannet et kryssplot for hver kontinuerlige forklaringsvariabel: *Page*, *Pieces*, *Unique_Pieces* og *Minifigures*. Med unntak av *Minifigures* virker det som det er en lineær sammenheng mellom alle forklaringsvariabler og pris.

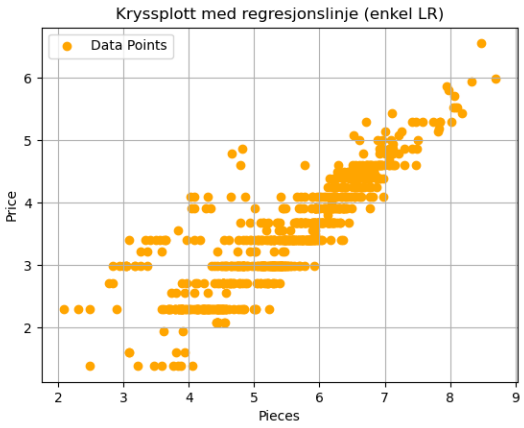

Fig. 1. Kryssplott av pris og antall brikker

En lineær sammenheng mellom f.eks. pris og antall brikker tilsier egentlig en pris per brikke. Derfor ble datasettet sortert på pris per brikke, pris per minifigur, osv., for å undersøke variasjonen i disse forholdene. Her ble det observert at LEGO-sett for barn i aldersgruppen 2+ (DUPLO) hadde en betydelig høyere pris per brikke. Årsaken er at LEGO-brikker i DUPLO-sett er to ganger større enn normal LEGO[2], og inneholder mange store ferdigstøpte figurer. I tabellen under er datasettet sortert etter pris per brikke.

Set_Name	Theme	Pieces	Price	Ages	Price_Per_Pieces
Ocean's Bottom	Classic	579.0	29.99	Ages_5-99	0.051796
Formula E ...	Speed Champions	565.0	29.99	Ages_8+	0.053080
Diagon Alley	Harry Potter	374.0	19.99	Ages_10+	0.053449
Fun Future	Classic	186.0	9.99	Ages_5-99	0.053710
Central Perk	Ideas	1070.0	59.99	Ages_16+	0.056065
...	...	...	...	...	...
Wrecking Ball ...	DUPLO	56.0	59.99	Ages_2+	1.071250
Cargo Train	DUPLO	105.0	119.99	Ages_2-5	1.142762
Elsa and Olaf ...	DUPLO	17.0	19.99	Ages_2+	1.175882
Police Bike	DUPLO	8.0	9.99	Ages_2+	1.248750
T. rex Tower	Jurassic World	22.0	29.99	Ages_2-5	1.363182

Dersom datasettet skal modelleres lineært burde pris/forklaringsvariabel ha et normalfordelt standardavvik. Til høyre er fordelingen av pris/brikke. Stolpene langt til høyre er et resultat av DUPLO settene

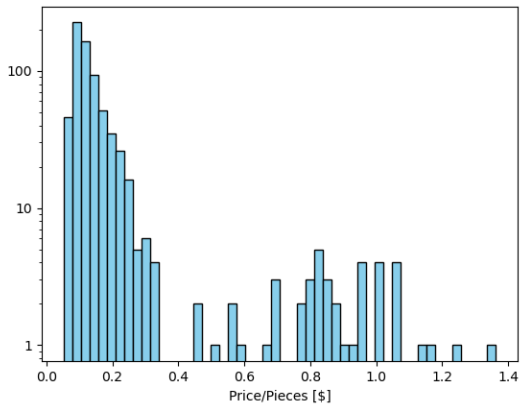

Fig. 2. Fordeling av pris per brikke på en log-skala

Lineær regresjon ble utført på *Pages*, *Pieces* og *Unique_Pieces* mot pris for å sammenligne de. Her er resultatet av modellene for *Pages* og *Pieces*, som er de beste kandidatene

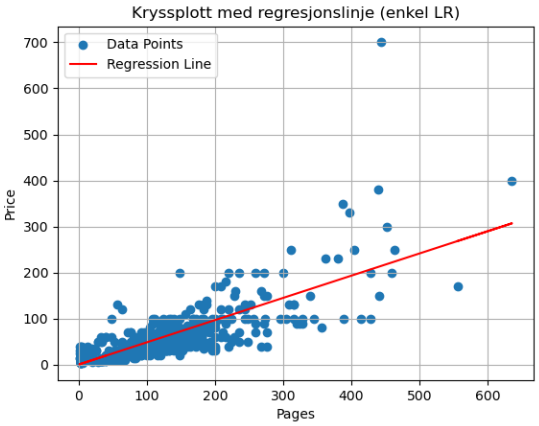

Fig. 3. Regresjon av $Price \sim Pages$

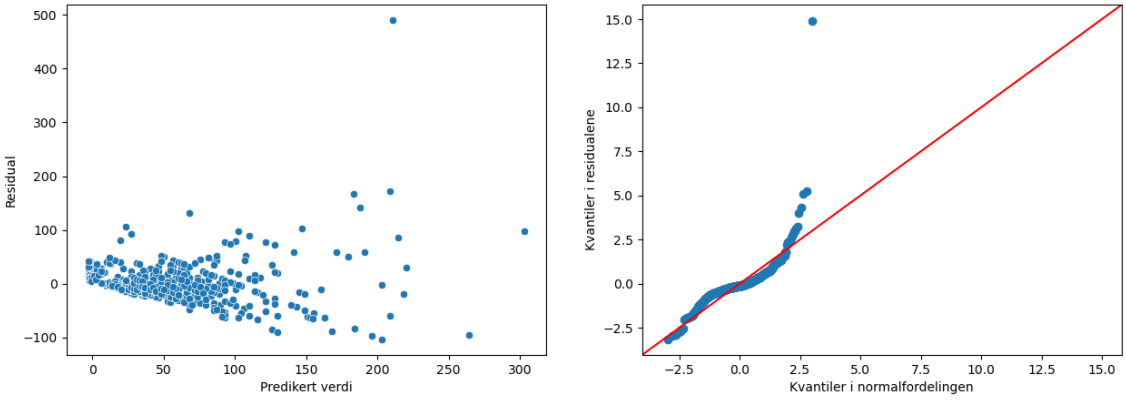

Fig. 4. QQ-plot av $Price \sim Pages$

Fordi området for residualer ikke er konstant, så anvendes en logaritmisk transformasjon av responsvariabelen for å stabilisere. Forklaringsvariabelen blir også transformert for å beholde den lineære sammenhengen. Etter log-transformasjonen blir området for residualer mye mindre variabelt enn før. Her er regresjonen av $Price \sim Pieces$ etter transformasjonen.

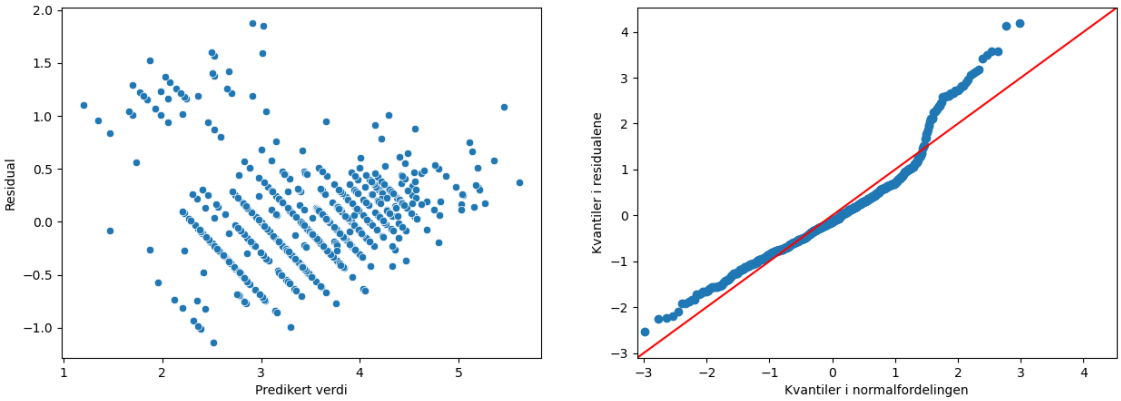

Fig. 5. Regresjon av $ln(Price) \sim ln(Pieces)$

At høyre side har en tung hale kan forklares av DUPLO sett [1]. *Pieces* har ellers den største justerte R^2 verdien av alle modellene i det transformerte datasettet, på 0.727. Dette vil si at modellen passer til 73% av datapunktene.

Dersom LEGO-sett assosiert med varemerker koster mere, burde en multippel lineær regresjon med varemerkestatus som dummy-variabel vise det. *Category* defineres som dummy-variabelen, og det lages en modell med interaksjon mellom *Category* og *Pieces*, som gir resultatet:

Table 2: OLS Regression Results

Dep. Variable:	Price	R-squared:	0.732
Model:	OLS	Adj. R-squared:	0.730
Method:	Least Squares	F-statistic:	385.9
Date:	Fri, 17 Nov 2023	Prob (F-statistic):	1.98e-199
Time:	20:09:24	Log-Likelihood:	-431.88
No. Observations:	714	AIC:	875.8
Df Residuals:	708	BIC:	903.2
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.2486	0.148	-1.675	0.094	-0.540	0.043
Category[T.not_licensed]	0.2237	0.187	1.196	0.232	-0.144	0.591
Category[T.uncertain]	-0.6037	0.293	-2.059	0.040	-1.179	-0.028
Pieces	0.6815	0.026	25.904	0.000	0.630	0.733
Pieces:Category[T.not_licensed]	-0.0451	0.034	-1.342	0.180	-0.111	0.021
Pieces:Category[T.uncertain]	0.0868	0.051	1.698	0.090	-0.014	0.187

Her ser vi at p-verdien til *Pieces:Category[T.not_licensed]* er over 0.05, og vi kan ikke forkaste nullhypotesen. Vi ikke med sikkerhet si om assosiasjon med varemerker har en innvirkning på pris.

[1] I Jupyter-notebooken vises en seksjon der DUPLO er fjernet

6 Konklusjon

I denne analysen ble det undersøkt om LEGO-sett som er assosierte med varemerker koster mer enn LEGO-sett som ikke er det. Ved bruk av multippel lineær regresjon der pris ble predikert av antall brikker i interaksjon med lisens, hadde interaksjonen mellom lisens og brikker en for høy p-verdi ($0.180 > 0.05$) til å avvise nullhypotesen. Det er da ikke mulig å si utifra analysen at LEGO-sett assosiert med varemerke koster mere enn sett ikke assosiert med varemerke.

Referanser

[1] Jim Frost. *Independent and Dependent Variables*. Accessed: November 18, 2023. Copyright © 2023 Jim Frost. 2023. URL: https://statisticsbyjim.com/regression/independent-dependent-variables/.

[2] Tony Hafner. *What are the exact dimensions of a DUPLO brick?* Accessed: November 18, 2023. Content licensed under CC BY-SA. 2023. URL: https://bricks.stackexchange.com/a/1600.

[3] Ingeborg Hem Sørmoen and Kenneth Aase. *Multippel lineær regresion: Interaksjonseffekter*. Accessed: November 18, 2023. 2023. URL: blackboard%20ISTG1003:%20notat_interaksjoner.pdf.

[4] Mette Langaas. *IST[A/G/T]1003: Statistisk læring og data science*. 2020. URL: https://www.math.ntnu.no/emner/IST100x/ISTx1003/Regresjon.html#Enkel_line%C3%A6r_regresjon.

[5] Anna D. Peterson and Laura Ziegler. 'Building a Multiple Linear Regression Model with LEGO Brick Data'. In: *Journal of Statistics and Data Science Education* 29.3 (2021), pp. 297–303. DOI: 10.1080/26939169.2021.1946450. URL: https://doi.org/10.1080/26939169.2021.1946450.

[6] Wikipedia contributors. *List of Lego Themes*. Accessed: November 18, 2023. 2023. URL: https://en.wikipedia.org/w/index.php?title=List_of_Lego_themes&oldid=1184716316.