NTNU

NTNU | Fakultet for informasjons-
teknologi og elektroteknikk
Institutt for datateknologi
og informatikk

Fordeling av unike brikker i LEGO-sett basert på kjønn

ISTT1003 Statistikk

Multippel lineær regresjon i analyse av LEGO-datasett

Faglærer

Ingeborg Hem Sørmoen

Innhold

1 Problemstilling

LEGO-sett er ofte designet med en bestemt målgruppe basert på kjønn. Denne rapporten ønsker å utforske den potensielle sammenhengen mellom kjønn og mengden av unike LEGO-brikker tilgjengelig i settene. Problemstillingen for denne rapporten er dermed formulert som følger: "Har LEGO-sett tilpasset gutter flere unike brikker enn LEGO-sett tilpasset jenter?".

Forklaringsvariablene som skal analyseres er "kjønn" og "det totale antallet brikker til stede i hvert LEGO-sett", samt samspillseffekten mellom kjønn og det totale antall brikker. Kjønn er delt inn i kategoriene gutt, jente og kjønnsnøytralt/unisex. Det er rimelig å anta at en økning i det totale antallet brikker i et sett også gir en økning i antallet unike brikker. Problemstillingen ønsker å utforske om "gutte-sett" har en høyere andel unike brikker enn "jente-sett", og ikke om "gutte-sett" generelt har flere brikker enn "jente-sett". Derfor inkluderes "antall totale brikker" som forklaringsvariabel, for å ta hensyn til forholdet mellom unike og totale brikker i LEGO-settene.

Responsvariabelen til problemstillingen er definert som "antallet unike brikker basert på kjønn", som gir en indikasjon på variasjonen av unike LEGO-brikker mellom målgruppene. Ved å analysere denne responsvariabelen, har rapporten som hensikt å undersøke om det er en betydelig variasjon i unike brikker basert på kjønn i LEGO-settene.

2 Begreper og forkortelser

Dataframe	Tabelloversikt over data
Kategorisk forklaringsvariabel	Forklaringsvariabel med diskret verdier
Kontinuerlig forklaringsvariabel	Forklaringsvariabel med kontinuerlige verdier
Korrelasjonskoeffisient	Lineær sammenheng mellom to variabler
MLR	Multippel Lineær Regresjon
One-hot koding	Konvertering av kategoriske verdier til numeriske

3 Teori

3.1 Enkel lineær regresjon

En enkel lineær regresjonsmodell er en modell som prøver å best mulig predikere nye data (respons, for eksempel sannsynlighet for å vinne en kamp) fra en forklaringsvariabel (predikater, for eksempel antall mål scoret) (Langaas, 2020). Regresjonsmodellen består av skjæringspunktet β_0 og stigningstallet β_1 hvor formelen for minste kvadratsums estimator minimerer avvikene mellom datapunktene og regresjonslinja.

Uansett hvor spredt eller komprimert dataen er kan det kalkuleres en regresjonsmodell som kan predikere nye data med variabel nøyaktighet. Jo mer lineær sammenhengen mellom prediktor og respons er, i tillegg til hvor nært dataen følger denne lineære sammenhengen, desto bedre blir regresjonsmodellen. Gode regresjonsmodeller kan predikere nye datapunkter basert på forklaringsvariabelen.

3.2 Multippel lineær regresjon

MLR er en utvidelse av enkel lineær regresjon hvor det kan være to eller flere forklaringsvariabler (for eksempel temperatur og årstid) som gir en respons (for eksempel hvor mye det vil regne) (Langaas,

2020). Forklaringsvariablene i MLR kan enten være kontinuerlige hvor variabelen er en mengde av noe (temperatur), eller kategoriske hvor diskre verdier kan aktivere effekten til forskjellige tilstander (om det regner eller ikke).

3.3 Interaksjonseffekter

MLR modellerer i utgangspunktet med antagelsen om at alle forklaringsvariablene er uavhengige, selv om effekten de har på hverandre bør tas hensyn til for å få en mer nøyaktig modell (Sørmoen & Aase, 2023). Ved å ta produktet fra to forklaringsvariabler får man en ny forklaringsvariabel som representerer interaksjonseffekten de består av.

3.4 R-kvadrert

I multippel lineær regresjonsanalyse er en relevant måleverdi "R-kvadrert", også kjent som determinasjonskoeffisienten (Løvås, 2018, s. 309). R-kvadrert er utledet av korrelasjonskoeffisienten, men kvadrert. Verdien til R-kvadrert vil si noe om hvor stor andel av variasjonen i datasettet som kan forklares av regresjonsmodellen/regresjonslinja. Determinasjonskoeffisientens verdi sier altså noe om hvor godt en regresjonsmodell passer dataene som undersøkes. R-kvadrert kan ta en verdi i intervallet [0, 1], og desto nærmere 1 verdien ligger, desto bedre passer regresjonsmodellen datasettet som undersøkes.

3.5 Hypotesetesting

Hypotesetesting benyttes hver gang en hypotese/påstand skal bevises/motbevises (Løvås, 2018, s. 255). Påstanden som skal prøves ut, kalles for den alternative hypotesen H_1, mens den "motsatte" hypotesen kalles for nullhypotesen H_0. Dersom en ønsker å bevise en hypotese som påstår at en verdi er høyere enn basisgrunnlaget H_0, så kan en benytte en høyresidig hypotesetest. I et slikt tilfelle vil H_1 for eksempel kunne være "verdien x er høyere enn antagelsen y" og H_0 kunne være "det er ingen forskjell i differansen mellom verdiene x og y".

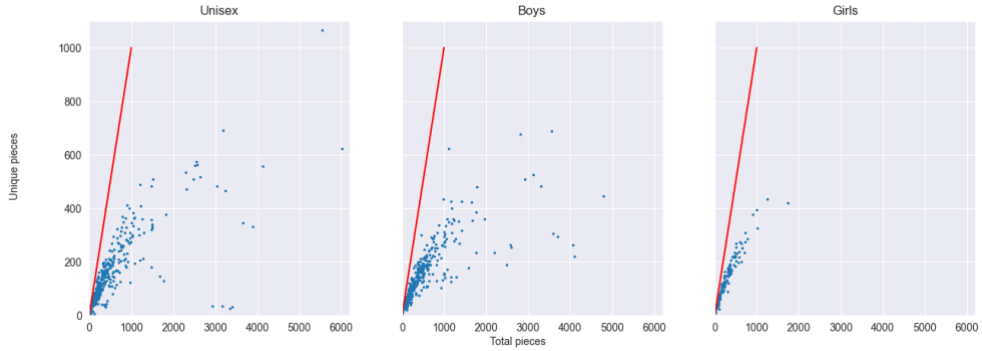
4 Datapreprosessering

Preprosesseringen av datasettet starter med å hente ut kolonnene "Theme", "Pieces" og "Unique_Pieces" og lagre disse i en dataframe. Deretter fjernes alle rader/datapunkter hvor minst én av kolonnene mangler data. Til slutt settes det inn en ny kolonne, "Gender", hvor de valgte verdiene baserer seg på kolonnen "Theme". Datapunkter i kolonnen "Gender" verdiene "Boy", "Girl" eller "Unisex".

Kategoriseringen av LEGO-settene sin «kjønnsmålgruppe» ble subjektivt fordelt ut ifra det enkelte settets tema. For å plassere hvert enkelt LEGO-sett inn i kategoriene «gutt», «jente» og «unisex» ble følgende parametere om settets tema vurdert. Temaets generelle fargepalett ble vurdert, der lyse farger samt rosa og pastellfarger vektes i retning av målgruppen «jente», mens mørke farger vektes i retningen av «gutt». LEGO-sett med nøytrale farger vektes mot «unisex». Videre ble temaets interesser vurdert. Her ble temaer med fokus på anleggsmaskiner, superhelter og slåssing plassert i gutt-kategorien, mens temaer som fokuserte på familierelasjoner, hester og vennskap ble vurdert i retning mot jente-kategorien. Til slutt ble tradisjonelle kjønnsassosiasjoner og historiske trender tatt med i beregningen. Her ble det vurdert at action- og eventyrtemaer appellerer mer til gutter, mens kreasjons- og vennskapsrelaterte temaer appellerer mer til jenter. Det er gruppens subjektive helhetsvurdering av alle overnevnte faktorer som har vært med å kategorisere temaene. Til slutt er det viktig å påpeke at denne inndelingen ikke er absolutt eller universell, ettersom kjønnsinndeling er komplekst og påvirkes av subjektivitet. Det er derfor viktig å ikke se på denne inndelingen som en

fasit, men som en generell indikator. Den endelige inndelingen kan ses i vedlegg 10.1 LEGO-sett kjønnsfordeling.

Etter kategoriseringen vil datasettet inneholde 283 unisex-sett, 279 guttesett og 152 jentesett.



Figur 1 Plott av antall brikker (x-akse) og unike brikker (y-akse), og en maksimumslinje (rød, y=x).

5 Modeller og hypoteser

For å svare på problemstillingen er det brukt en multippel lineær regresjonsanalyse (3.2 Multippel lineær regresjon). Det er tatt i bruk en kontinuerlig forklaringsvariabel for antall brikker i settet totalt og to kategoriske one-hot kodede variabler: én for gutt og én for unisex. Det er også to interaksjonseffekter som beskriver hvor mye antall brikker påvirker responsen, om det er gutt eller unisex. Jente er basisvariabelen der responsen kan formuleres som:

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

Interaksjonseffektene/samspillseffektene ligger i leddene $(\beta_4 x_{i,Boy} \cdot x_{i,Pieces})$ og $(\beta_5 x_{i,Unisex} \cdot x_{i,Pieces})$ fordi disse er bygd opp av produktet av to andre forklaringsvariabler (3.3 Interaksjonseffekter). Det er passende med en samspillseffekt i denne analysen fordi regresjonslinjene for hver kjønnskategori er relativt ulik. Ved å introdusere en samspillseffekt, vil variasjonen i responsen med hensyn til forklaringsvariabelen "kjønn" også være avhengig av en annen forklaringsvariabel, nemlig "antall brikker totalt".

Før problemstillingen kan besvares må det bestemmes om det er en forskjell mellom gutter og jenter med hensyn til antall brikker totalt. For å finne ut av dette, benyttes hypotesetesten under (3.5 Hypotesetesting).

H_0: β_4 er *ikke* positiv med signifikant margin
H_1: β_4 er positiv med signifikant margin

Skjæringspunktene er vurdert til å ikke ha noen relevant påvirkning på problemstillingen og er derfor ikke en del av hypotesetesten.

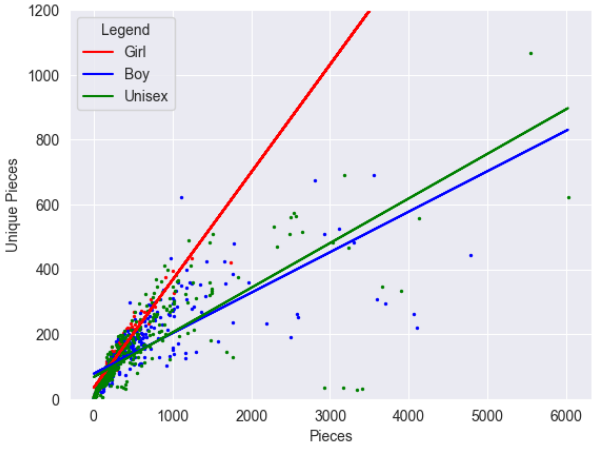
6 Modelltilpasning og evaluering

Opprinnelig var ikke interaksjonseffekter inkludert i regresjonsmodellen, noe som raskt resulterte i motsigende resultater fra det som var observert med visuell analyse. Det ble derfor bestemt å inkludere interaksjonseffekten mellom kjønn og antall brikker.

Med interaksjonseffekter inkludert blir det i praksis tre forskjellige regresjonsanalyser hvor stigningstallet til hvert kjønn beskriver hvor mange unike brikker et legosett har basert på det totale antallet brikker. Estimatene ble som følger:

$$\beta_0 = 34{,}1858,\ \beta_1 = 43{,}3865,\ \beta_2 = 33{,}1274,\ \beta_3 = 0{,}3324,\ \beta_4 = -0{,}2074,\ \beta_5 = -0{,}1946$$

$$Y_{i,Girl} = \beta_0 + \beta_3 x_{i,Pieces} = 34{,}1858 + 0{,}3324 x_{i,Pieces}$$
$$Y_{i,Boy} = (\beta_0 + \beta_1) + (\beta_3 + \beta_4) x_{i,Pieces} = 77{,}5723 + 0{,}125 x_{i,Pieces}$$
$$Y_{i,Unisex} = (\beta_0 + \beta_2) + (\beta_3 + \beta_5) x_{i,Pieces} = 67{,}3132 + 0{,}1378 x_{i,Pieces}$$

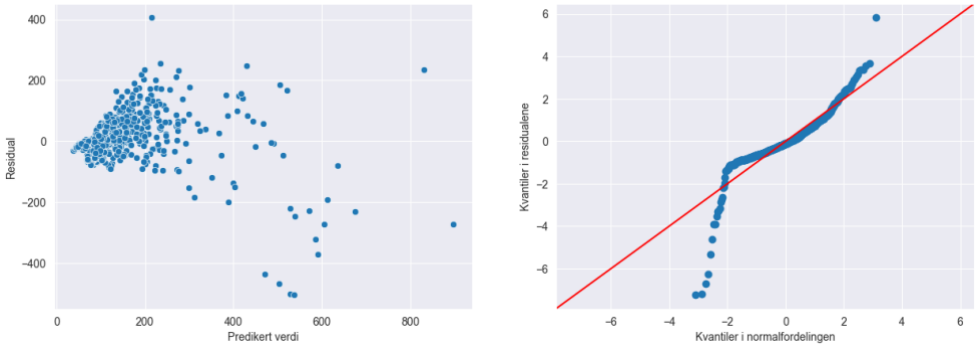


Figur 2 Multippel lineær regresjonsmodell.

Stigningstallet for jenter er 0,3324 og kan tolkes som at generelt sett er 33,24% av brikkene i legosett for jenter unike. For gutter blir stigningstallet $(0{,}3324 - 0{,}2074) = 0{,}125 = 12{,}5\%$ som illustrerer at gutter generelt sett har mer enn to ganger færre unike brikker enn jenter.

Den tosidige OLS-analysen i Python viser at både skjæringspunktene og stigningstallene til alle kjønn er statistisk signifikant forskjellig fra basis hvor alle p-verdiene er mindre enn 0.001. Ettersom p-verdiene til en ensidig analyse er halvparten av p-verdiene til en tosidig analyse, viser p-verdiene at gutter ikke har flere unike brikker enn jenter.

R-kvadrert får en verdi på 0,627. Dette betyr at 62,7% av variasjonen i datasettet kan forklares av regresjonsmodellen (3.4 R-kvadrert).



Figur 33 Residual- og QQ-plott.

Ved analyse av datapunktene viser residualene at regresjonsmodellen feilestimerer på prediksjonene sine på LEGO-sett med mange unike brikker. Dette betyr at forklaringsvariablene inneholder informasjon som ikke er tatt hensyn til. QQ-plottet viser store haler på begge sider som indikerer at dataen ikke er normalfordelt, noe som tilsier at dataen ikke egner seg for lineær regresjonsanalyse.

På bakgrunn av overnevnte resultater kan det til en viss grad påstås at det ikke foreligger tilstrekkelig med bevis for å forkaste H_0. Dette vil dog ikke medføre at H_0 er bevist.

7 Diskusjon

Både visuell analyse av datasettet og verdier i regresjonsmodellen viser at LEGO-sett for jenter har flere unike brikker enn LEGO-sett for gutter. Det er få indikasjoner på at regresjonsanalysen er unøyaktig grunnet de nåværende klassifiseringene av kjønn hos LEGO-settene, sammenlignet med det faktum at det er selve datasettet som er lite egnet for en regresjonsanalyse som vi kan se i (Figur 33 Residual- og QQ-plott.).

Datasettet har en skjevfordeling av datapunkter innenfor de ulike kjønnskategoriene. Det er nærmere dobbelt så mange datapunkter for gutte- og unisex-sett sammenlignet med jente-sett. I tillegg finnes det 18 gutte-sett og 21 unisex-sett med flere antall brikker totalt enn det største jente-settet. Dette kan medføre at regresjonsmodellen fungerer dårligere på LEGO-sett for jenter med et høyt antall totale brikker.
På en annen side, så faller de fleste datapunktene innenfor et begrenset intervall med hensyn til antall brikker totalt. Dette medfører at påvirkningen fra datapunktene med et høyt antall brikker totalt er begrenset sammenlignet med datapunktene innenfor "hovedintervallet" som ligger mellom [0, 500]. Med andre ord, blir modellen påvirket i mindre grad av de få settene med et høyt antall totale brikker sammenlignet med resten av settene.

En mulig forbedring av modellen kunne vært å dele inn hvert enkelt sett inn i en kjønns-kategori istedenfor å kun kategorisere basert på settets tema slik denne modellen gjør. Dette vil muligens gi en mer nøyaktig kjønnsfordeling og dermed en mer nøyaktig modell. En slik inndeling ville kunne øke antall LEGO-sett i kategoriene "gutt" og "jente" og minke antall sett i "unisex" kategorien. Siden temaer som appellerer til begge kjønn kunne vært delt opp ytterligere ved å se på hvert enkelt sett. En slik inndeling ville vært tidkrevende siden datasettet ikke inneholder beskrivende informasjon om hvert enkelt sett, og denne informasjonen måtte derfor vært innhentet for å muliggjøre en slik inndeling.

Et naturlig dilemma for regresjonsanalyser er om alle variabler som faktisk påvirker resultatet er tatt med. I denne analysen brukes to forklaringsvariabler «kjønn» og «antall totale brikker». Selv om dette trolig nok er tilfredsstillende til denne problemstillingen, kan det i videre forskning være aktuelt å inkludere flere variabler for å få mer komplett analyse. Prisen på hvert enkelt sett kunne vært en aktuell variabel i denne sammenhengen. Dersom jenter for eksempel hadde flere unike brikker fordi jente-sett generelt er dyrere enn gutte-sett ville «pris» vært en naturlig forklaringsvariabel å inkludere.

QQ-plottet (Figur 33 Residual- og QQ-plott.) til regresjonsmodellen har tunge haler både ved veldig små og veldig høy verdier. Dette tilsier at modellen er dårligere på å predikere antall unike brikker ved ytterpunktene av datamengden. I denne modellen er det brukt lineær regresjon, en mulig forbedring av modellen kunne vært og tatt i bruk en annen type regresjon istedenfor den lineære modellen.

8 Konklusjon

Med utgangspunkt i et datasett som ikke spesielt egner for lineær regresjonsanalyse, viser analysen at gutter ikke har flere unike brikker enn jenter og dermed er problemstillingen feil. Det viser seg heller at jenter har flere unike brikker enn gutter, og det med en statistisk signifikant margin. Til tross for at H_0 ikke forkastes, foreligger det altså ikke nok bevis for å kunne garantere at den er korrekt.

9 Referanser

Langaas, M. (2020, Oktober 19). IST[A/G/T]1003: Statistisk læring og data science. *Kompendium, Regresjon*. IMF/NTNU.

Løvås, G. G. (2018). *Statistikk for universiteter og høgskoler.* Oslo: Universitetsforlaget.

Sørmoen, I. H., & Aase, K. (2023, November 17). *Multippel lineær regresjon: Interaksjonseffekter.* Hentet fra math.ntnu.no: https://www.math.ntnu.no/emner/IST100x/ISTx1003/notat_interaksjoner.pdf

10 Vedlegg

10.1 LEGO-sett kjønnsfordeling

Architecture	Unisex
Batman	Gutt
BrickHeadz	Unisex
City	Unisex
Classic	Unisex
Creator 3-in-1	Unisex
Creator Expert	Unisex
DC	Gutt
Disney	Jente
DOTS	Unisex
DUPLO	Unisex
Friends	Jente
Harry Potter	Unisex
Hidden Side	Unisex
Ideas	Unisex
Jurassic World	Gutt
Juniors	Unisex
LEGO Art	Unisex
LEGO Brick Sketches	Unisex
LEGO Education	Unisex
LEGO Frozen 2	Jente
LEGO Super Mario	Unisex
Marvel	Gutt
Minecraft	Unisex
Minifigures	Unisex
Minions	Unisex
Monkie Kid	Gutt
NINJAGO	Gutt
Overwatch	Gutt
Powered UP	Gutt
Powerpuff Girls	Jente
Speed Champions	Gutt
Spider-Man	Gutt
Star Wars	Gutt
Stranger Things	Unisex
Technic	Gutt
THE LEGO MOVIE 2	Unisex
Trolls World Tour	Jente

Unikitty	Jente
Xtra	Unisex

10.2 Kildekode/GitLab repository

https://gitlab.stud.idi.ntnu.no/joachigw/istt1003-gruppe-11.git