Innledning og sammendrag

Denne rapporten er en besvarelse for prosjektoppgaven i emnet ISTT1003. Utgangspunktet til oppgaven var et datasett som bestod av informasjon tilknyttet legosett. Ut ifra dette, har gruppen formulert en problemstilling og utarbeidet teori og metode for å kunne besvare denne.

Rapporten består først av den underliggende teorien bak besvarelsen. Deretter presenterer vi problemstillingen og datasettet, og tar for oss pre-prosesseringen av dette, før vi går over på metodene og hypotesene vi har brukt for å kunne svare på problemstillingen. Til slutt drøfter vi resultatene og modellene, og hvorvidt disse tilstrekkelig besvarte problemstillingen.

Teori

I statistikk ønsker vi ofte å undersøke hvordan verdien til en rekke faktorer (forklaringsvariabler), kan "predikere" verdien til andre variabler (responsvariabel).

Lineær regresjon er en statistisk metode som blir brukt til å modellere forholdet mellom en responsvariabel og en eller flere forklaringsvariabler. Det finnes to hovedtyper av lineær regresjon; enkel lineær regresjon (ELR) og multippel lineær regresjon (MLR). I ELR brukes bare én forklaringsvariabel til å modellere verdien til responsvariabelen. Da blir formelen til modellen slik:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y er responsvariabelen, β_0 er skjærings punktet med y-aksen, β_1 er stigningstallet til linjen, x er forklaringsvariabel og ε er feilleddet (Mette Langaas. 2020). I MLR er det to eller flere forklaringsvariabler som er med i modellen. Da blir formelen slik:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

Når vi har ELR prøver vi å finne den beste rette linjen som beskriver dataen vår, mens med MLR prøver vi enten å finne det beste planet (2 forklaringsvariabler) eller det beste hyperplanet (3+ forklaringsvariabler) som beskriver dataen vår (Mette Langaas. 2020).

En interaksjonseffekt er en ny forklaringsvariabel som er produktet mellom to forklaringsvariabler. Praktisk sett gjør dette det mulig å la forklaringsvariabler påvirke hverandre. For eksempel, så kan man ha en interaksjonseffekt mellom en kategorisk variabel "beliggenhet" og en kontinuerlig variabel "areal". Dette passer hvis vi prøver å modellere prisen til leiligheter, siden øking av areal kan være mer

betydningsfullt i leiligheter med god beliggenhet sammenlignet med leiligheter som har dårlig beliggenhet (Ingeborg Hem Sørmoen, Kenneth Aase. “Multippel lineær regresjon: Interaksjonseffekter” Notat. N.d).

Problemstilling

I dette prosjektet hadde gruppen som ønske å undersøke eventuelle betydelige prisforskjellene mellom lego-produkter rettet mot forskjellige kjønn, noe som resulterte i den endelige problemstilling: “Er LEGO for gutter dyrere enn LEGO for jenter?”. For å svare på dette var det nødvendig ta i bruk multippel lineær regresjon med tre nye kategoriske variabler, hvor en av dem (kjønnskategori) blir definert og utdypet i følgende seksjon.

Pre-prosessering av data

For å svare problemstillingen, måtte forklaringsvariablene vurderes til hvilken grad de påvirket resultatene hver for seg. Problemstillingen stiller spørsmål til prisen på LEGO med hensyn til hvordan målgrupper (kjønn) produktene appellerer til. Temaet LEGO-settene tilhører er i datasettet en forklaringsvariabel som ble brukt til å strukturere datasettet i tre ulike kategorier av kjønn. Historisk sett har noen temaer i LEGO blitt utviklet og markedsført spesifikt mot én av de to klassiske kjønnsrollene basert på farger, innhold, reklamemateriale og formål. Dette ble studert og kollektivt diskutert innad i gruppen for alle temaer, med oppsøk på internett for å finne informasjon om de ulike typene. Den tredje kategorien “Kjønnsnøytralt” fikk LEGO-settene som ikke visste tydelige tegn til en spesifisert markedsstrategi mot en kjønnsrolle. Resultatet av dette var en rekke subjektive vurderinger for inndeling av kategoriene:

Gutte-dominert (11 totalt): NINJAGO, Star Wars, Spider-Man, Batman, Marvel, DC, Hidden Side, Speed Champions, Monkie Kid, City, Jurassic World

Jente-dominert (5 totalt): Friends, Unikitty, LEGO Frozen 2, Powerpuff Girls, Trolls World Tour

Kjønnsnøytralt (14 totalt): DUPLO, Disney, Harry Potter, THE LEGO MOVIE 2, Overwatch, Minecraft, Juniors, Minions, Stranger Things, Powered UP, Creator 3-in-1, Creator Expert, Ideas, Classic

Dette er listen over LEGO-temaene og respektivt til hvilken kategori den tilhører. Det er flest kjønnsnøytrale temaer som inneholder totalt 14 temaer, etterfulgt av gutte-dominert med 11, og jente-dominert med 5.

Modell og hypotese

Respons = pris, Forklaringsvariabler = kjønn, antall brikker, antall sider i manualen, unique pieces, minifigures

A = 'Price ~ Pieces'

B = 'Price ~ Pieces + Pages'

C1 = 'Price ~ Pieces + cat'

C2 = 'Price ~ Pieces*cat'

Målet med modellen var utviklet med formal om å finne svar på problemstillingen. Forklaringsvariablene ble selektert basert på korrelasjonsfaktoren individuelt per variabel, der «Pieces» og «Pages» var forklaringsvariablene med høyest korrelasjon til resulterende pris per enhet. Hypotesen vi vil teste, er om det er en korrelasjon mellom pris, og hvilket kjønn legosettene er markedsført til.

Modellevaluering og tilpasning

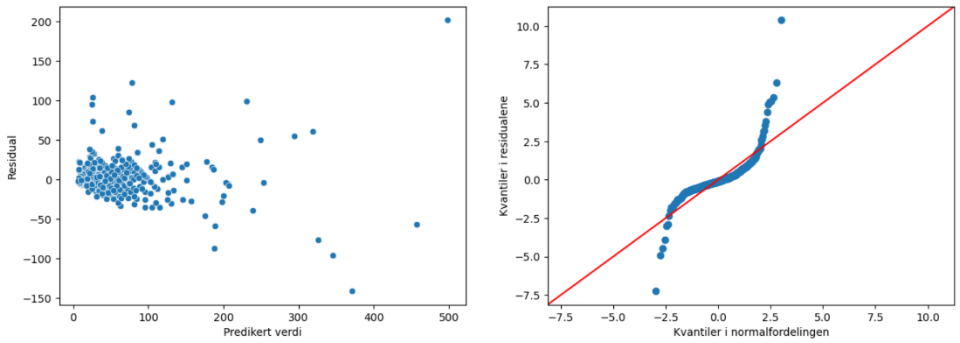
OLS Regression Results

Dep. Variable:	Price	**R-squared:**	0.867
Model:	OLS	**Adj. R-squared:**	0.866
Method:	Least Squares	**F-statistic:**	925.5
Date:	Mon, 13 Nov 2023	**Prob (F-statistic):**	1.27e-307
Time:	11:48:27	**Log-Likelihood:**	-3132.5
No. Observations:	714	**AIC:**	6277.
Df Residuals:	708	**BIC:**	6304.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.6251	1.288	3.592	0.000	2.097	7.153
Gender[T.jente]	-0.8768	3.053	-0.287	0.774	-6.871	5.118
Gender[T.nøytral]	12.6068	2.072	6.085	0.000	8.539	16.675
Pieces	0.1031	0.002	52.203	0.000	0.099	0.107
Pieces:Gender[T.jente]	-0.0049	0.008	-0.615	0.539	-0.021	0.011
Pieces:Gender[T.nøytral]	-0.0301	0.003	-11.242	0.000	-0.035	-0.025

Omnibus:	451.106	**Durbin-Watson:**	2.031
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	22517.946
Skew:	2.146	**Prob(JB):**	0.00
Kurtosis:	30.175	**Cond. No.**	3.47e+03

Figur 1: Regresjonsresultat modell C2



Figur 2: 1Evaluering av modell C2

Vi prøvde først ut en modell med mange variabler, men forklaringsvariablene gav lite endring til resultatene. Det ble dermed avklart å følge modellen med få forklaringsvariabler, da dette gav tilnærmet likt resultat, og videre er mer lesbart og introduserer mindre støy.

Diskusjon

For oss kunne problemstillingen ikke bli besvart for modellene som ble lagd, fordi QQ og residual plotter for alle tilpasninger av modellen gir inkonsistente resultater. Grunnen til dette er at modellen bare gir et gyldig resultat, dersom man kan anta at forklaringsvariablene er normalfordelte, noe som de ikke var. Gruppen vurderte fortløpende om det var noe som kunne endres med modellen, eller med dataene, men kom til slutt frem til at dette var utenfor hensikten til oppgaven, i tillegg til at forsøk på å "håndplukke" dataene til å være mer normalfordelte, vil resultere i datamanipulasjon, noe som ikke er forsvarlig.

Om dataene hadde vært normalfordelt, kunne resultatene gitt en konklusjon. Om et slikt utfall er antatt, kan man på figur.1 se at koeffisientene (to nederste) er veldig nærme null, som betyr at prisen sannsynligvis ikke er avhengig hvilke kjønnsgrupper som siktes for ulike temaer.

Gruppen fant ut at i datasettet er det noen legosett som bryter trenden, og har svært høy pris og antall deler enn resten av legosettene i datasettet. Dette er blant annet en av grunnene til den skjeve fordelingen i datasettet.

Konklusjon

"Er LEGO for gutter dyrere enn LEGO for jenter?"

Ja	Nei	Uklart
		X

Gruppen konkluderer med at det er usikkert at LEGO for gutter er dyrere en LEGO for jenter. Ved å utarbeide en modell og hypotese tilpasset problemstillingen, var målet å kunne svare på denne, men antagelsene som modellen krever for å kunne være gyldig, var altså ikke oppfylt ved nærmere undersøkelse. For å kunne svare videre på problemstillingen, vil det være nødvendig og finne en ny modell, men dette er utenfor hensikten til oppgaven. Gruppen mener at oppgaven har vært en tilstrekkelig og lærerik anvendelse av metoder og teori som er blitt undervist i faget.

Referanser

Mette Langaas. (2020). Regresjon. IST[A/G/T]1003: Statistisk læring og data science. NTNU

Appendix

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from scipy import stats
import statsmodels.formula.api as smf
import statsmodels.api as sm
df = pd.read_csv("lego.population.csv", sep = ",", encoding = "latin1")
# fjerner forklaringsvariabler vi ikke trenger
df2 = df[['Set_Name', 'Theme', 'Pieces', 'Price', 'Pages', 'Minifigures', 'Unique_Pieces']]
```

```
# fjerner observasjoner med manglende datapunkter

df2 = df2.dropna()


# gjør themes om til string og fjern alle tegn vi ikke vil ha med

df2['Theme'] = df2['Theme'].astype(str)

df2['Theme'] = df2['Theme'].str.replace(r'[^a-zA-Z0-9\s-]', '', regex = True)


# fjerner dollartegn og trademark-tegn fra datasettet

df2['Price'] = df2['Price'].str.replace('\$', '', regex = True)


# og gjør så prisen om til float

df2['Price'] = df2['Price'].astype(float)

# Gruppere temaer i nye grupper:

# (Harry Potter, NINJAGO og Star Wars havner i én gruppe, City og Friends i en annen, og alle andre i en tredje)

df2['Gender'] = np.where(df2['Theme'].isin(["NINJAGO", "Star Wars", "Spider-Man", "Batman", "Marvel", "DC", "Hidden Side", "Speed Champions", "Monkie Kid", "City", "Jurassic World"]), 'gutt',

      np.where(df2['Theme'].isin(["Friends", "Unikitty", "LEGO Frozen 2", "Powerpuff Girls", "Trolls World Tour"]), 'jente', 'nøytral'))

df2.groupby(['Gender']).size().reset_index(name = 'Count')

#regresjon

A = 'Price ~ Pieces'

B = 'Price ~ Pieces + Pages'

C1 = 'Price ~ Pieces + Gender'

C2 = 'Price ~ Pieces*Gender'


formel = C2


modell = smf.ols(formel, data = df2)

resultat = modell.fit()
```

```
resultat.summary()
# Plotte predikert verdi mot residual
figure, axis = plt.subplots(1, 2, figsize = (15, 5))
sns.scatterplot(x = modell.fit().fittedvalues, y = modell.fit().resid, ax = axis[0])
axis[0].set_ylabel("Residual")
axis[0].set_xlabel("Predikert verdi")

# Lage kvantil-kvantil-plott for residualene
sm.qqplot(modell.fit().resid, line = '45', fit = True, ax = axis[1])
axis[1].set_ylabel("Kvantiler i residualene")
axis[1].set_xlabel("Kvantiler i normalfordelingen")
plt.show()
```