LUNDS UNIVERSITET
Ekonomihögskolan

Hållbarhetsbetyget ESG - är spelreglerna desamma för alla?

En analys bland Europeiska industri- och dagligvaruproducenter med en explorativ ansats baserad på kvantitativa statistiska metoder

Victor Feilberg & Oscar Båth Viderström

Statistiska institutionen
Handledare: Peter Gustafsson
Kandidatuppsats i Statistik (15 hp)

Förord

Vi vill rikta ett oerhört stort tack till vår handledare Peter Gustafsson för både råd och vägledning under uppsatsens gång. Vidare vill vi även tacka den statistiska institutionen på Ekonomihögskolan vid Lunds Universitet, och i synnerhet Joel Danielsson för hans råd om Ridge- och Lasso-teknik.

Lund den 5:e januari 2023

Victor Feilberg & Oscar Båth Viderström

Abstract

The growing expansion of importance amid financial stakeholders regarding sustainability has led to an extended demand and need of reliable reporting amongst sustainability metrics. Previous research and information about ESG reporting and how ESG ratings are calculated are flawed. This thesis investigates if there are differences between how ESG ratings are assessed between companies producing daily goods and manufacturing companies in Europe. To answer this question, multiple linear regression has been used with ESG data from Bloomberg where models have been built for both industries using different techniques in order to detect differences in the model components. Ordinary Least Squares, Ridge and Lasso regression was used with different diagnostic tools to perform variable selection and determine the reliability of each model. This thesis concludes that there is a difference between the industries large manufacturing companies and daily goods producers in which variables that explains the ESG score.

Keywords: ESG, Regression**,** Lasso, Lund University, data visualization

Innehållsförteckning

1 Inledning

1.1 Bakgrund

ESG - *Environmental, Social and corporate Governance* är ett paraplybegrepp som handlar om hur man inom bolagsstyrning och investering behandlar frågor om miljö och socialt ansvar. Produktion och konsumtion är nära sammanlänkad med hur vår miljö påverkas där global uppvärmning, höjda havsnivåer, skogsavyttring och förlust av biologiskt mångfald är en del av klimatkrisen som råder. För att motarbeta detta hot är det av största vikt att bolag och investerare börjar göra medvetna val för att begränsa förstörelsen av vår miljö. Angående de sociala parametrarna ingår här frågor om orättvisor, diskriminering och skyldigheter gentemot sina medarbetare vad gäller arbetsvillkor. Både för privata och offentliga organisationer är det inom bolagsstyrning viktigt att ta hänsyn till dessa faktorer. Beslutsfattandeprocessen bör genomsyras av säkerställandet av att man är mån om och transparent angående miljömässiga och sociala överväganden. (Europeiska Kommissionen, 2022a)

Europeiska kommissionen (2019) har i sin taxonomiplan ställt upp en rad mål i sin strävan att bli den första klimatneutrala kontinenten. Bland annat vill man inför år 2050 att nettoutsläppen ska vara lika med noll, ekonomisk tillväxt ska vara frikopplad från resursanvändning och att inga människor eller platser ska vara kvarlämnade i denna process (Europeiska Kommissionen, 2019). En grön taxonomiförordning har införts av Europeiska kommissionen (2022b) vilket är ett klassificeringssystem av olika hållbara ekonomiska aktiviteter. Den blev först publicerad den 22:e juni 2020 och trädde i kraft den 12:e juli samma år. Förordningen fastställer ett ramverk för vilka förutsättningar en ekonomisk aktivitet behöver för att räknas som hållbar. Vidare består den av sex miljömässiga mål:

- Klimatförändringslindring
- Klimatförändringsanpassning
- Ett hållbart användande av och beskyddning av vatten och marina resurser
- Övergång till cirkulär ekonomi
- Föroreningskontroll och prevention
- Beskyddning och återskapande av biologisk mångfald och ekosystem

Sammanfattningsvis innebär detta att kraven för hur publika bolag och organisationer rapporterar stramas åt inom EU. Från och med januari 2023 är det hårda krav på att bolag måste redovisa en rad olika mått som behandlar punkterna ovan. Exempelvis hur mycket icke-förnybar energi man förbrukar, hur mycket svinn som förekommer, parallellt med könsfördelning. (Europeiska Kommissionen, 2022b)

1.2 Problematisering och frågeställning

Det råder tvetydigheter vad gäller hur man kalkylerar fram ESG-betyg med tanke på att det inte föreligger någon internationell oberoende myndighet som granskar dessa. Således kan trovärdigheten för betyget svikta vilket kan få investeraren att känna sig vilseledd vid ett investeringsbeslut. Det finns många avvägningar att göra där det är tämligen godtyckligt när man ska besluta om exempelvis hur tungt vissa parametrar ska vägas mot resterande, samt vilka som bör inkluderas för att få fram ett ESG-betyg. Hur ESG-betyg är kalkylerade är tämligen komplicerat att få information om och det finns flera olika ratinginstitut som alla har sina egna metode. Detta leder till frågor som; hur bör företag ställa sig till ett ESG-betyg? Vilka åtgärder kan man vidta för att få ett så högt betyg som möjligt? Vet man vilka åtgärder man bör ta? Är spelreglerna desamma för samtliga företag inom olika branscher eller skiljer de sig åt? Detta leder således till uppsatsfrågan:

> *Föreligger det någon skillnad i vilka variabler som förklarar ett företags ESG-betyg mellan branscherna dagligvaruproducenter och tillverkande industribolag?*

Med tanke på att ESG-betygsättningen är ett tämligen vitt fält utförs en kvantitativ studie med en explorativ ansats där stor vikt läggs vid kartläggningen av betygsstrukturen. Författarna har inte funnit någon liknande studie som tidigare utförts.

1.3 Avgränsning

Det finns flera olika ratinginstitut som har olika metoder för att kalkylera ESG-betyg. Valet föll på Bloomberg med tanke på dess tillförlitlighet, tillgänglighet och allmänna acceptans inom finansindustrin. Vidare har avgränsning till ett kombinerat dataset av börsnoterade tillverkande industribolag samt dagligvaruproducenter i EU gjorts. Datasetet är avgränsat från åren 2016 till 2021 för att kunna konstruera rättvisande imputationer, där modellframställningen är baserad på 2021 års data.

2 Data

Den data som kommer ligga till grund för undersökningen är hämtad från den finansiella databasen Bloomberg. Databasen Bloomberg (2022) är allmänt accepterad på den finansiella marknaden och ses som den primära informationskällan och analytiska verktyg globalt. För många företag är numera ESG ett viktigt beslutsunderlag, vilket enligt EU:s taxonomiplan (Europeiska kommissionen, 2022b) måste redovisas för från och med januari 2023. ESG innefattar många komplexa variabler vilka i ett sammanviktat mått leder till en betygsättning. Bloomberg (2022) har listat bolag i olika kategorier givet den redovisningstransparens som tillgodosetts utefter diverse parametrar. I denna studie kommer

uteslutande Bloombergs lista och redovisade resultat att nyttjas för den Europeiska marknaden av industribolag och dagligvaruproducenter. Bloomberg (2022) har listat 62 bolag som redovisar tillräcklig data för att kunna betygsätta bolagen. Till höger om varje variabel i kursivt skådas benämningen på variabeln som kodats om i R för att kunna utföra analysen. När datan sedermera delats upp i vardera bransch har beteckningarna *.ind* för tillverkande industribolag och *.mat* för dagligvaruproducenter använts. För att specificera vilket år datan är inhämtad från har även *.21* använts som beteckning. En stjärna (*) har satts till höger om de variabler som valts att inte inkluderas i analysen. Detta redogörs mer ingående för i metodavsnittet. Parametrarna som inkluderas och kommer att anses som förklarande variabler är enligt databasens betygsättning följande kategorier:

- Resurseffektivitet
- Klimatförändring
- Hälsa, säkerhet, miljö
- Arbetskraft
- Bolagsstyrning
- Ersättning

2.1 Resurseffektivitet

Vad gäller resurseffektivitet mäts hur väl bolagen utnyttjar resurser i relativitetsmått. Under paraplyet för resurseffektivitet ingår:

- Energiintensitet per försäljning (MWh per miljon) - *energy.int.per.sales*
- Avfallsintensitet per försäljning (ton per miljon) - *waste.int.per.sales*
- Procentuellt återvunnet avfall*

2.2 Klimatförändring

För att avgöra klimatförändringen enligt databasen används följande mått:

- Intensitet av växthusgasutsläpp per försäljning (metrisk ton per miljon) - *Emissions per sales*
- Procentuell användning av förnybar energi*
- Vattenintensitet per försäljning (kubikmeter per miljon) - *water.int.per.sales*
- Procentuellt återvunnet vatten*

2.3 Hälsa, säkerhet, miljö

Under denna kategori ingår följande mått:

- Nox-utsläpp per försäljning (ton per miljon)*
- Soxutsläpp per försäljning (ton per miljon)*
- Antal incidenter med förlorad tid*
- Totalt antal händelser som kan registreras*

2.4 Arbetskraft

För arbetskraften används relativitetsmått för prestation i relation till antal anställda och fackligt anslutna enligt nedan:

- Omsättning per anställd - *turnover.per.emp*
- Personalkostnad per anställd - *personnel.cost.per.emp*
- EBITDA per anställd - *ebitda.per.emp*
- Nettoresultat per anställd - *net.income.per.emp*
- Procentuell personalomsättning*
- Procentuell anställda som är fackligt anslutna*

2.5 Bolagsstyrning

Vad gäller bolagsstyrningen används nedanstående relativitetsmått för typen av bolagsstyrning:

- Procentuell oberoende styrelseledamöter - *independent*
- Procentuell icke verkställande direktörer - *non.ex.board*
- Procentuell närvaro vid styrelsemöten - *board*
- Procentuellt antal kvinnor i styrelsen - *women.board*
- Procentuellt antal kvinnliga chefer - *women.ex*
- Gemenskapens utgifter i procent av vinsten före skatt*

2.6 Ersättning

En del av ESG-måttet är hur ersättningsstrukturen ser ut inom bolagen, enligt nedan följer hur Bloomberg valt att kvantifiera det:

- Total VD-ersättning i procent av den totala ersättningen till ledande befattningshavare - tot.*ceo.comp.as.tot.ex*
- Ledningens lön + bonus i procent av den totala ersättningen till ledningen - *ex.salary.bonus.tot.comp*
- VD:ns all annan ersättning i % av total ersättning - *ex.salary.other.tot.comp*
- VD-lön + bonus i procent av den totala VD-ersättningen - *ceo.salary*
- VD Alla andra ersättningar i % av den totala VD-ersättningen - *ceo.comp*

2.7 ESG-betyg

Ett viktat mått utefter ovanstående parametrar resulterar i ett betyg mellan 0-100, där 0 är det svagaste resultatet och 100 det starkaste (Bloomberg, 2022).

3 Teori och statistiska metoder

I följande avsnitt presenteras samtliga teorier som ligger till grund för de statistiska metoder som används i uppsatsen.

3.1 Imputation

Imputationsteknik syftar till att hantera de icke redovisade datavärdena för diverse variabler samt för att få ett fullständigt material att arbete med. Imputation är processen att ersätta värden som saknas, är ofullständiga eller har blivit felaktiga på något vis. Dessa kan skattas med hjälp av olika imputationstekniker. (SCB, 2017)

3.1.1 Enkel linjär regressions-imputation

I fall där man kan utläsa en trend i hur datan förändrats under en viss period kan man använda sig utav en imputationsteknik via enkel linjär regression:

$$Y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x$$

Anledningen till detta är att tekniken kan ge en liten felmarginal om det predikterade värdet i realiteten följer trenden från år till år. (Subrahmanya, 2018)

3.1.2 Cold deck imputation

När man inte kan utläsa någon trend och det bara finns några få icke redovisade datapunkter kan man använda sig utav *cold deck imputation*. Tekniken innebär att man kopierar data från året innan eller efter. (OECD, 2013)

3.1.3 Mean imputation

Mean Imputation är en imputationsteknik där man tar medelvärdet av de värden som inte saknas för en variabel (Wicklin, 2017).

3.1.4 Problematik med imputation

Om man jobbar med ofullständig data där man tvingas göra imputationer kan detta leda till problem. Dels kan det leda till en hög bias när man genomför analysen samt att variansen kan bli hög. (Mittag, 2013)

3.2 Multipel linjär regression

Multipel linjär regression är enligt Sheather (2009) en teknik där man använder sig av flera prediktorvariabler för att förklara en responsvariabel. Enligt honom ser väntevärdesfunktionen för en sådan modell ut som:

$$E(Y|X_1 = x_1\,,\, X_2 = x_2\,,\, \dots\,,\, X_p = x_p) \;=\; \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \dots + \boldsymbol{\beta}_p x_p$$

Följaktligen ser den multipla linjära regressionsmodellen ut som:

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{1i} + \boldsymbol{\beta}_2 x_{2i} + \dots + \boldsymbol{\beta}_p x_{pi} + e_i$$

där e_i är ett slumpmässigt fel med väntevärde 0 betingat på slumpvariabel **X**:

$$E(e_i|x) \;=\; 0$$

(S.J. Sheather, 2009)

Denna ansats går även enligt Sheather (2009) att formulera med hjälp av vektorer och matriser där man låter Y vara en $(n \times 1)$ vektor, X vara en $(n \times (p+1))$ matris, $\boldsymbol{\beta}$ vara en $((p+1) \times 1)$ vektor av de okända parametervärdena för varje prediktorvariabel och **e** vara en $(n \times 1)$ vektor för slumpfelen:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ \dots & \dots & & \\ 1 & X_{n1} & \dots & X_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \dots \\ \boldsymbol{\beta}_p \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix}$$

Med hjälp av dessa menar han att ekvationen för multipel linjär regression kan skrivas med matrisansats genom:

$$Y = X\boldsymbol{\beta} + e$$

Om man vidare enligt honom låter x_i beteckna den i:te raden av matris X så att:

$$x_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})$$

är en $(1 \times (p + 1))$ vektor så att väntevärdet av slumpvariabel Y betingat på slumpvariabel X = x kan skrivas som:

$$E(Y_i|X = x) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \dots + \boldsymbol{\beta}_p x_p = x_i\boldsymbol{\beta}$$

Residualkvadratsumman kan även med stöd av Sheather (2009) skrivas i matrisform:

$$RSS(\boldsymbol{\beta}) = (Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta})$$

Vidare menar Sheather (2008) att $(AB)' = B'A'$ samt $B'A = A'B$ om A och B är $(n \times 1)$ vektorer. När resultatet är (1×1) kan man genom expandering av föregående ekvation visa att:

$$RSS(\boldsymbol{\beta}) = Y'Y + (X\boldsymbol{\beta})'X\boldsymbol{\beta} - Y'X\boldsymbol{\beta} - (X\boldsymbol{\beta})'Y = Y'Y + \boldsymbol{\beta}'(X'X)\boldsymbol{\beta} - 2Y'X\boldsymbol{\beta}$$

Härnäst differentieras enligt Sheather (2009) den senaste ekvationen med avseende på $\boldsymbol{\beta}$ för att hitta minska kvadrat skattningarna. Sätts detta lika med noll och de två gemensamma på var sida tar ut varandra menar han att följande ekvation fås:

$$(X'X)\boldsymbol{\beta} = X'Y$$

Om antagandet görs att inversen till $(X'X)$ existerar menar han att man kan visa att minsta kvadrat skattning av $\widehat{\boldsymbol{\beta}}$ ges av ekvationen:

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$$

I enlighet med Sheather (2009) ges det predikterade värdet $\widehat{Y}$ därmed av:

$$\widehat{Y} = X\widehat{\boldsymbol{\beta}}$$

och residualerna från:

$$\widehat{e} = Y - Y' = Y - X\widehat{\boldsymbol{\beta}}.$$

För att räkna ut R^2 som är den totala variationen i Y förklarad av regressionsmodellen tar man:

$$R^2 = \frac{SSreg}{SST} = 1 - \frac{RSS}{SST}$$

och för R^2_{adj} tar man:

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

där SSreg, RSS och SST, som är regressionskvadratsumman, residualkvadratsumman och den totala kvadratsumman av varianserna i modellen fås av:

SSreg	$\sum_i^n (\widehat{y}_i - \bar{y})^2$
RSS	$\sum_i^n (y_i - \widehat{y}_i)^2$
SST	$SSreg + RSS$

Tabell 1: Formell för SSreg, RSS och SST

(S.J. Sheather, 2009).

3.3 Modellvalidering och variabelselektion

För att passa en så givande modell som möjligt bör man med hjälp av en rad diagnostiska verktyg utföra variabelselektion. Det är svårt att definiera vilken som är den bästa modellen, men det råder konsensus inom den akademiska statistiska sfären om vad som bör anses vara bra. En rad antaganden bör uppfyllas och om så är fallet kan modellen betraktas vara tillförlitlig. Dessa antaganden kommer redogöras för i efterföljande avsnitt. (Sheather, 2009)

3.3.1 Antaganden

För att bygga en multipel linjär regressionsmodell som beskriver verkligheten på ett rimligt vis är det viktigt att en rad antaganden är uppfyllda (The Pennsylvania State University, 2018). Ett intuitivt sätt att kolla ifall modellen går att lita på är att visuellt kontrollera antaganden genom grafisk analys med hjälp av plottar.

Den första plotten kallas *Residuals vs Fitted*. I denna låter man residualerna vara på y-axeln och det passade värdet på x-axeln i ett spridningsdiagram. Vad man utläser i denna plott är ifall antagandet om linjäritet och homoskedasticitet uppfylls, vilket är om residualerna ser horisontella ut kring ett värde på y-axeln för att se om variansen av feltermerna e_i är konstant och oberoende av nivån på x eller y. (The Pennsylvania State University, 2018)

Den andra plotten kallas *Normal QQ*. Här låts två uppsättningar kvantiler från stickprovet samt den teoretiska vara på x- och y-axeln i ett spridningsdiagram. Detta görs i syfte att se om datamaterialet ser ut att komma från samma fördelning, i det här fallet normalfördelningen. Om de gör det ska en någorlunda rak diagonal linje formas. (Ford, 2015)

Den tredje plotten kallas *Scale-location*. På y-axeln låts roten ur de standardiserade residualerna vara med dess passade värden på x-axeln. Detta är ett verktyg för att kolla så att residualerna över prediktorvariablerna har ungefär samma spridning där det inte ska finnas någon tydlig trend att utläsa, det vill säga antagandet om linjäritet och homoskedasticitet. (Kim, 2015)

Det fjärde och sista diagnostiska verktyget är den så kallade *Residuals vs Leverage* plotten. Denna används för att upptäcka om det finns kraftigt inflytelserika outliers i datamaterialet. För att avgöra detta kollar man uppe i det högra samt vänstra hörnet för att se om någon datapunkt befinner sig innanför *Cook's distance.* (Kim, 2015)

3.3.2 AIC

I den vanligaste definitionen av Akaikes Informationskriterium (*AIC*) vill man åstadkomma en balans mellan att straffa modellers komplexitet samtidigt som man mäter hur väl en modell är passad. Desto lägre värde på AIC, desto bättre är modellen. Ett mått på hur välanpassad modellen är; desto mindre värde på AIC desto bättre är modellen. Man tar två gånger den negativa log-likelihoodfunktionen av den passade modellen så att:

$$AIC = 2\left[- log(L(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1, ..., \widehat{\boldsymbol{\beta}}_p, \widehat{\boldsymbol{\sigma}}^2 | Y) + K \right]$$

där $K = p + 2$ är ett mått på komplexitet där p är antalet variabler. (S.J. Sheather, 2009)

3.3.3 MSPE

MSPE - Mean Square Prediction Error, eller medelkvadratfelet, är ett mått där man tar väntevärdet av medelkvadratfelet mellan ett antal observationer som man bygger en modell med och jämför dessa med ett antal observationer som kommer från samma dataset:

$$MSPE(L) = E[(g(x_i) - \widehat{g}(x_i))^2].$$

(James, Witten, Hastie & Tibshirani, 2021)

3.3.4 Korsvalidering med Monte-Carlo simulation

Ett tillvägagångssätt för att göra goda skattningar av MSPE är att utföra korsvalidering. En teknik som kallas för *k-fold cross-validation* går ut på att man delar in det ursprungliga datamaterialet i k lika stora delar där en delmängd blir träningsdata som man bygger modellen med och testar detta mot testdatan för att se hur väl den tränade datan passar testdatan. (Picard & Book, 1984)

Monte-Carlo simulation kan utföras genom upprepade slumpmässiga urval för att få fram ett numeriskt resultat (Kroese, Brereton, Taimre & Botev, 2014). Vid korsvalidering kommer tränings- respektive testdatan utgöras av olika observationer för varje simulering då dessa väljs slumpmässigt (Xu & Liang, 2011). Därför blir det rimligt att med hjälp av en Monte-Carlo simulering köra korsvalideringen flera gånger för att få ett mer robust resultat. (Xu & Liang, 2011)

3.3.5 Transformationer

Transformationer genomförs ifall problem uppstår med exempelvis icke-konstant varians och olinjäritet i modellen (Sheather, 2009). *Logaritmiska tranformationer* kan vara användbart när man stöter på variabler som är skevt fördelade, till exempel om intervallet är brett (West, 2021).

För vissa typer av data kan det vara av värde att koda om variabler till binära (0 och 1) koder, så kallade *dummyvariabler*. Man ger observationer en etta då den uppfyller ett slags kriterium och en nolla om den inte uppfyller det angivna kriteriet. (Hardy, 1993)

3.3.6 Multikollinearitet

När det förekommer stark korrelation mellan flera prediktorvariabler kan man få problem med *multikollinearitet* (Sheather, 2009). För att upptäcka multikollinearitet kan man göra detta grafiskt genom att visualisera en matris av spridningsdiagram mellan variablerna (Sheather, 2009).

Korrelationen kan även räknas ut numeriskt och anta ett värde mellan -1 och 1 för två variabler x och y med hjälp av:

$$r_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \Sigma(y-\bar{y})^2}}$$

(Körner & Wahlgren, 2016)

Multikollinearitet kan leda till att den statistiska signifikansen hos en variabel som korrelerar med en eller flera i modellen blir försvagad vilket medför att det blir svårare att göra statistisk inferens. Det kan även innebära att standardavvikelsen blir hög och att koefficienterna kan variera mellan olika stickprov. (Allen, 1997)

Masseffekt handlar i multipel linjär regression om hur en enskild prediktorvariabel kan kvantifieras i form av dess koefficient. Koefficienten representerar den genomsnittliga förändringen i responsvariabeln när prediktorvariabeln förändras med en enhet, givet att alla andra prediktorvariabler hålls konstanta. (James et. al, 2021)

3.3.7 Variance Inflation Factor

Ett givande diagnostiskt verktyg för att upptäcka multikollinearitet kallas för Variance Inflation Factor (*VIF*) vilken man får genom:

$$\frac{1}{1-r_{xy}^2}$$

från r_{xy} i föregående ekvation (Sheather, 2009).

O'brien (2007) menar att det finns en enighet i den akademiska sfären där konsensus är att ett VIF värde på över 10 bör betraktas som ett tecken på att det råder kraftig multikollinearitet och att någon åtgärd måste vidtas. Vidare argumenterar han för att man kan ta det med viss reservation då åtgärderna som vidtas kan innebära att större problem skapas än de som eventuellt finns med multikollinearitet.

3.3.8 Marginal Model Plot

Marginal Model Plot är ytterligare ett verktyg för att avgöra om man har en väl passad modell. Ponera en modell som ser ut som:

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{1i} + \boldsymbol{\beta}_2 x_{2i} + \ldots + \boldsymbol{\beta}_p x_{pi} + e_i \qquad \text{(A)}$$

med flera prediktorvariabler (Sheather. 2008). För att visuellt bedöma om modellen ovan är väntevärdesriktig:

$$E_A(Y|x_i)$$

så vill man jämföra den med den icke-parametriska regressionsmodellen:

$$Y = f(x_1, x_2, \ldots, x_p) + e_i \qquad \text{(B)}$$

I modell B skattas väntevärdet av:

$$E_B(Y|x_i)$$

genom att lägga till en icke-parametrisk passform där Y plottas mot x_i. Poängen med detta är att man vill se ifall skattningen $E_A(Y|x_i)$ ligger nära skattningen av $E_B(Y|x_i)$. (Sheather, 2009)

3.3.9 Added Variable Plot

För att avgöra vilken påverkan en enskild prediktorvariabel har på responsvariabeln kan man enligt Sheather (2009) göra detta visuellt med ett diagnostiskt verktyg kallat *Added Variable Plot*. Genom att kolla på denna kan man med stöd av honom konstatera effekten en prediktorvariabel har justerat för effekten av de andra prediktorvariablerna. Om man går tillbaka till den ursprungliga matrisansatsen för multipel linjär regression där:

$$Y = X\boldsymbol{\beta} + e$$

vill nu ytterligare en prediktorvariabel Z introduceras till modellen så att:

$$Y = X\boldsymbol{\beta} + Za + e$$

där i synnerhet a koefficienten är intressant för att se hur stor påverkan Z har på Y när denna justerats för effekten X har på Y. Plotten framställs genom att plotta residualerna från den ursprungliga modellen mot residualerna:

$$Z = X\boldsymbol{\delta} + e$$

(Sheather, 2009)

3.4 Regularisering

När man jobbar med komplex data där det finns många observationer och/eller variabler kan det vid tillfällen vara fördelaktigt att förenkla resultatet vid komplicerade problem samt när det finns risk för *overfitting* (Bühlmann & Van De Geer, 2011). Overfitting är att man har en modell som efterliknar datan för nära och kan därför vara olämplig när man ska passa ytterligare data och/eller prediktera framtida värden på ett tillförlitligt sätt (Porta, 2016).

3.4.1 Lasso

Lasso; least absolute shrinkage and selection operator, är en metod som på ett effektivt sätt skattar regressionskoefficienter samtidigt som variabelselektion utförs (Sheather, 2009). Om man utgår från en vanlig minsta-kvadrat regressionsmodell:

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{1i} + \boldsymbol{\beta}_2 x_{2i} + \ldots + \boldsymbol{\beta}_p x_{pi} + e_i$$

så beräknas Lasso enligt Sheather (2009) genom följande begränsade framställning av minsta-kvadratmetoden:

$$min \sum_{i=1}^{n} (y_i - \{\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{1i} + \boldsymbol{\beta}_2 x_{2i} + \ldots + \boldsymbol{\beta}_p x_{pi}\})^2 \text{under begränsningen att } \sum_{j=1}^{p} \left|\boldsymbol{\beta}_j\right| \leq s$$

där $s \geq 0$. Med hjälp av ett lagrangemultiplikatorargument kan man med stöd av honom sedan visa att föregående ekvation är ekvivalent med minimeringen av residualkvadratsumman där man inför en straffterm på absolutvärdet av regressionskoefficienterna, så att:

$$min \sum_{i=1}^{n} (y_i - \{\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{1i} + \boldsymbol{\beta}_2 x_{2i} + \ldots + \boldsymbol{\beta}_p x_{pi}\})^2 + \lambda \sum_{j=1}^{p} \left|\boldsymbol{\beta}_j\right|$$

För något $\lambda \geq 0$. (Sheather, 2009)

3.4.2 Ridge Regression

Till skillnad från minsta-kvadrat regression är *Ridge Regression* enligt Taboga (2021) en metod där variablerna istället är skattade så att den har mindre varians än en minsta-kvadrat skattning med något bias. Medelkvadratfelet hos en skattning med Ridge (summan av variansen och bias i kvadrat) är enligt honom i många fall mindre än hos en minsta-kvadrat skattning. Skattningen av $\widehat{\boldsymbol{\beta}_{\lambda}}$ angriper minimeringsproblemet lite annorlunda än vad minsta-kvadrat skattaren gör där:

$$\widehat{\boldsymbol{\beta}_{\lambda}} = arg\,min_{b} \sum_{i=1}^{N} (y_i - x_i b)^2 + \lambda \sum_{k=1}^{K} b_k^2$$

för $\lambda > 0$ (Taboga, 2021). Man har alltså adderat en straffterm till minsta-kvadratkriteriet i form av en kvadrerad vektornorm:

$$||b||^2 = \sum_{k=1}^{K} b_k^2$$

till den minimerade residualkvadratsumman:

$$SSR = argmin_{x}\,(SSR) \sum_{i=1}^{N} (y_i - x_i b)^2$$

där man väljer storleken på λ utefter hur mycket koefficienterna ska straffas (Taboga, 2021). Viktigt att tänka på här är att variablerna bör standardiseras så de är på samma skala (Taboga, 2021). Ridge används lämpligen i fall där det förekommer hög korrelation mellan de oberoende variablerna (Hilt & Seegrist, 1977).

4 Metod

Uppsatsens metod bygger på både en top-down ansats och en best-subset metod. För modellframställningen av de egenkonstruerande modellerna har en top-down ansats applicerats varpå för Ridge- och Lasso-teknikerna har best-subset tillämpats.

4.1 Datahantering

Som tidigare nämnts består datan av en rad olika oberoende variabler för år 2016-2021. Datan som är hämtade från Bloomberg har ej varit fullständig vilket ansatte imputation för att få en så korrekt bild av datamaterialet som möjligt. För de förklarande variabler som valdes bort hade dessa mindre än hälften redovisade siffror inom respektive bransch. Nedan förklaras de imputationstekniker som använts utefter datans struktur.

4.1.2 Enkel linjär regressions-imputation

I de fall en trend kunnat utläsas i hur datan förändrats i perioden mellan 2016-2021 användes en imputationsteknik via enkel linjär regression.

4.1.2 Cold deck imputation

När ingen trend kunnat utläsas och enbart några få tomma dataceller för ett och samma bolag skådats där datan inte förändrats speciellt mycket har cold deck imputation använts.

4.1.3 Mean imputation

Denna teknik har använts när det skådats stora variationer i datan och inget mönster kunnat utläsas. För vissa bolag som inte redovisat data för somliga av variablerna har ett industri-medelvärde tagits för varje år.

4.1.4 Dummyvariabler

Med tanke på att det endast fanns tillräckligt med data för två av fyra oberoende variabler inom kategorin klimatförändring samt två av tre inom kategorin resurseffektivitet skapades två extra dummyvariabler med kriterierna; rapporterade och icke rapporterade, för att se vilka bolag som redovisade datan för följande variabler:

- Rapportering av vattenintensitet per försäljning (kubikmeter per miljon) - *reported.-water.int.per.sales*
- Rapportering av avfallsintensitet per försäljning (ton per miljon) - *waste.int.per.sales*

4.2 Regressionsanalys

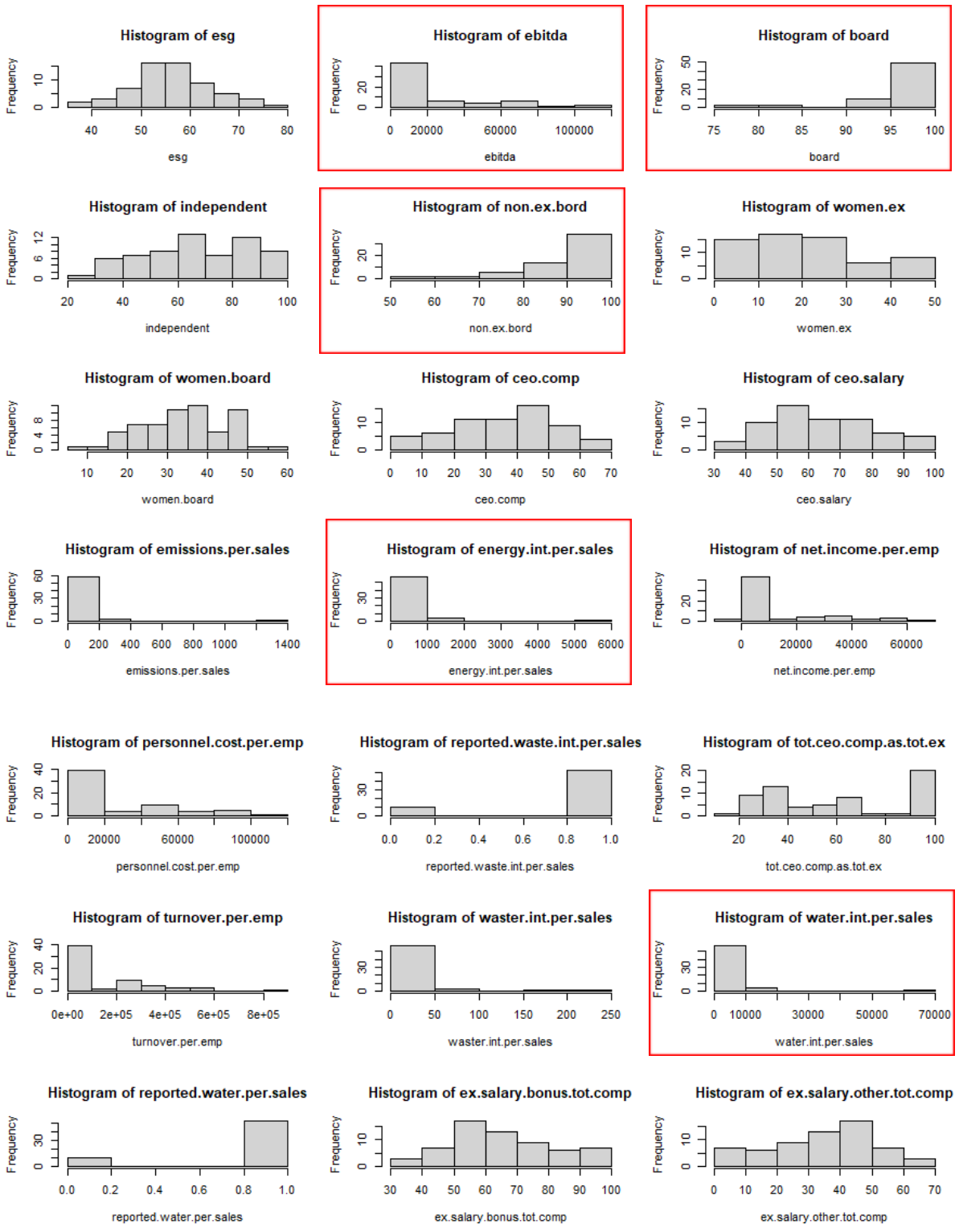
Under detta avsnitt kommer det redogöras för hur analysen är uppbyggd.

4.2.1 Variabeltransformation

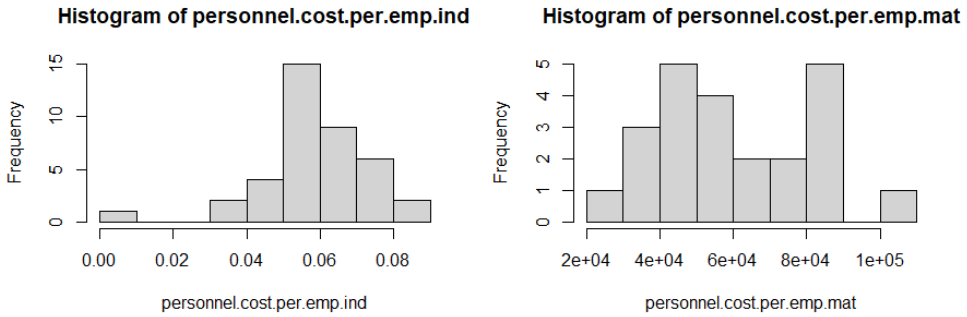
I figur 1 & 2 nedan avläses en fullständig bild av datamaterialets beroende variabel (*esg*) och de övriga oberoende variablerna i form av histogram. En relativt ojämn spridning observeras i majoriteten av de oberoende variablerna för det gemensamma datasetet. Figur 2 illustrerar ett exempel på skillnaden i den oberoende variabeln *personnel.cost.per.emp* uppdelad i ett nytt dataset för endast tillverkande industribolag och ett för dagligvaruproducenter. De två nya oberoende variablerna för respektive dataset har således blivit *personnel.cost.per.emp.ind* och *personnel.cost.per.emp.mat*.

Figur 2 åskådliggör skillnaden i spridning för respektive dataset vilket talar för att en uppdelning av dataseten bör göras för att förbättra variansen och linjäriteten i kommande regressionsmodeller.

De variabler som har genomgått en variabeltransformation är *ebitda.per.emp, board, non.ex.board, waste* och *water* (inringade i rött i figur 1 nedan) för respektive datamaterial.



Figur 1: Histogram över förklarande variabler för dagligvaruproducenter och tillverkande industribolag efter transformation



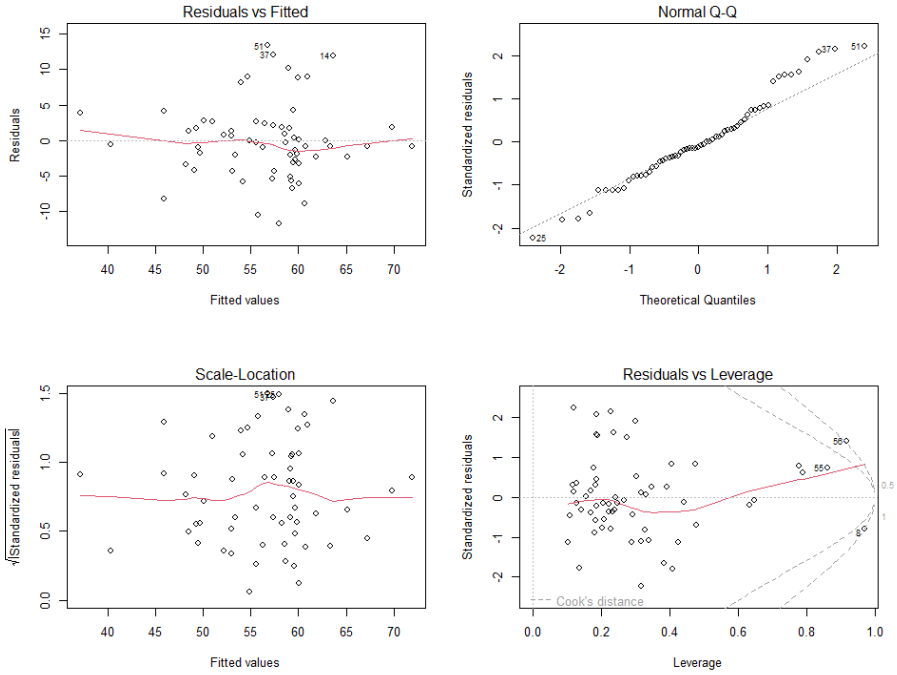
Figur 2: Histogram för variabeln personalkostnad per anställd för respektive dataset; livsmedelsproducent och tillverkande industribolag

4.2.2 Modellframställning

Vid modellframställning för en multipel linjär regressionsanalys eftersträvar man alltid en hög förklaringsgrad, lågt MSPE och signifikanta variabler. Kraven för linjäritet, homoskedacitet, normalfördelade residualer, observationer inom Cook's distance och icke korrelerade variabler måste vara uppfyllda. Inledningsvis har en modell byggts med det sammansatta datamaterialet dagligvaruproducenter och tillverkande industribolag för att se hur ovanstående krav och förklaringsgrad presenteras. Modellframställningen kommer att grundas på en top-down princip.

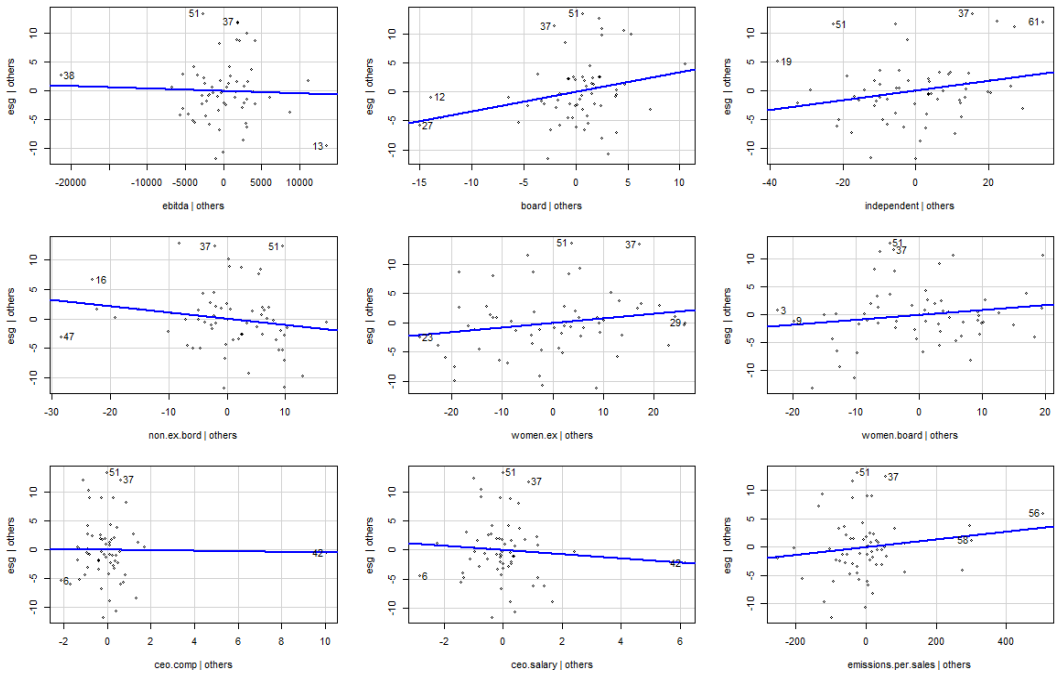
4.2.2.1 Modell 1 - det sammansatta datamaterialet dagligvaruproducenter och tillverkande industribolag

I plotten Residuals vs Fitted nedan åskådliggörs i figur 3 den första plotten för modellvalidering ett relativt linjärt samband. I den andra plotten för Normal QQ ser dessutom residualerna någorlunda normalfördelade ut med ett par avstickare i den högre kvartilen. Scale-Location plotten visar ett linjärt samband som påvisar homoskedasticitet, det vill säga att samtliga observationer har samma inbördes varians. Slutligen i den sista grafen för Residuals vs Leverage observeras två observationer med kraftigt inflytande på modellen; observation 8 och 55. På grund av de få observationer i datamaterialet görs bedömningen att analysen blir starkare av att ha kvar dessa.

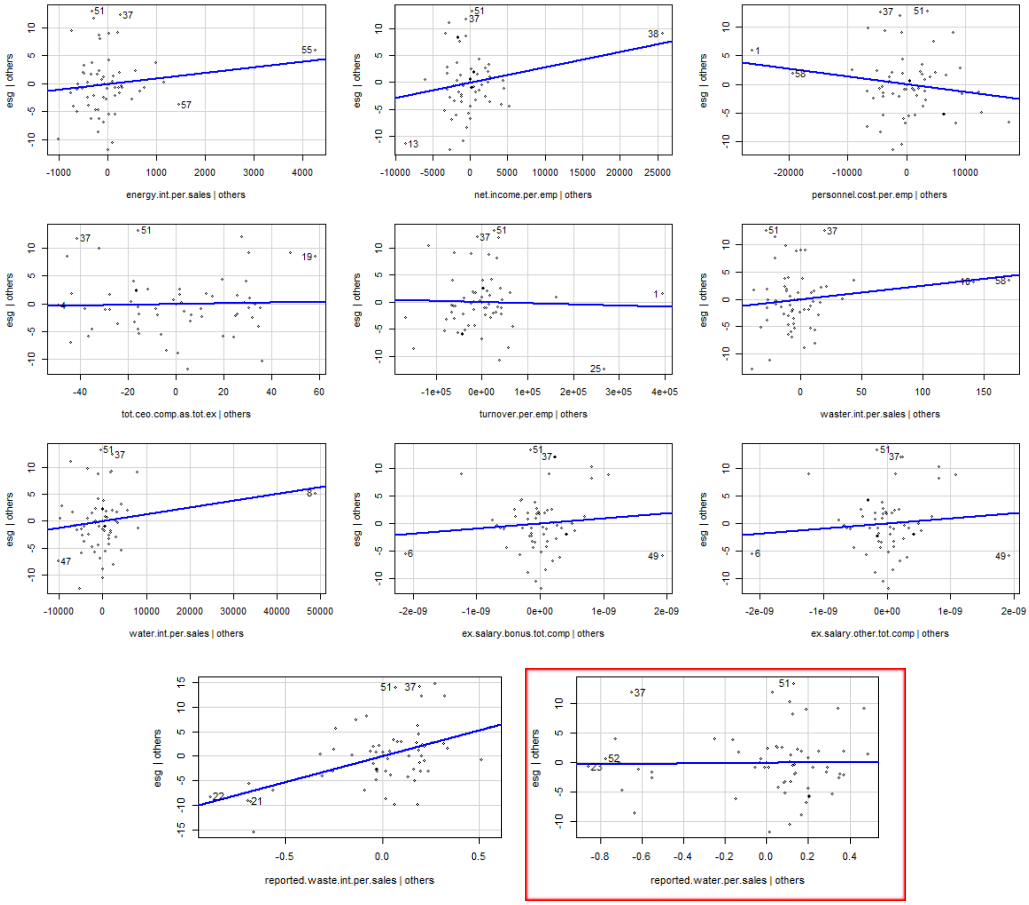


Figur 3: Residualdiagram över modellen för datamaterialen dagligvaruproducenter och tillverkande industribolag efter transformation

I figur 4 och 5 nedan illustreras Added Variable Plot som visar för respektive oberoende variabel dess bidragande information till modellen, det vill säga med högre eller lägre lutning beroende på hur mycket information som adderas till modellen. Värt att notera är hur variabeln *reported.water.per.sales* visuellt visar på ett relativt horisontellt samband, det vill säga ingen påverkan.

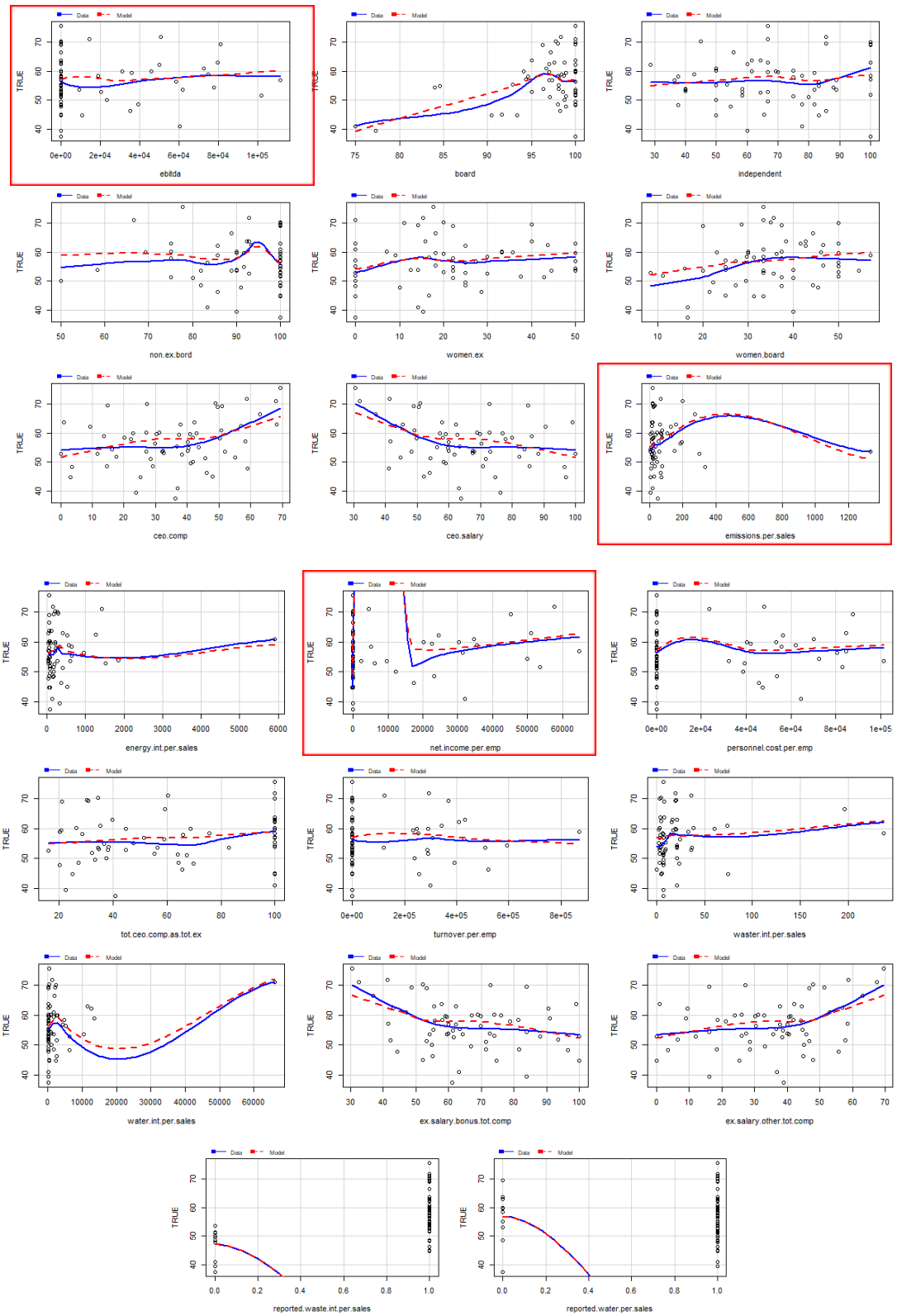


Figur 4: Added Variable Plot för samtliga variabler i datasetet för dagligvaruproducenter och tillverkande industribolag



Figur 5: Added Variable Plot för samtliga variabler i datasetet för dagligvaruproducenter och tillverkande industribolag

För figur 6 på nästa sida visualiseras Marginal Model Plots vilka beskriver hur väl modellen förklarar datan för respektive oberoende variabel. I plottarna går det att avläsa hur vissa observationer skiljer sig stark beroende på vilket dataset de kommer från. För att exemplifiera kan det observeras i plotten för *net.income.per.emp* hur det bildas ett kluster av data nära nollan på x-axeln och ett annat kluster en bit ifrån nollan. Likaså kan detta observeras för variablerna *ebitda.per.emp* och *emission.per.sales*. På grund av detta samt att det under variabeltransformation gick att åskådliggöra skillnader i spridning mellan de olika dataseten kommer nästkommande modeller skapas för respektive dataset i förhoppning om att nå en högre förklaringsgrad, fler signifikanta oberoende variabler och en bättre prediktionsförmåga.



Figur 6: Marginal Model Plots för modell 1 med båda dataseten

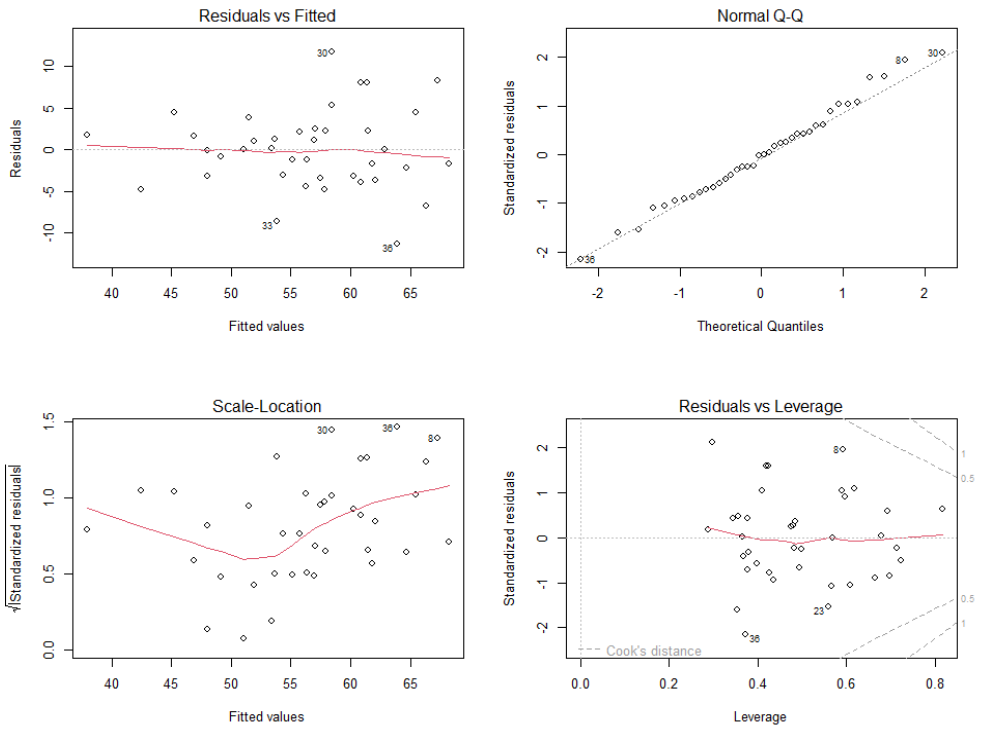
I den första modellen för datasetet med både dagligvaruproducenter och tillverkande industribolag kan det i tabellen nedan observeras en till synes hög förklaringsgrad R^2 om 0,5355 och förhållandevis låg R^2_{adj} om 0,3574 med endast en signifikant koefficient (*reported.waste.int.per.sales*). Modellen i sin helhet är signifikant med ett p-värde om 0,0022.

Modell	R^2	R^2_{adj}	p-värde	Antal signifikanta koefficienter
Modell 1 med båda dataseten för tillverkande industribolag och dagligvaruproducenter ihopslagna	0,5366	0,3574	0,0022	1 (reported.waste.int.per.sales)

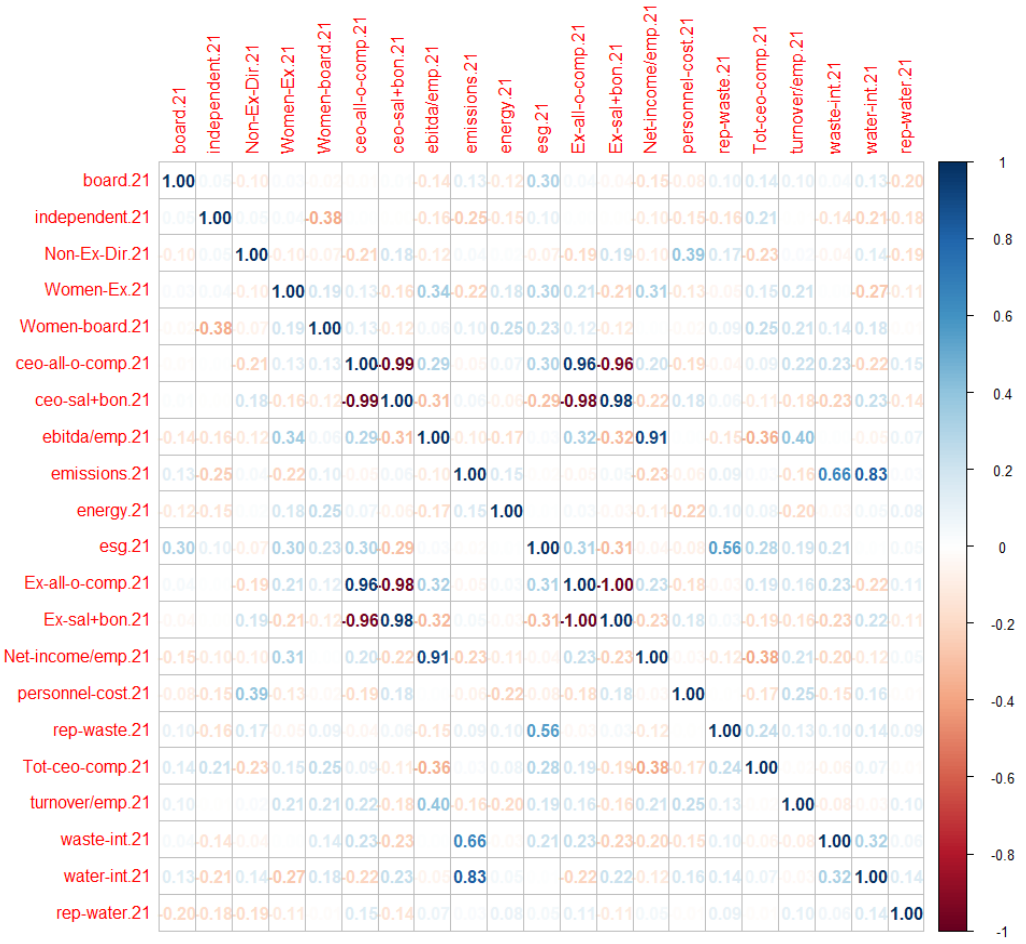
Tabell 2: Sammanställning av resultat för modell 1

4.2.2.2 Modell 2 - tillverkande industribolag

I modell 2 har datasetet innehållande både dagligvaruproducenter och tillverkande industribolag delats upp till respektive dataset vari denna modell endast innehåller observationer (bolag) som är kategoriserade som tillverkande industribolag. Figur 7 illustrerar; god linjäritet, normalfördelade residualer, relativ homoskedacitet och inga observationer som har för starkt inflytande enligt Cook's distance. Dock föreligger det en del multikollinearitet vilket avläses i korrelationsmatrisen för figur 8.

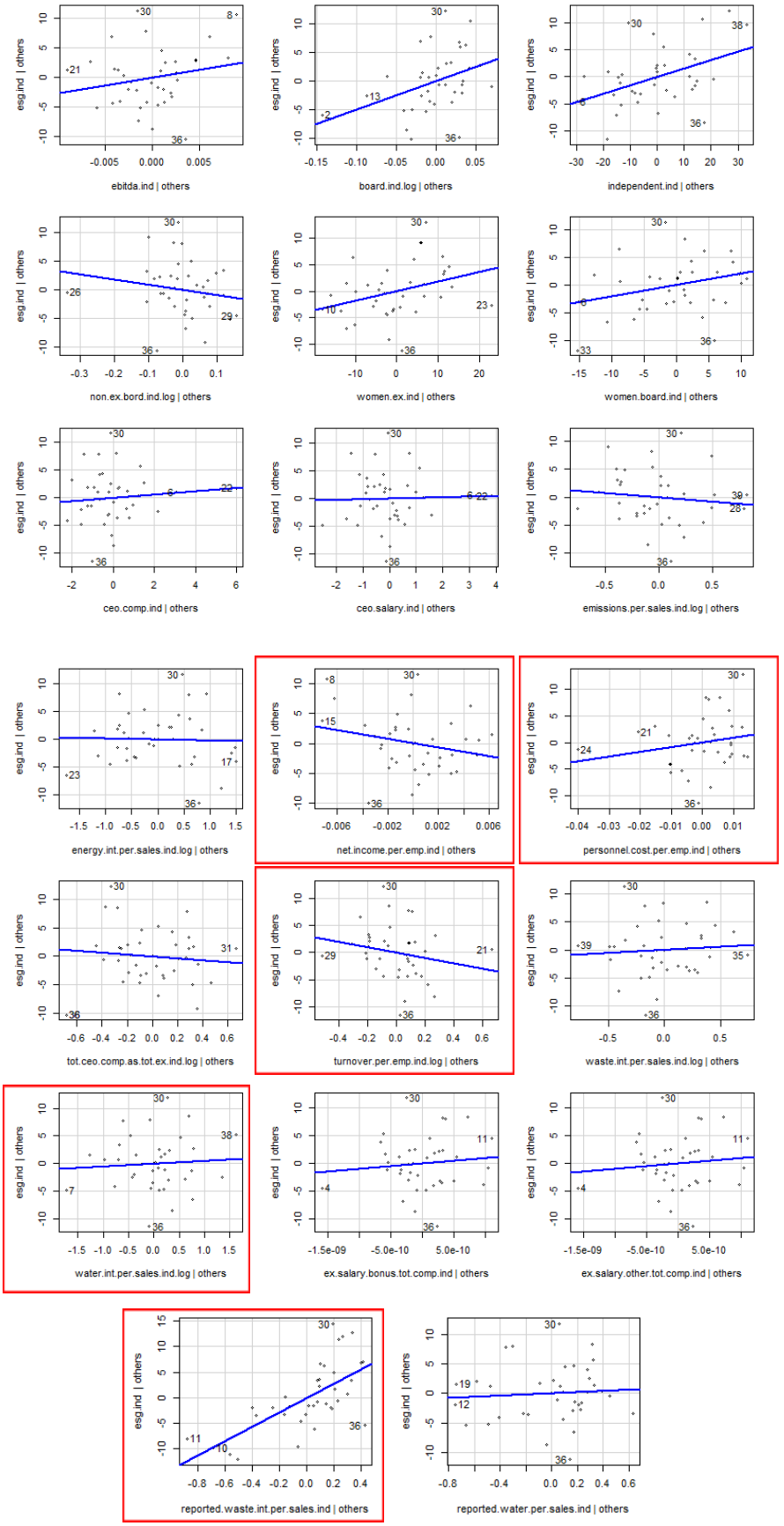


Figur 7: Residualdiagram för modell 2 samt korrelationsmatris för samtliga variabler



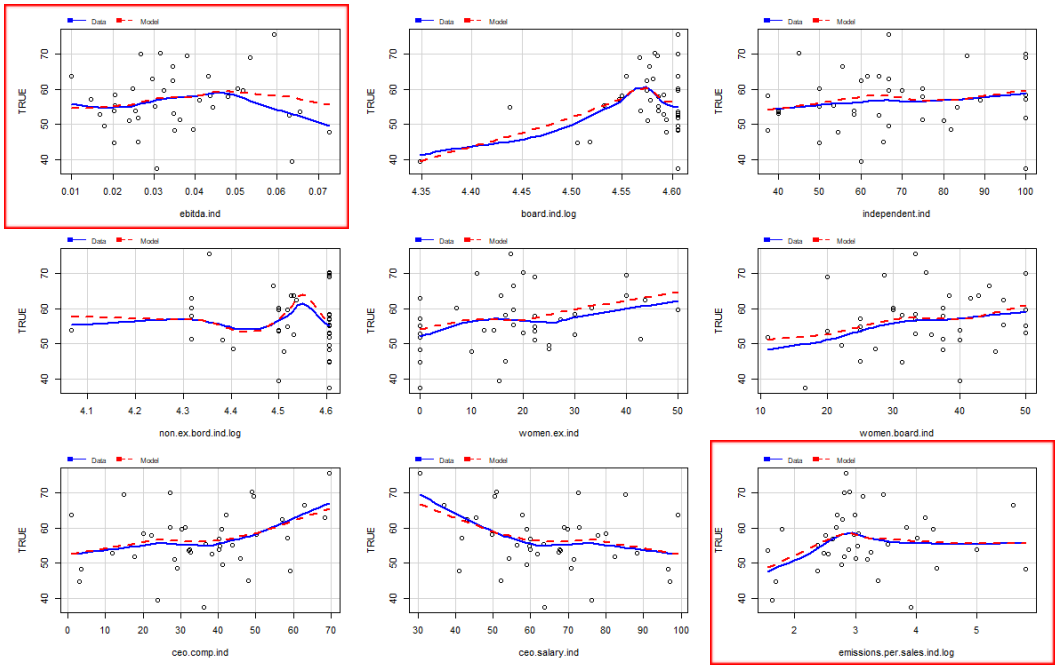
Figur 8: Korrelationsmatris för samtliga variabler

På nästa sida visualiseras Added Variable Plot i figur 9 för de olika variablerna. Värt att notera är hur några av variablerna, sedan det ursprungliga datasetet delades upp, har fått en ny påverkan på modellen. Variabeln *net.income.per.emp.ind* gick från att ha ett positivt samband i modellen till nu ett negativt. *Personnel.cost.per.emp.ind* gick från ett negativt till ett positivt samband. För variabeln *turnover.per.emo.ind* erhölls ett starkare negativt samband. Vidare fick variabeln *water.int.per.sales.ind.log* ett mindre positivt samband och *reported.waste.int.per.sales.ind* erhöll ett starkare samband än den tidigare modellen.

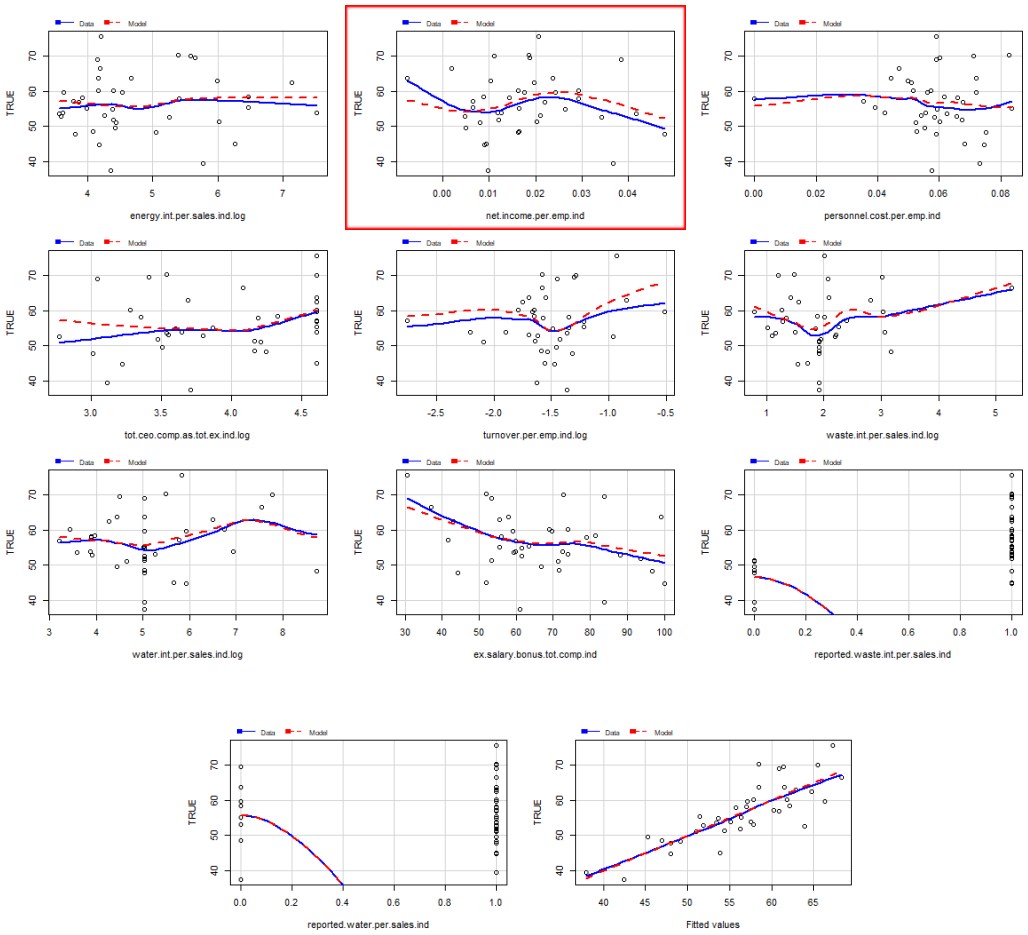


Figur 9: Added Variable Plot för samtliga variabler i datasetet för industri

Figur 10 och 11 presenterar nedan hur väl passad modellen är. Från modell 1 där båda dataseten var inkluderade kunde variablerna *ebitda.per.emp.ind, net.income.per.emp* och *emissions.per.sales* synliggöras med två kluster av observationer. Observeras den första variabeln *ebitda.per.emp.ind* ses en bättre spridning av datan, dock infinner modellen en avvikelse mot datan ju högre värden på x-axeln. För *net.income.per.emp.ind* observeras inte heller en optimal modell för datan, men något mer representativt än modell 1. För variabeln *emissions.per.sales* har modellen passats väl gentemot observationerna.



Figur 10: Marginal Model Plots för modellen med tillverkande industribolag



Figur 11: Marginal Model Plots för modellen med tillverkande industribolag

Efter uppdelningen av dataseten erhåller denna modell enligt tabellen nedan ett R^2 om 0,6841, vilket är 0,1486 enheter högre än modell 1. För R^2_{adj} erhöll modellen ett värde om 0,3682, vilket är en ökning om 0,0108 enheter. Modellen kom i besittning av två signifikanta koefficienter; *reported.waste.int.per.sales.ind* och *independent.ind* där ett p-värde om 0,0503 noteras från tidigare 0,0022 vilket observeras som en försämring.

Modell	R^2	R^2_{adj}	p-värde	Antal signifikanta koefficienter
Modell 2 med data från tillverkande industribolag	0,6841	0,3682	0,0503	2 (reported.waste.int.per.sales.ind) (independent.ind)

Tabell 3: Sammanställning av resultat för modell 2

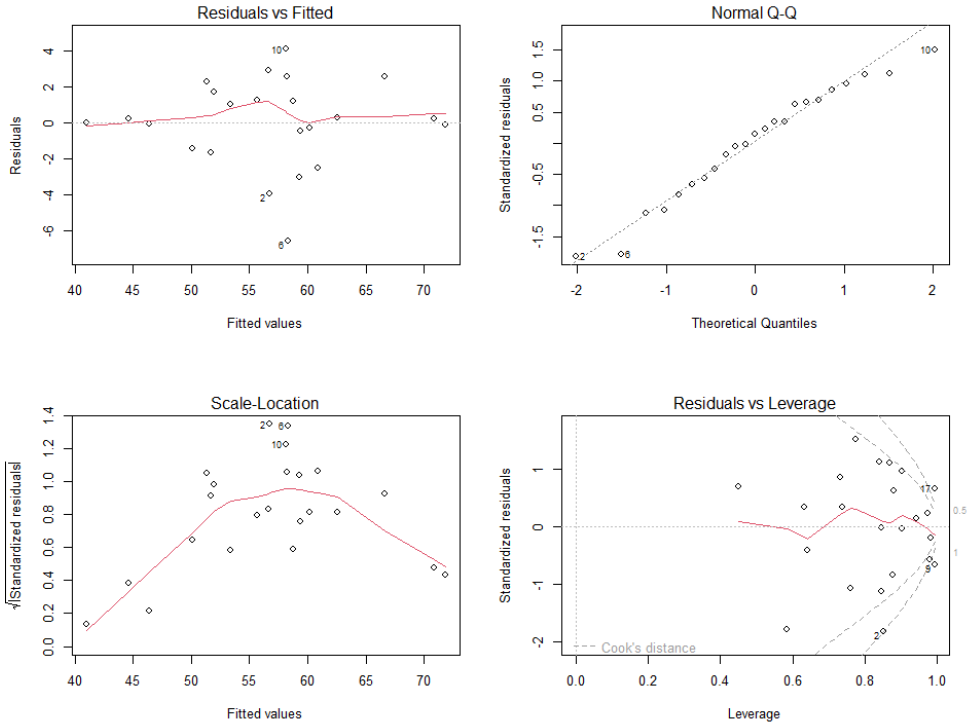
4.2.2.3 Modell 3 - egen modell av tillverkande industribolag

Nedan redogörs för modell 3 med datamaterialet från tillverkande industribolag. I denna modell har några oberoende variabler valts bort då det råder multikollinearitet där analys utförts numeriskt med hjälp av VIF och grafiskt genom korrelationsmatris. Inledningsvis kan det konstateras att modell 3 erhåller relativt god linjäritet och normalfördelade residualer med en någorlunda god homoskedacitet,

där observation 9 och 17 ligger på Cook's linje. Modellen kommer inte utesluta några observationer då det dels är för få och dels föreligger risk för overfitting.

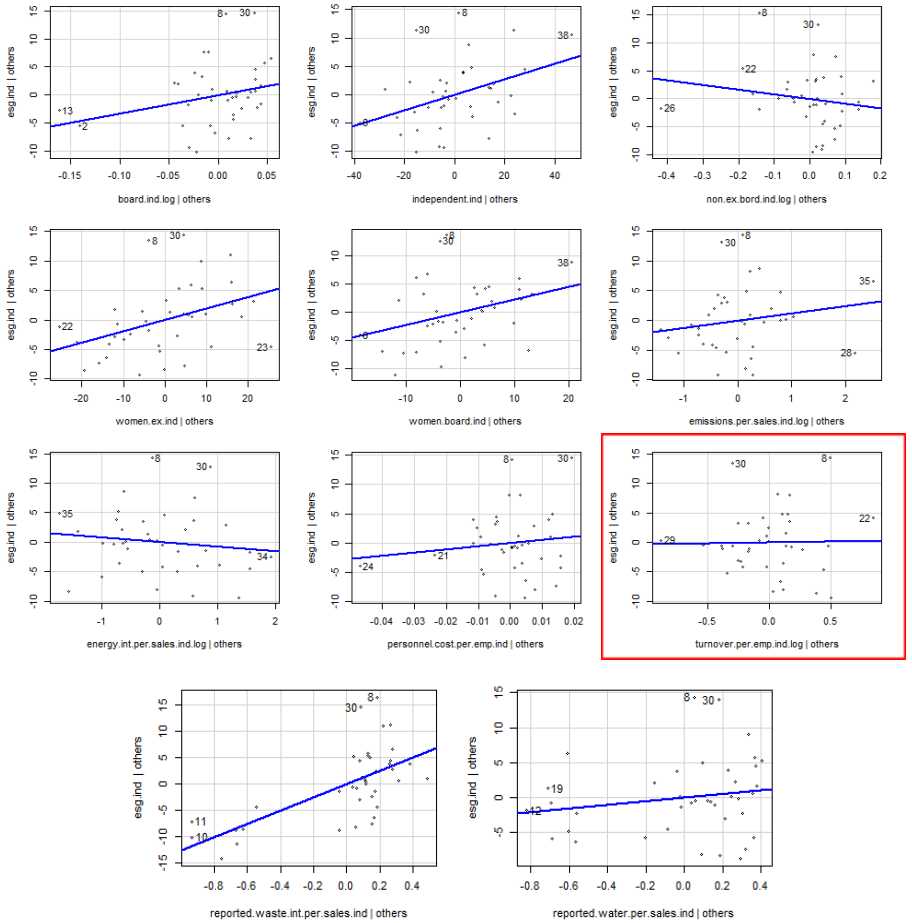
Åtta variabler som på grund av multikollinearitet, enligt figur 8, och erhöll värden över 10 på VIF, togs bort inför framställande av denna modell:

- *net.income.per.emp.ind*
- *waste.int.per.sales.ind.log*
- *water.int.per.sales.ind.log*
- *ex.salary.bonus.tot.comp.ind*
- *ceo.comp.ind*
- *ceo.salary.ind*
- *ebitda.ind*
- *tot.ceo.comp*



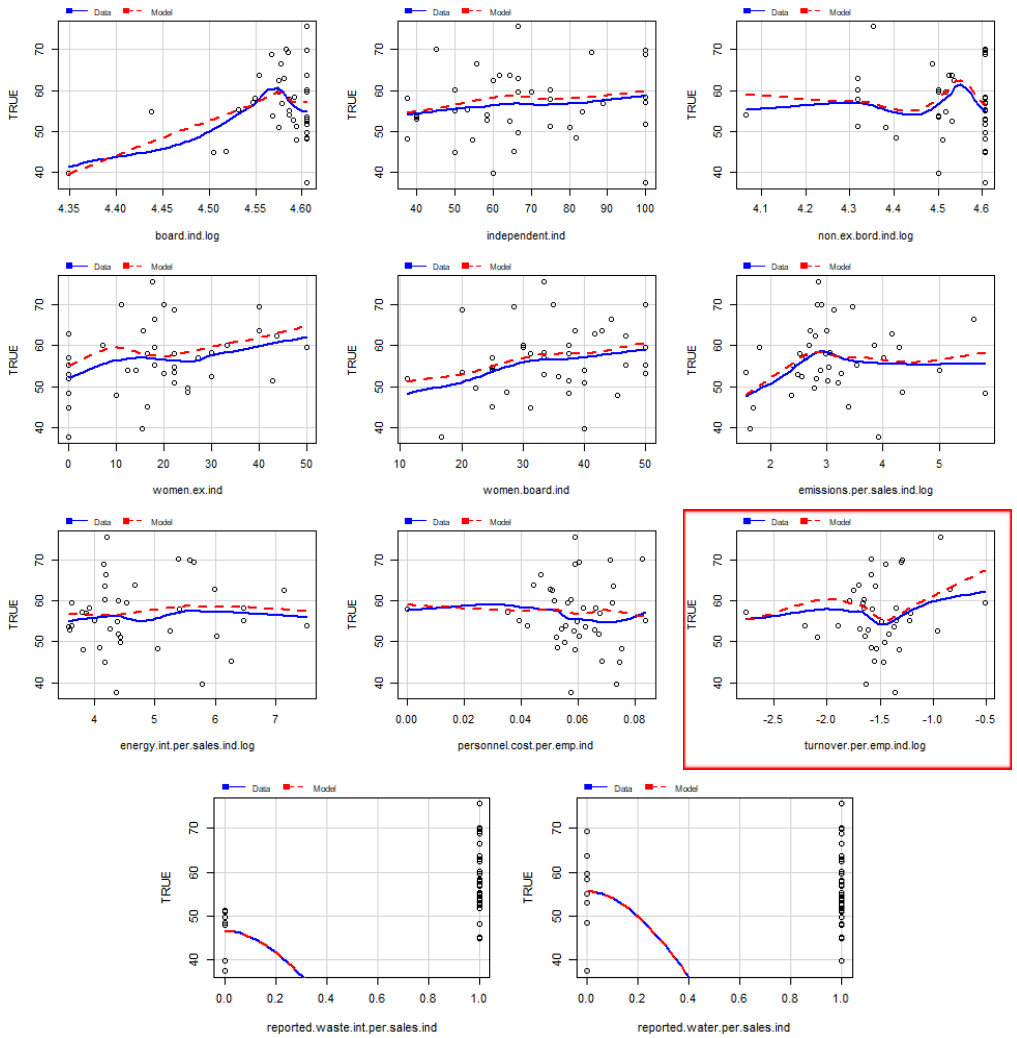
Figur 12: Residualdiagram över den egna modellen för datamaterial tillverkande industribolag

I figur 13 nedan går det att avläsa hur samtliga variabler har en stark inverkan på modellen, med undantag för den oberoende variabeln *turnover.per.emp.ind.log* som endast påvisar en marginellt ökande trend. Oberoende så väljs den till trots att vara med, givet balansgången med att göra sig av med för många variabler kontra pay-offen gentemot komplexitet för modellen.



Figur 13: Added Variable Plot för den egna modellen för tillverkande industribolag

Vad gäller modellens representation av datan kan man i enlighet med figur 14 nedan fastställa att samtliga oberoende variabler representeras relativt väl av modellen där endast variabeln *turnover.per.emp.ind.log* med värden mellan -1,5 och -0,5 avviker från modellens passning.



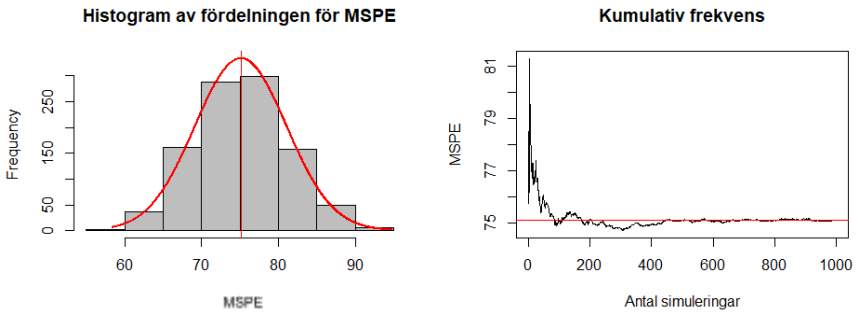
Figur 14: Residualdiagram över den egna modellen för tillverkande industribolag

Sammantaget går det att konstatera från den tredje modellen i tabell 4 att förklaringsgraden R^2 om 0,5866 erhölls vilket är en sänkning med 0,0975 enheter från modell två, varpå R^2_{adj} fick 0,4181 vilket är en ökning om 0,0499 enheter i jämförelse med modell 2. Modellen i sin helhet fick ett lågt p-värde om 0,0040 och kom i besittning av tre signifikanta variabler; *reported.waste.int.per.sales, independent.ind* och *women.ex.ind.*

Modell	R^2	R^2_{adj}	p-värde	AIC	MSPE	Antal signifikanta koefficienter
Modell 3 - egen modell av tillverkande industribolag	0,5866	0,4181	0,0040	281,88	75,1195	3 (reported.waste.int.per.sales) (independent.ind) (women.ex.ind)

Tabell 4: Sammanställning av resultat för modell 3

För att kunna avgöra modellens prediktionsfel och sedermera jämföra med Lasso och Ridge har MSPE beräknats med korsvalidering genom Monte Carlo Simulation där ett MSPE om 75,1195 kan observeras. Vidare går det i figur 15 att observera hur fördelningen av MSPE följer normalfördelningen och att ett AIC värde om 281,88.

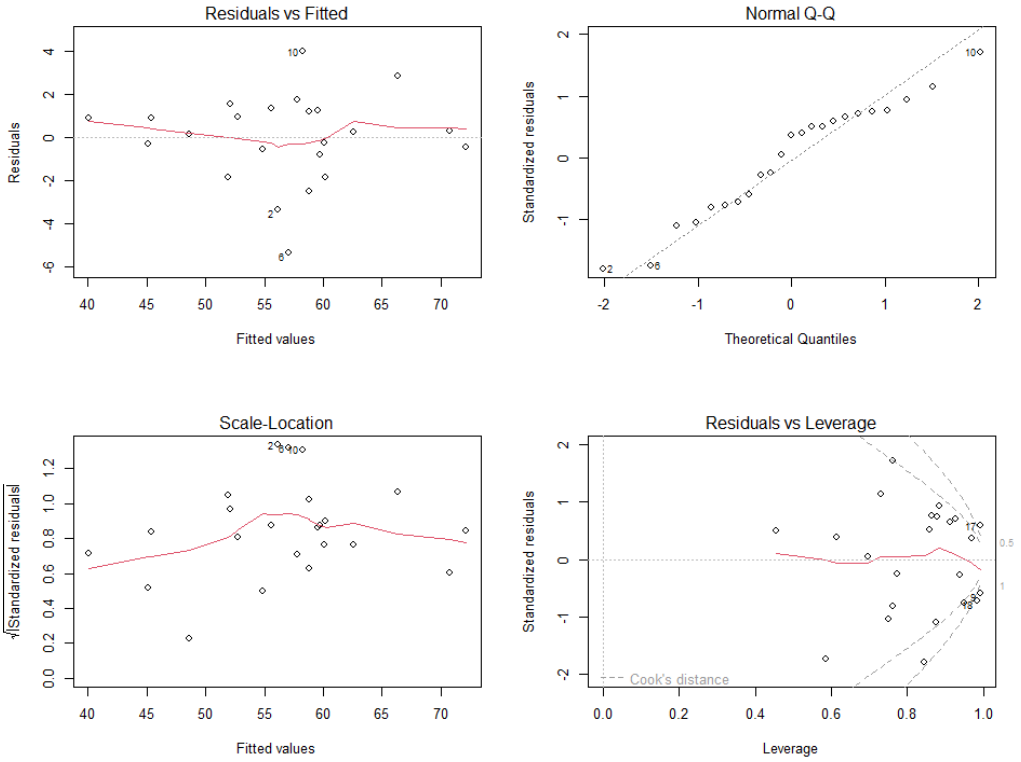


Figur 15: Histogram över fördelningen av ESG-betyget för 1000 simuleringar samt den kumulativa frekvensen

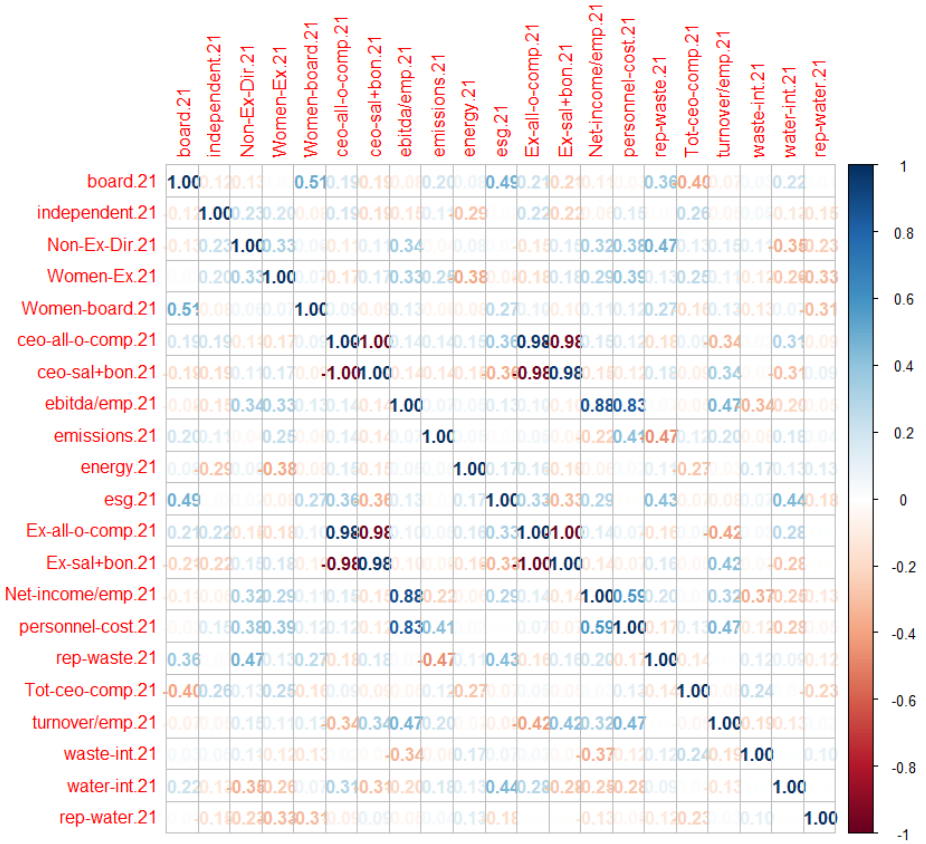
4.2.2.4 Modell 4 - dagligvaruproducenter

För denna modell är det istället endast datasetet för dagligvaruproducenter som undersöks. Istället för 39 observationer i industridatan så är det endast 23. Den första grafen i figur 16 nedan indikerar att modellen är relativt linjär, erhåller några residualer som inte följer normalfördelningen, någorlunda homoskedastisk och har ett par variabler som ligger innanför Cook's distance. Modellen i sin helhet är inte optimal för att kunna anses uppfylla samtliga grundantaganden om multipel linjär regression.

I korrelationsmatrisen i figur 17 går det att avläsa en del variabler som lider av multikollinearitet vilka kommer att tas hänsyn till när den egna modellen kommer framställas.

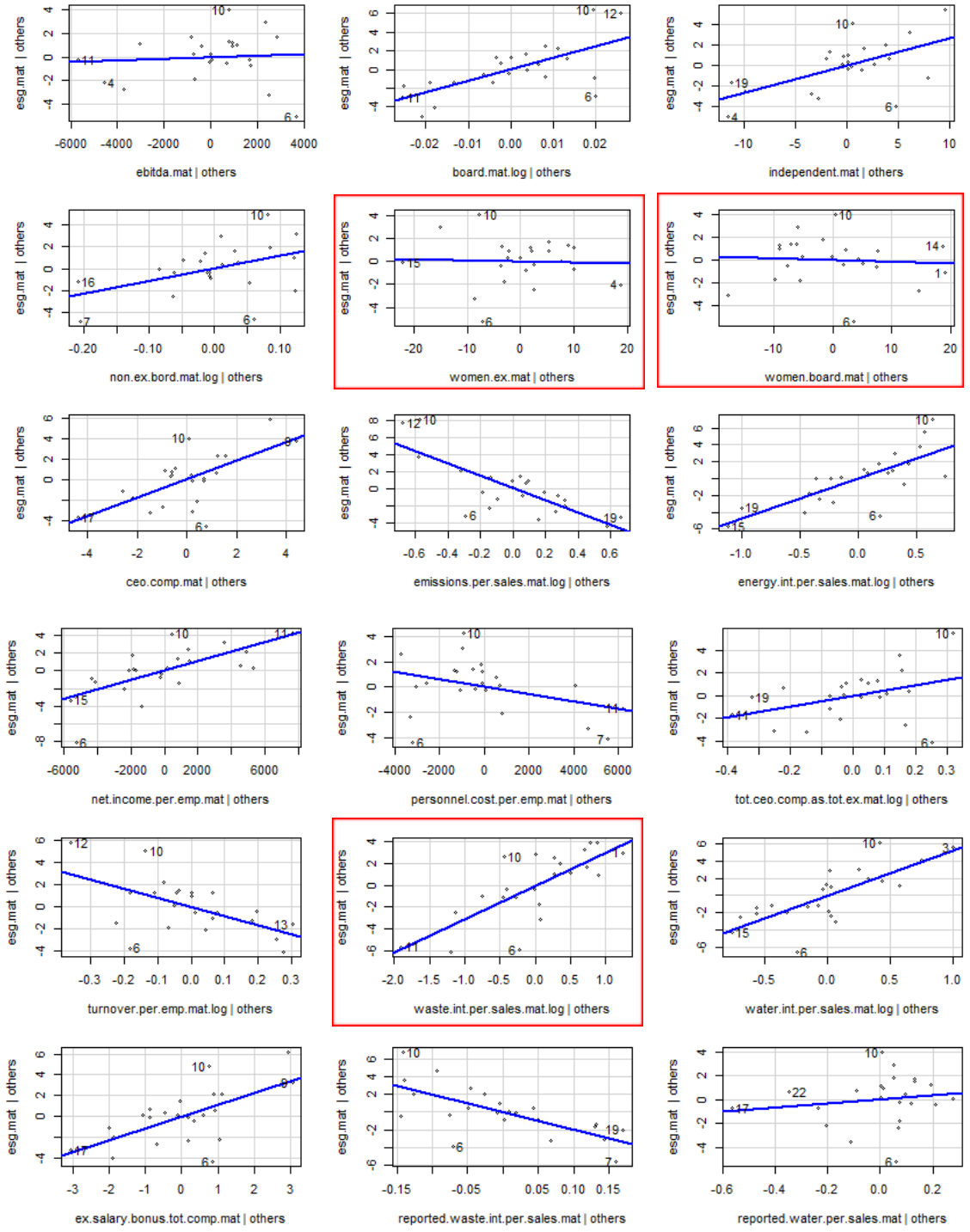


Figur 16: Residualdiagram över den egna modellen för datamaterial dagligvaruproducenter



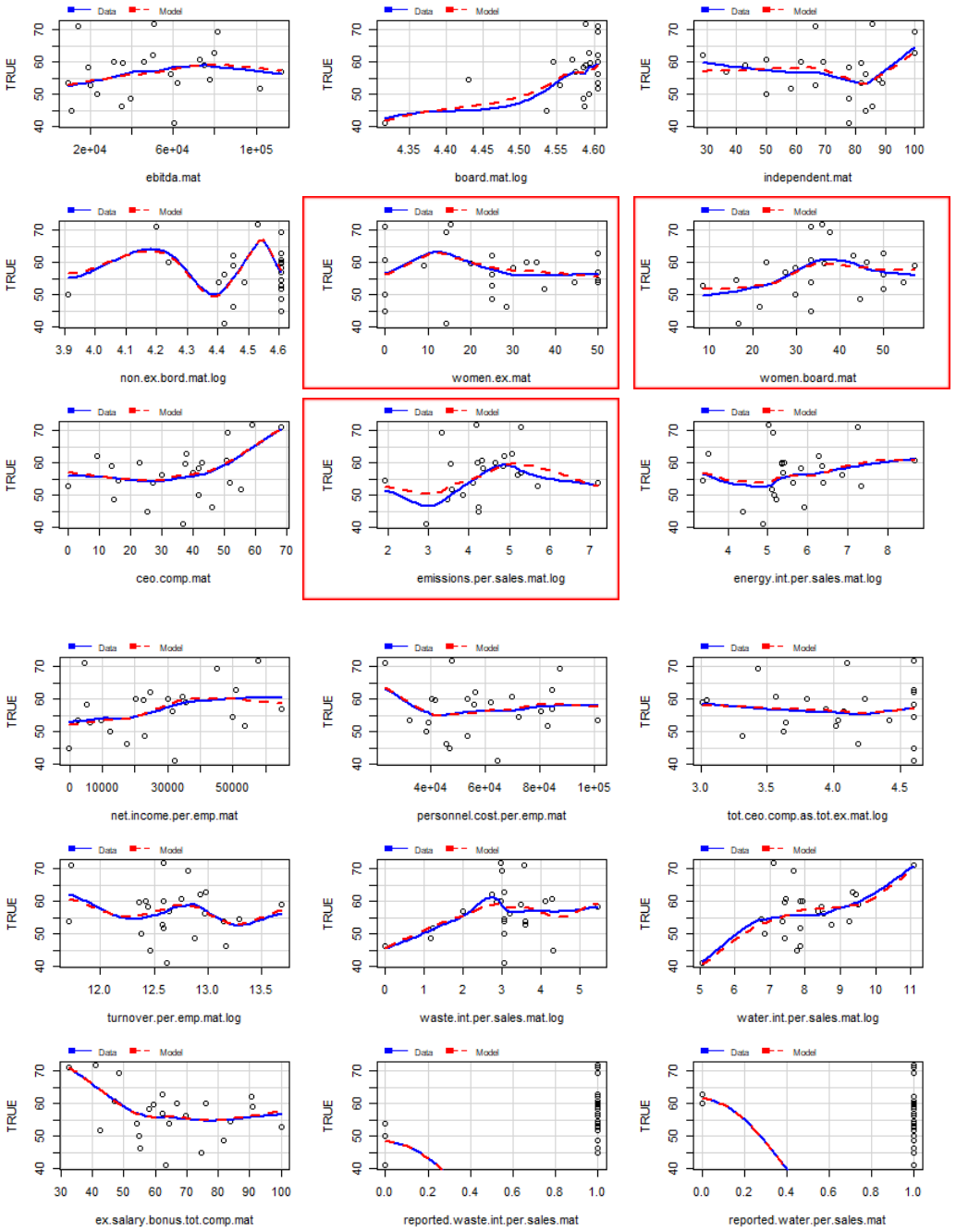
Figur 17: Korrelationsmatris för dagligvaruproducenter

I figur 18 nedan går det att avläsa samtliga oberoende variabler utom *women.ex.mat* och *women.board.mat* som kraftigt bidragande till modellen. Värt att notera är hur starkt den oberoende variablerna *waste.int.per.sales.mat.log* bidrar till modellen.



Figur 18: Added Variable Plot för samtliga variabler i datasetet för dagligvaruproducenter

Nedan går det i figur 19 att avläsa hur, mer eller mindre, samtliga oberoende variabler förklaras väl av modellen med *women.ex.mat, women.board,* och *emissions.per.sales.mat.log* som möjliga avstickare. En intressant observation är hur är både *women.ex.mat* och *women.board* inte har en starkt bidragande roll till modellen, samtidigt som modellen har svårt att representera datan.



Figur 19: Marginal Model Plots för modellen med dagligvaruproducenter

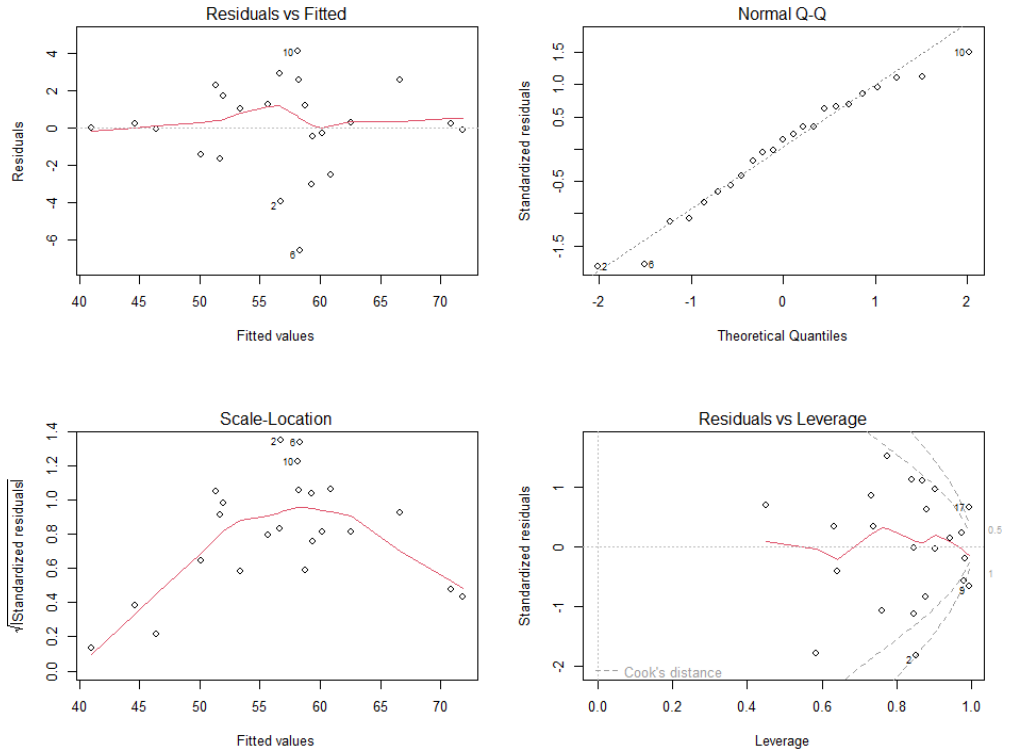
Sammanfattningsvis går det att konstatera för denna modell att antagandena i sin helhet inte är uppfyllda men till trots erhåller en relativt hög förklaringsgrad på ett R^2 om 0,9346, vilket är en ökning om 0,3980 från modell 1 och ett R^2_{adj} om 0,6403 som i sin tur är en ökning om 0,2829 enheter. Dock går det inte att utläsa att några signifikanta koefficienter från modellen.

Modell	R^2	R^2_{adj}	p-värde	Antal signifikanta koefficienter
Modell 4 med data från industri och mat	0,9346	0,6403	0,1357	0

Tabell 5: Sammanställning av resultat för modell 4

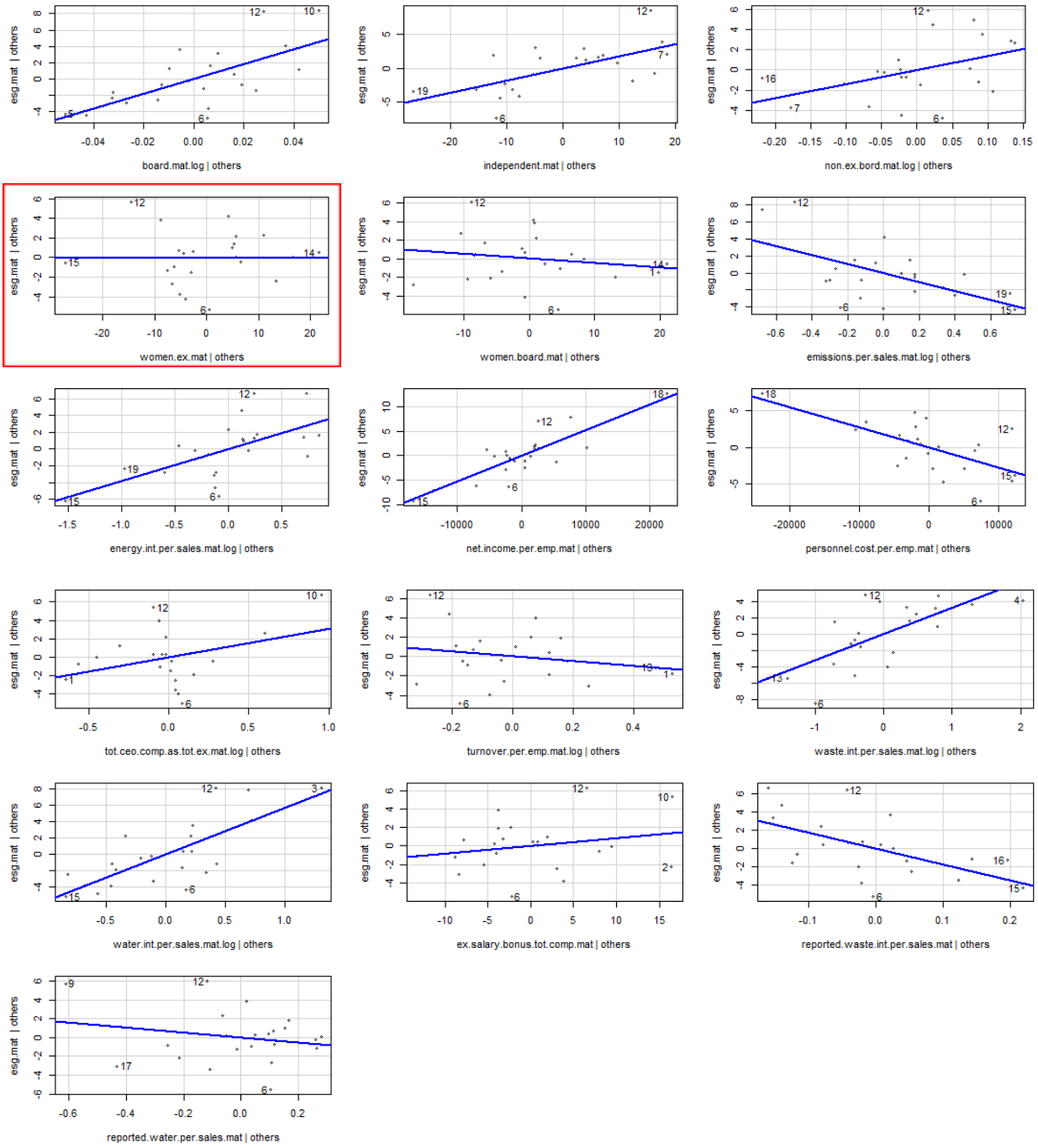
4.2.2.5 Modell 5 - egen modell av dagligvaruproducenter

För den egenkonstruerade modellen för kategorin dagligvaruproducenter kan linjäritet, normalfördelade residualer, homoskedasticitet och ett par observationer som ligger inom Cook's distance observeras i figur 20 nedan. I sin helhet konstateras att modellen är acceptabel att göra vidare undersökning på.



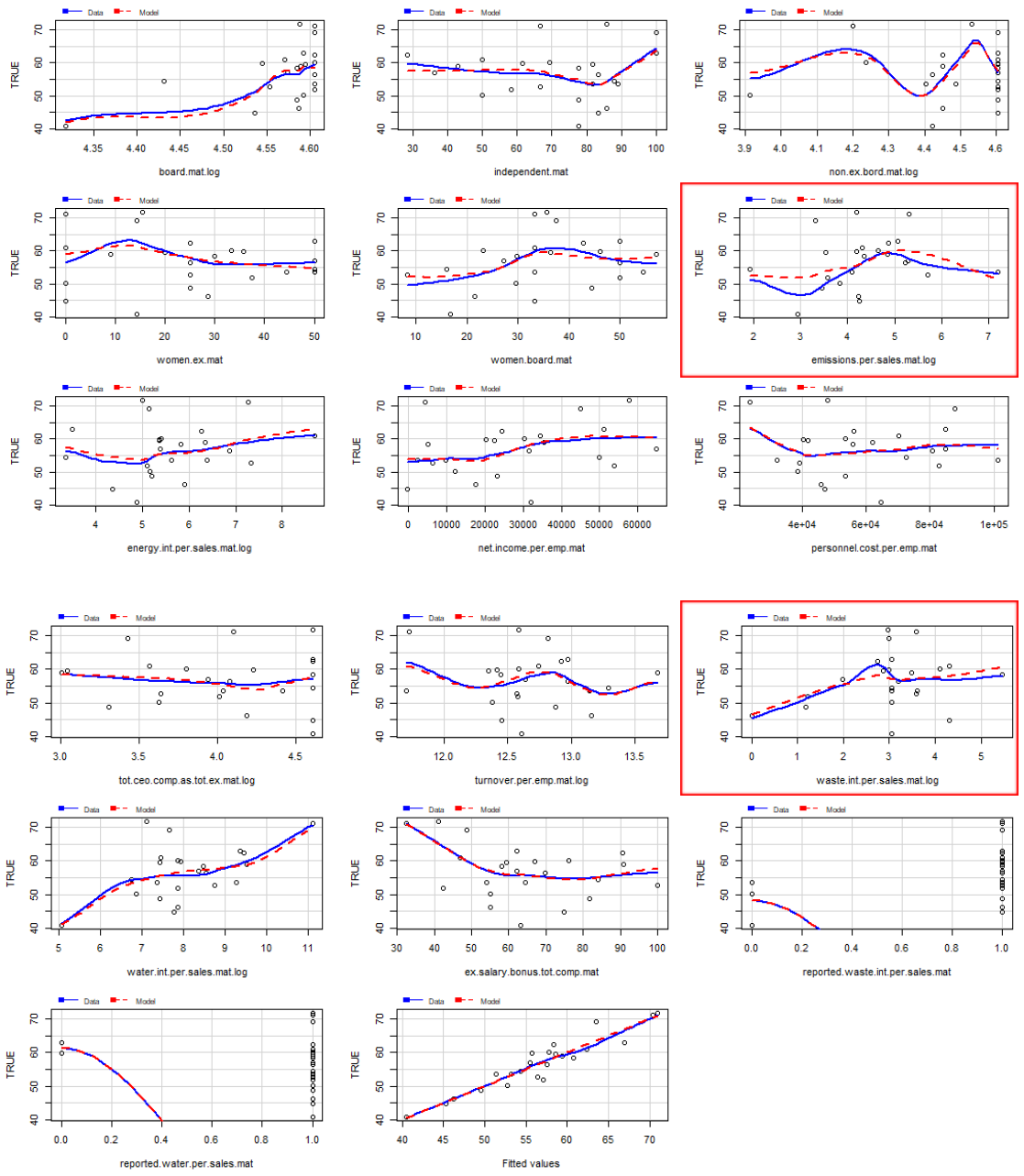
Figur 20: Residualdiagram över egna modell för dagligvaruproducenter

För variabler med en bidragande faktor till modellen kan det i figur 21 nedan konstateras att majoriteten av dem har en bidragande effekt. Den oberoende variabel som bidrar minst till modellen kan genom visuell analys konstateras vara *women.ex.mat* vilken till trots väljs att ha kvar, för att undvika overfitting.



Figur 21: Added Variable Plot för den egna modellen av dagligvaruproducenter

I figur 22 kan man vid en första anblick se att vi har en relativt väl passad modell för samtliga observationer och oberoende variabler. Två avstickare värda att notera är *emissions.per.sales.mat* och *waste.int.per.sales* vilka modellen får svårt att göra en bra passning av.

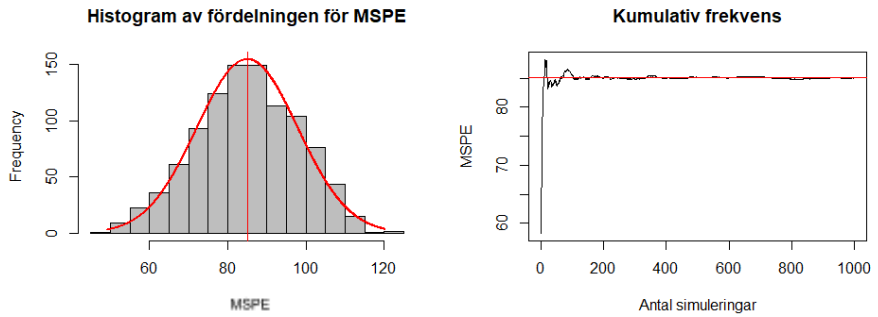


Figur 22: Marginal Model Plots för den egna modellen av dagligvaruproducenter

Sammanfattningsvis kan ett något lägre R^2 om 0,8808 observeras till skillnad från föregående modell om 0,9346 vilket är en minskning på 0,0538 enheter. Modellen erhåller ett R^2_{adj} om 0,5844 vilket är en minskning om 0,0559 enheter från föregående modell, däremot med tre signifikanta koefficienter. Vidare går det att observera ett MSPE om 85,0875. Likaså går det att observera ett AIC värde om 316,33. Även här iakttas att MSPE är normalfördelat efter korsvalideringen enligt figur 23 nedan.

Modell	R^2	R^2_{adj}	p-värde	AIC	MSPE	Antal signifikanta koefficienter
Modell 5 - egen modell av dagligvaruproducenter	0,8867	0,5844	0,0949	316,33	85,0875	3 (net.income.per.emp.mat) (waste.int.per.sales.mat.log) (water.int.per.sales.mat.log

Tabell 6: Sammanställning av resultat för modell 5



Figur 23: Histogram över fördelningen av ESG-betyget för 1000 simuleringar samt den kumulativa frekvensen

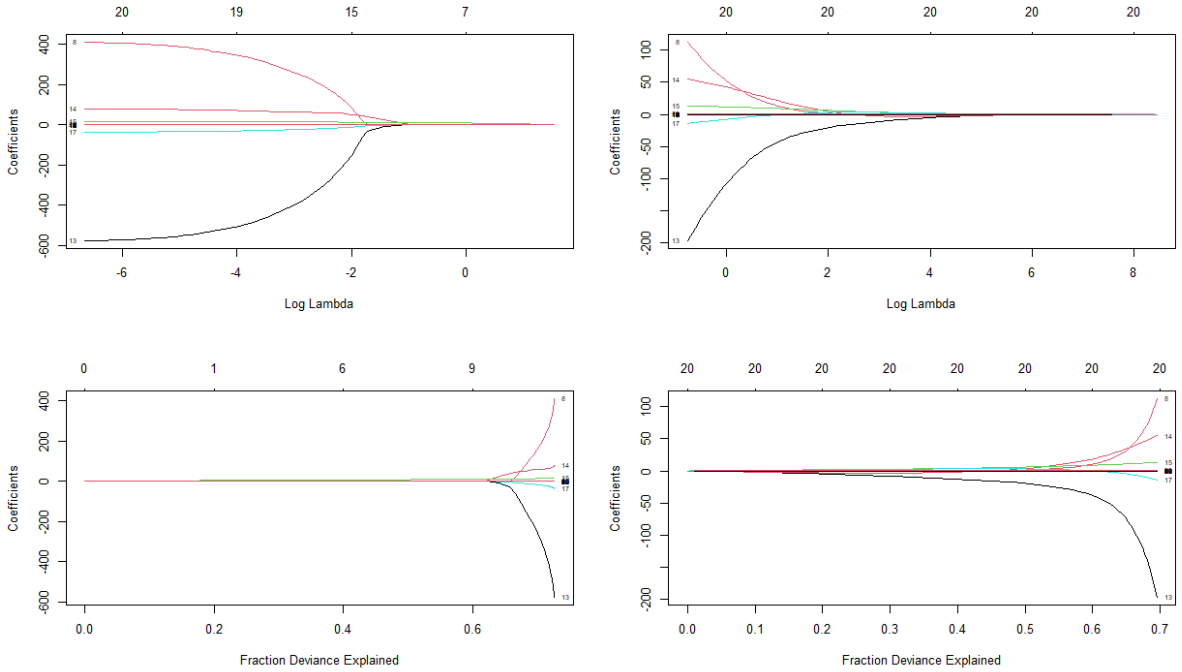
4.3 Lasso & Ridge

I detta avsnitt kommer de två regulariseringstekniker, Ridge och Lasso, användas för att hantera och motverka overfitting i modellerna till träningsdata och förbättra generaliseringsförmågan hos modellen. Detta för att kunna analysera skillnader mellan de förklarande variablerna av de olika dataseten för dagligvaruproducenter och tillverkande industribolag. Verktyget AIC har inte kunnat användas för Lasso och Ridge.

4.3.1 Lasso och Ridge med datasetet för tillverkande industribolag

I figur 24 nedan går det att avläsa i de två översta rutorna, Ridge i den vänstra och Lasso i högra, värdena på strafftermen λ för att krympa värdet på de förklarande variablerna. På y-axeln avläses koefficientens storlek för de olika förklarande variablerna och på x-axeln avläses storleken på strafftermen λ i logaritmerad skala. Axeln i toppen redogör för antalet nollskilda koefficienter. Nollskilda koefficienter visar hur många av koefficienter som inte är noll för ett givet värde på strafftermen λ. Vad gäller båda termerna blir modellen allt mindre komplex beroende på om strafftermen tas med, eller storleken på denne. För Ridge modellen kommer samtliga förklarande variabler vara kvar oberoende storleken på λ, där Lasso istället kommer utesluta variabler som påvisar multikollinearitet. Det vill säga att antalet variabler minskar för Lasso när log(λ) ökar medans i Ridge bibehålls samtliga variabler, med risk att de kan gå mot/nära noll. Värt att notera i figur 24 är hur de oberoende variablerna *ebitda.per.emp* (8) och *Ex.sal+bonus* (13) erhåller höga värden på de förklarande variablerna och därmed får en högre straffterm där variablerna sätts till noll vid ungefär log(-1.8) för Lasso. För Ridge sätts båda termerna för ett log(λ) om fem.

Av vad som går att avläsa i de två nedre rutorna av figur 24, Lasso i den högra och Ridge i den vänstra, är modellens förklaringsgrad (Fraction Deviance Explained) på x-axeln och koefficientens storlek på y-axeln. På topp-axeln i x-led redogörs för antalet variabler skilda från noll. För Lasso är det värt att notera hur förklaringsgraden som uppgår till strax över 0,6 när multikollinearitet råder och för Ridge strax innan 0,6 där variabel *Ex.sal+bonus* (13) är en avstickare.



Figur 24: De två översta graferna visar för vilket λ som sätter koefficienterna till noll, Lasso till vänster och Ridge till höger. De två nedersta graferna visar förklaringsgraden i relation till koefficienterna storlek, Lasso till vänster och Ridge till höger.

I figur 25 nedan går det att avläsa termen measure vilken representerar MSPE för både Ridge och Lasso. Inledningsvis går det att konstatera att Lasso har ett lägre MSPE än Ridge med en differens om 3,48 enheter. För den egenkonstruerade modellen för datan av tillverkande industribolag med ett MSPE om 75,1195 ger Lasso-modellen ett lägre prediktionsfel om 53,32.

```
> cv.LASSO

Call:  cv.glmnet(x = X, y = Y, foldid = indelning, alpha = 1)

Measure: Mean-Squared Error

    Lambda Index Measure    SE Nonzero
min 0.7927    20   53.32 11.72       7
1se 2.9160     6   64.40 17.29       1
> cv.ridge

Call:  cv.glmnet(x = X, y = Y, foldid = indelning, alpha = 0)

Measure: Mean-Squared Error

    Lambda Index Measure    SE Nonzero
min   10.0    67   56.80 12.93      20
1se  284.9    31   69.64 17.90      20
```

Figur 25: MSPE för Lasso och Ridge för tillverkande industribolag

Nedan går det i figur 26 att avläsa vilka variabler Lasso väljer att sätta till noll och vilka variabler Ridge väljer att sätta näst intill noll. Värt att poängtera är hur Lasso väljer att endast ta med sju variabler givet trade-off:n komplexitet och förklaringsgrad. Två variabler som gör sig märkbara för Lasso, med relativt stora koefficient, är *Women-Ex.21* och *rep.waste.21*. För Ridge är det värt att uppmärksamma de förklarande variablerna *turnover/emp.21* och *rep-water.21* vilka är de två variabler som erhåller de två största koefficienterna.

```
21 x 1 sparse Matrix of class "dgCMatrix"                    21 x 1 sparse Matrix of class "dgCMatrix"
                                  s1                                                           s1
(Intercept)             10.118587072          (Intercept)             -2.042002e+00
board.21                 0.286606693          board.21                 4.146047e-01
independent.21           0.039736199          independent.21           1.106300e-01
`Non-Ex-Dir.21`          .                    `Non-Ex-Dir.21`         -8.391352e-02
`Women-Ex.21`            0.110653398          `Women-Ex.21`            1.655775e-01
`Women-board.21`         0.058315519          `Women-board.21`         1.773344e-01
`ceo-all-o-comp.21`      0.092957642          `ceo-all-o-comp.21`      7.626747e-02
`ceo-sal+bon.21`         .                    `ceo-sal+bon.21`        -4.724888e-02
`ebitda/emp.21`          .                    `ebitda/emp.21`          6.737937e+01
emissions.21             .                    emissions.21            -1.770708e-02
energy.21                .                    energy.21               -1.673962e-03
`Ex-all-o-comp.21`       .                    `Ex-all-o-comp.21`       4.572597e-03
`Ex-sal+bon.21`          .                    `Ex-sal+bon.21`         -5.028405e-03
`Net-income/emp.21`      .                    `Net-income/emp.21`     -1.316252e+02
`personnel-cost.21`      .                    `personnel-cost.21`      4.650277e+01
`rep-waste.21`          10.210164762          `rep-waste.21`           1.231153e+01
`Tot-ceo-comp.21`        .                    `Tot-ceo-comp.21`       -2.232870e-02
`turnover/emp.21`        .                    `turnover/emp.21`       -9.372301e+00
`waste-int.21`           0.005387802          `waste-int.21`           2.520560e-02
`water-int.21`           .                    `water-int.21`           8.195938e-04
`rep-water.21`           .                    `rep-water.21`           1.091888e+00
```

Figur 26: Storleken på respektive koefficient för modellerna samt om den sätts till noll, Lasso till vänster och Ridge till höger

I tabellen nedan redogörs för förklaringsgraden för respektive modell av Ridge och Lasso, där Ridge erhåller en lägre förklaringsgrad om 0,4828 och Lasso en högre förklaringsgrad om 0,6138. Som tidigare konstaterats erhåller dessutom Lasso ett lägre prediktionsfel än Ridge.

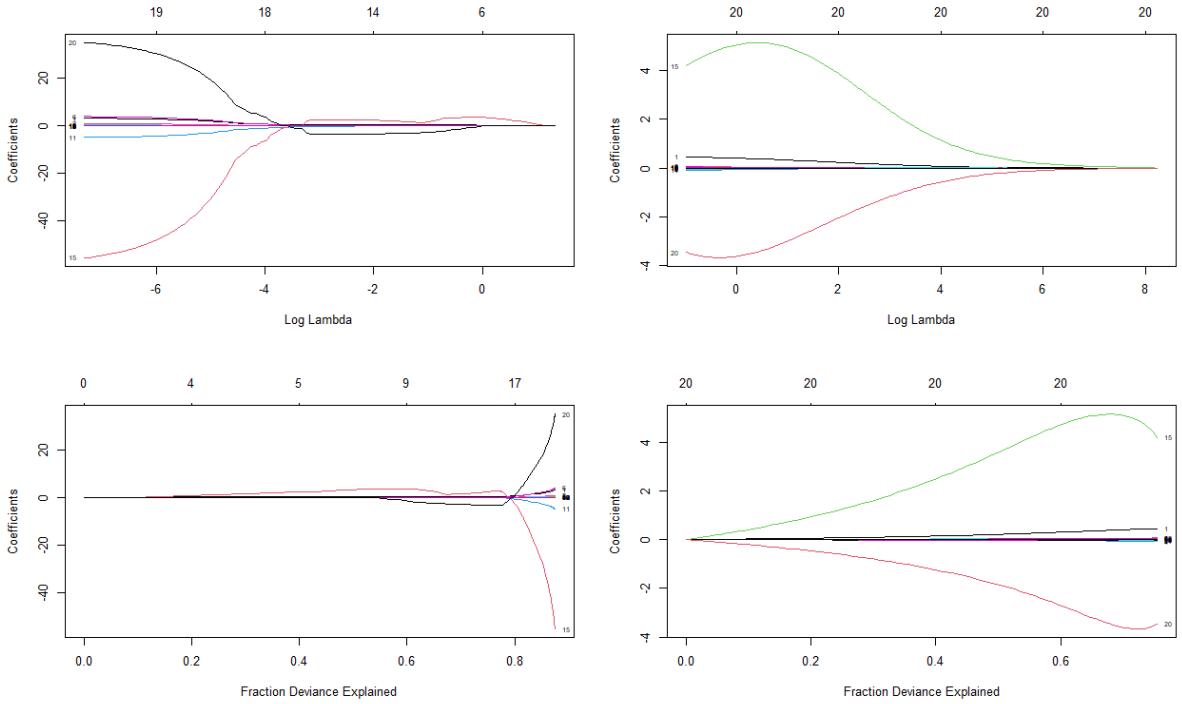
Modell	R^2	AIC	MSPE
Ridge för tillverkande industribolag	0,4828	N/A	56,80
LASSO för tillverkande industribolag	0,6138	N/A	54,53

Tabell 7: Sammanställning av resultat för modellerna med Ridge och Lasso för tillverkande industribolag

4.3.2 Lasso och Ridge med datasetet för dagligvaruproducenter

Av vad som går att avläsas i figur 27 nedan är de oberoende variablerna *water.int* (20) och *personnel-cost* (15) för både Ridge och Lasso vilka erhåller höga värden på koefficienterna, och därav en högre straffterm.

Av vad som går att avläsa i figur 27, Lasso i den högra, och Ridge i den vänstra, att för Lasso är förklaringsgraden strax över 0,8 när multikollinearitet råder och för Ridge strax innan 0,6 där variable *personnel-cost* och *water.int* är avstickare.



Figur 27: De två översta graferna visar för vilket λ som sätter koefficienterna till noll, Lasso till vänster och Ridge till höger. De två nedersta graferna visar förklaringsgraden i relation till koefficienterna storlek, Lasso till vänster och Ridge till höger.

Värt att notera i figur 28 nedan är hur prediktionsfelet (measure) för Lasso-modellen, om 51,03, är något lägre än för Lasso-modellen med datamaterialet för tillverkande industribolag, om 53,32. Ridge-modellen erhåller ett prediktionsfel om 60,83 som är något högre än för datamaterialet för tillverkande industribolag, vilket var 56,80.

```
> cv.LASSO.2

Call:  cv.glmnet(x = X.2, y = Y.2, foldid = indelning.2, alpha = 1)

Measure: Mean-Squared Error

    Lambda Index Measure    SE Nonzero
min 0.7054    19   51.03 12.92       8
1se 1.4848    11   63.84 16.93       5
> cv.ridge.2

Call:  cv.glmnet(x = X.2, y = Y.2, foldid = indelning.2, alpha = 0)

Measure: Mean-Squared Error

    Lambda Index Measure    SE Nonzero
min      7    69   60.83 16.10      20
1se   3765     1   69.24 18.96      20
```

Figur 28: MSPE för Lasso och Ridge

I den vänstra tabellen i figur 29 nedan går Lasso-modellen att avläsa vilken erhåller sju oberoende variabler. Värt att lägga märke till är de relativt stora koefficienterna *rep-waste.21, rep-water.21* och *board.21*. I Ridge-modellen erhåller denne dessutom stora värden för *rep-water.21* och *rep-waste.21*.

```
21 x 1 sparse Matrix of class "dgCMatrix"
                                s1
(Intercept)         13.1991818221
board.21             0.3628113376
independent.21       .
`Non-Ex-Dir.21`      .
`Women-Ex.21`        .
`Women-board.21`     .
`ceo-all-o-comp.21`  0.0439726357
`ceo-sal+bon.21`     .
`ebitda/emp.21`      .
emissions.21         .
energy.21            .
`Ex-all-o-comp.21`   .
`Ex-sal+bon.21`      .
`Net-income/emp.21`  0.0001067139
`personnel-cost.21`  .
`rep-waste.21`       3.6290194956
`Tot-ceo-comp.21`    .
`turnover/emp.21`    .
`waste-int.21`       0.0051359079
`water-int.21`       0.0001782079
`rep-water.21`      -0.6844219201
```

```
21 x 1 sparse Matrix of class "dgCMatrix"
                                 s1
(Intercept)          1.124557e+01
board.21             4.232780e-01
independent.21       4.916369e-02
`Non-Ex-Dir.21`     -4.082071e-02
`Women-Ex.21`       -7.177977e-02
`Women-board.21`     1.564857e-02
`ceo-all-o-comp.21`  3.058613e-02
`ceo-sal+bon.21`    -2.971384e-02
`ebitda/emp.21`     -6.277334e-06
emissions.21         5.700456e-03
energy.21            5.278250e-04
`Ex-all-o-comp.21`  -2.757243e-02
`Ex-sal+bon.21`      2.648632e-02
`Net-income/emp.21`  2.524311e-04
`personnel-cost.21` -6.337519e-05
`rep-waste.21`       4.930595e+00
`Tot-ceo-comp.21`    4.977808e-03
`turnover/emp.21`   -4.424552e-06
`waste-int.21`       3.065221e-02
`water-int.21`       1.828664e-04
`rep-water.21`      -3.668227e+00
```

Figur 29: Storleken på respektive koefficient för modellerna samt om den sätts till noll, Lasso till vänster och Ridge till höger

I tabell 8 nedan presenteras MSPE, AIC och förklaringsgraden R^2 för Ridge och Lasso i datamaterialet för tillverkande industribolag. Av vad som kan urskiljas från tidigare avsnitt av datamaterialet för dagligvaruproducenter erhölls en förklaringsgrad om 0,4828 för Ridge och 0,6138 för Lasso, varpå det för dagligvaruproducenter nu erhålls ett värde om 0,5372 för Ridge och 0,5832 för Lasso. Värt att notera är hur MSPE för Lasso av dagligvaruproducenter förbättrades från 54,53 för tillverkande industribolag till 51,03. För Ridge steg MSPE för dagligvaruproducenter från 56,80 för industri till 60,83.

Modell	R^2	AIC	MSPE
Ridge för dagligvaruproducenter	0,5372	N/A	60,83
Lasso för dagligvaruproducenter	0,5832	N/A	51,03

Tabell 8: Sammanställning av resultat för modellerna med Ridge och Lasso för dagligvaruproducenter

5 Resultat

	R^2	R^2_{adj}	p-värde	AIC	MSPE	Antal signifikanta koefficienter	Koefficienter i modellen
Modell 1: All data kombinerat	0,5366	0,3574	0,0022	N/A	N/A	1 (reported.waste.int.per.sales)	Samtliga
Modell 2: All data för tillverkande industribolag	0,6841	0,3682	0.0503	N/A	N/A	2 (reported.waste.int.per.sales.ind) (independent.ind)	Samtliga
Modell 3: Egen modell för tillverkande industribolag	0,5866	0,4181	0,0040	281,88	75,1195	3 (reported.waste.int.per.sales) (independent.ind) (women.ex.ind)	board.ind.log independent.ind non.ex.bord.ind.log women.ex.ind women.board.ind emissions.per.sales.ind.log energy.int.per.sales.ind.log personnel.cost.per.emp.ind turnover.per.emp.ind.log reported.waste.int.per.sales.ind reported.water.per.sales.ind
Modell 4: All data dagligvaru-producenter	0,9346	0,6403	0,1357	N/A	N/A	0	Samtliga
Modell 5: Egen modell för dagligvaru-producenter	0,8867	0,5844	0,0949	316,33	85,0875	3 (net.income.per.emp.mat) (waste.int.per.sales.mat.log) (water.int.per.sales.mat.log	board.mat.log independent.mat tot.ceo.comp.as.tot.ex.mat.log turnover.per.emp.mat.log waste.int.per.sales.mat.log women.ex.mat women.board.mat emissions.per.sales.mat.log ex.salary.bonus.tot.comp.mat reported.waste.int.per.sales.mat energy.int.per.sales.mat.log net.income.per.emp.mat personnel.cost.per.emp.mat reported.water.per.sales.mat
Ridge: Tillverkande industribolag	0,4828	N/A	N/A	N/A	56,80	N/A	Samtliga
Lasso: Tillverkande industribolag	0,6138	N/A	N/A	N/A	54,53	N/A	board.21 independent.21 women-Ex.21 women-board.21 ceo-all-o-co,p-21 rep.waste.21 waste-int.21
Ridge: Dagligvaru-producenter	0,5372	N/A	N/A	N/A	60,83	N/A	Samtliga
Lasso: Dagligvaru-producenter	0,5832	N/A	N/A	N/A	51,03	N/A	board.21 ceo-all-o-comp.21 net-income/emp.21 rep.waste.21 waste-int.21 water-int.21 rep-water.21

Tabell 9: Sammanställning av resultat för samtliga modeller

Ovan finns en sammanställning av alla väsentliga modellvalideringsparametrar för samtliga modeller. När hela datamaterialet användes på samma gång resulterade detta i enbart en signifikant variabel för hela modell 1 som var signifikant med ett p-värde under 0,05.

Därefter delades materialet så att en analys utfördes för dagligvaruproducenter och en för tillverkande industribolag. I modell 2 där samtliga variabler var inkluderade för industribolag erhölls ett p-värde på 0,0503 med två variabler som var signifikant skilda från noll.
För livsmedelsproducenter i modell 4 var motsvarande siffra 0,1357, det vill säga en modell som inte är signifikant. R^2 blev även större när datan delades upp för både tillverkande industribolag (modell 2) och dagligvaruproducenter (modell 4).

Härnäst utfördes variabelselektion och modellvalidering för tillverkande industribolag och dagligvaruproducenter separat. I den egna modellen för industribolag kunde modell 3 skådas vilken var signifikant i sin helhet med tre variabler signifikant skilda från noll.

Elva beroende variabler användes med ett MSPE på 75,1195, R^2 på 0.5866 och R^2_{adj} på 0,4181. Modellen i sin helhet var även signifikant.

Vår egna modell för livsmedelsproducenter (modell 5) fick ett något lägre R^2 samt R^2_{adj} efter utförd variabelselektion, däremot med ett lägre p-värde och fler signifikanta koefficienter. MSPE landade på 85,0875. Värt att notera är även att modellerna efter utförd variabelselektion ser ut att följa antaganden för multipel linjär regression och framför allt ej lider av multikollinearitet.

För Ridge och Lasso kunde ingen nämnbar skillnad i R^2 skådas men däremot en stor förbättring i MSPE. Lasso filtrerade bort alla förutom sju oberoende variabler för såväl dagligvaruproducenter som tillverkande industribolag.

När variabelselektion utfördes på egen hand kom en modell för tillverkande industribolag med elva oberoende variabler till stånd och för livsmdelsproducenter 14 oberoende variabler. Det finns en skillnad i vilka variabler som inkluderades i modellerna. För tillverkande industribolag återfinns *non.ex.bord.ind.log* och *energy.int.per.sales.ind.log* som inte finns med i modellen för dagligvaruproducenter. I modellen för dagligvaruproducenter återfinns *tot.ceo.comp.as.tot.ex.mat.log, waste.int.per.sales.mat.log, ex.salary.bonus.tot.comp.mat, reported.waste.int.per.sales.mat, net.income.per.emp.mat* och *reported.- water.per.sales.mat* som inte finns i modellen för tillverkande industribolag. Den största märkbara skillnaden är att för dagligvaruproducenter tenderar vikten av att man faktiskt rapporterat svinn och vattenintensitet vara viktigt för att förklara ESG-betyget. Något som kan utläsas är att båda påverkar ESG-betyget negativt, vilket indikerar att de bolag som faktiskt rapporterat dessa får ett lägre ESG-betyg.

Med hjälp av Lasso identifierades två modeller för dagligvaruproducenter och tillverkande industribolag där båda har sju oberoende variabler. Även här finns det skillnader, för tillverkande industribolag ingår *independent.21, women-Ex.21* och *women-board.21* som inte finns med för dagligvaruproducenter.

Dagligvaruproducenterna innehar i sin tur *net-income/emp.21, water.int.21* och *rep-water.21.* som inte finns med för de tillverkande industribolagen.

Det finns alltså en variation i vilka variabler som för dessa bolag tycks kunna förklara ESG-betyget år 2021 med olika hög precision.

6 Diskussion

ESG-betyget är ett till synes vitt fält med många olika tolkningar och indexeringar från diverse självutnämnda expertinstitutioner. Företag och investerare har således haft svårt att navigera denna uppsjö av information då vissa bolag valt att redovisa för vissa nyckeltal och andra inte, vilket minst sagt utmanat betygets trovärdighet. EU har därav från och med januari 2023 lagstadgat krav för hur företag ska redovisa för diverse nyckeltal för att möjliggöra en större transparens med målsättningen om noll koldioxidutsläpp år 2050. Således har syftet med denna uppsats utmynnats till att undersöka ifall det föreligger några skillnader mellan ESG-betyget för dagligvaruproducenter och tillverkande industribolag under 2021 i Europa.

Uppsatsen har varit begränsad till Bloombergs dataregister vilket begränsat författarna till endast 62 observationer. Med så få observationer, 39 inom tillverkade industribolag och 23 för dagligvaruproducenter, föreligger det en stor risk för fel i modellerna. Det skulle potentiellt kunna argumenteras för både overfitting och underfitting vilket skulle kunna vara fallet i modell 4. Dock har Ridge- och Lasso-teknikerna, ut efter bästa förmåga, försökt filtrera bort multikollinearitet och anpassat modeller där man balanserar för komplexitet och förklaringsgrad. Vidare har modellframställningen baserats på en top-down princip.

Inledningsvis har Bloombergs dataregister, oberoende bransch, en mängd olika variabler denne kollar efter vilket kan skapa en masseffekt vid betygframställningen. Frågan man bör ställa sig är hur företag och investerare bör förhålla sig till tolkandet av ett ESG-betyg. Vilka åtgärder ska man som företag vidta för att erhålla ett högre betyg och hur viktar Bloomberg de olika förklarande variablerna. Vid modellframställning var förväntningen att majoriteten av de förklarande variablerna skulle uppvisa signifikans, men det visade sig för den första modellen att det endast var en förklarande variabel, *reported.waste.int.per.sales,* som uppvisade signifikans. Vidare för uppdelningen av datasetet blev det för den andra modellen, tillverkande industribolag, endast två signifikanta variabler, *reported.waste.int.per.sales.ind* samt *independent.ind*, och för dagligvaruproducenter, noll signifikanta variabler. Trots detta erhöll både modellerna relativt höga förklaringsgrader. Ovanstående kan således styrka fenomenet om masseffekt av de 21 variabler vi utfört analysen på.

Av vad som kan konstateras mellan dagligvaruproducenter och tillverkande industribolag, både för Lasso- och Ridge-teknikerna samt de egna modellerna, är att olika variabler förklarar ESG-betyget inom respektive bransch. Betygsframställningen kan tolkas vara styrande där signifikanta variabler ändras efter hand i takt med att strukturförändringar sker inom branscher.

Inom dagligvaruproducenter var bland annat de förklarande variablerna *water.int.21* och *rep-water.21* en del av Lasso-modellen, vilket de inte var för tillverkande industribolag. Variabeln *rep-water.21* är en egengjord dummyvariabel där vikten av det negativa inflytandet denne fick på modellen observerades, vilket kan ge en indikation på varför många andra bolag inte valt att redovisa detta för att undgå en sämre rating. Vidare påvisar variablerna *women-Ex.21* och *women-board.21* inflytande för Lasso-modellen för tillverkande industribolag, men inte dagligvaruproducenter. Detta kan bero på att daglivaruproducenter redan erhåller en relativt hög grad av kvinnlig representation i jämförelse med tillverkande industribolag. Om dagligvaruproducenter skulle vilja erhålla ett högre ESG-betyg kan detta göra det missvisande ifall de väljer att ta in en större andel kvinnlig representation på styrelsenivå. Sistnämnda kan fortsatt påvisa hur betyget är styrande och ändras efterhand beroende på bransch vilket gör det utmanande för företag att navigera i vilka nyckeltal som måste förbättras för att erhålla ett högre betyg.

Vidare är det värt att notera hur uppdelningen av datasetet för tillverkande industribolag och dagligvaruproducenter gjorde att dagligvaruproducenter, med endast 23 observationer erhöll den bästa prediktionsförmåga med ett MSPE om 51,03 med Lasso-tekniken. Av samtliga modeller var båda Lasso-modellerna för respektive dataset de som erhöll en totalt sett högre prediktionsförmåga än Ridge-modellerna.

Hur Bloomberg väljer att vikta deras olika förklarande variabler är fortsättningsvis ett frågetecken, dock kan variablernas signifikans och koefficienternas storlek ge en indikation på hur viktningen sett ut för den branschen. Med tanke på att Bloomberg inte har offentligt gått ut med hur de viktar ESG-betyget har det varit intressant att undersöka vilka variabler som har en påverkan ur ett statistiskt perspektiv.

Givet utmaningen med hanteringen av datan och filtreringen av de förklarande variablerna och hela underkategorin hälsa, säkerhet och miljö missar modellerna en aning information. Angående imputationen är det sannolikt att denna bidragit till en högre varians och bias vilket gör analysen mindre tillförlitlig. Detta faller sig naturligt då vi för en del av datan enbart jobbar med skattningar av det riktiga värdet.

AIC är ett givande verktyg när man ska utvärdera en modells prediktionsförmåga. För att styrka analysen och ytterligare motivera Ridge- och Lasso-modellernas prediktionsförmåga hade det varit bra att inkludera detta mått för att jämföra dessa med resterande modeller.

Denna uppsats har gett upphov till flera forskningsområden som skulle vara intressanta att undersöka vidare. Exempelvis skulle man kunna kolla på fler branscher och försöka se om det finns andra intressanta skillnader i vad ESG-betyget härstammar från. Det hade även varit intressant att jämföra olika betygskällor för att se om det finns skillnader och/eller likheter mellan de olika ratinginstituten angående hur betyget blivit framtaget. Vi har avgränsat oss till att uteslutande använda en rad olika analystekniker relaterat till regressionsanalys, däremot finns det en uppsjö av diverse statistiska metoder som möjligtvis kan lämpa sig bättre och/eller bidra med nya perspektiv i analysen.

Fördelarna med att använda en multipel linjär regression är att den till synes är en enkel modell att förstå, använda och analysera. Den ger en bra översikt över vilka förklarande variabler som erhåller störst påverkan på responsvariabeln, vilket ansetts vara en viktigt faktor i utfallet av denna uppsats. Dock förutsätter denna modell att det finns en linjär relation mellan responsvariabeln och förklaringsvariablerna, vilket kan ses som en begränsning då det skulle kunna återfinnas en mer komplex relation mellan dessa variabler. Därför finns möjligheten för framtida studier att undersöka samma frågeställning med en mixed effect modell för att kunna ta hänsyn till en mer komplex relation mellan respons- och förklaringsvariablerna där både fasta och slumpmässiga effekter kan analyseras. Något som även skulle kunna undersökas är ifall Elastic Net skulle kunna användas för att modellera betyget, vilket är en kombination av Ridge och Lasso. En annan synvinkel vi saknar är att vi enbart analyserat data över ett år. En jämförelse över flera år hade möjliggjort att se om det finns skillnader över tid i hur betyget tas fram och om det skett olika grader av kulturell mognad mellan branscher. Exempelvis om en viss bransch ligger mer i framkant i att ha en jämställd arbetsplats och hur detta påverkar betyget. Som vi tidigare diskuterat hade en liknande analys kunnat styrkas med hjälp av att inkludera fler observationer. Med större stickprov fås högre styrka och ett mer tillförlitligt resultat. För att få större insyn i hur processen går till för datainsamlingen vid beräkning av ESG-betyget utförs skulle en kvalitativ forskningsansats kunna lämpa sig väl. Att till exempel utföra intervjuer med beslutsfattare inom instituten skulle kunna ge en djupare förståelse för ämnet.

7 Slutsats

Uppsatsens frågeställning ställdes enligt följande:

> *Föreligger det någon skillnad i vilka variabler som förklarar ett företags ESG-betyg mellan branscherna dagligvaruproducenter och tillverkande industribolag?*

Av vad som kan konstateras enligt uppsatsens resultat är att det föreligger skillnader mellan vilka variabler som förklarar ESG-betyget mellan branscherna dagligvaruproducenter och tillverkande industribolag i Europa för 2021. ESG-betyget kan konstateras som en styrande indexering vilken kan komma att förändras beroende på branschens olika nivåer av nyckeltal. Med tanke på att det föreligger en ovetskap gällande hur ESG-betyget viktas externt kan det tolkas som en svart låda. För att exemplifiera valdes de förklarande variablerna *women-Ex.21* och *women-board.21* för Lasso-tekniken inom tillverkande industribolag, men inte inom dagligvaruproducenter. För dagligvaruproducenter var det bland annat *water.int.21* och *rep-water.21* som valdes, men inte för tillverkande industribolag. Detta kan påvisa någon form av mognad för respektive variabel inom branscherna vilket betyget tar i beaktning. Vidare kan en masseffekt observeras gällande betygsättningen där masseffekten består av olika variabelkombinationer för respektive bransch. Inledningsvis förväntades samtliga oberoende variabler erhålla signifikans vid modellframställningen, men det visades sig endast vara en handfull oberoende variabler, och dessutom skillnader mellan vilka dessa var inom respektive bransch.

8 Referenser

Allen, M.P. (1997). Understanding Regression analysis, New York: Plenum Press, Tillgänlig online: https://link.springer.com/book/10.1007/b102242 [Hämtad 4 januari 2023]

Bloomberg. (2022). Bloomberg Intelligence ESG [Hämtad 4 januari 2023]

Bommae, K. (2015). Understanding Diagnostic Plots for Linear Regression Analysis, Tillgänglig online: https://data.library.virginia.edu/diagnostic-plots/ [Hämtad 4 januari 2023]

Bühlmann, P. & van de Geer, S. (2011). Statistics for High-Dimensional Data, Springer Series in Statistics, [e-book] Berlin: Springer Berlin Heidelberg, Tillgänglig Online: https://link.springer.com/content/pdf/10.1007/978-3-642-20192-9.pdf?pdf=button [Hämtad 4 januari 2023]

European Commission. (2022a). Overview of sustainable finance, Tillgänglig online: https://finance.ec.europa.eu/sustainable-finance/overview-sustainable-finance_en [Hämtad 4 januari 2023]

European Commission. (2022b), EU Taxonomy for sustainable activities, https://finance.ec.europa.eu/sustainable-finance/tools-and-standards/eu-taxonomy-sustainable-activities_en [Hämtad 4 januari 2023]

European Commission. (2019). A European Green Deal, Tillgänglig online: https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en [Hämtad 4 januari 2023]

Ford, C. (2015). Understanding Q-Q Plots, Tillgänglig online: https://data.library.virginia.edu/understanding-q-q-plots/ [Hämtad 4 januari 2023]

Hardy, M. A. (1993). Regression with Dummy Variables, [e-bok] Newbury Park: Sage Publications, Tillgänglig online: https://books.google.se/books?hl=sv&lr=&id=EzLffJlYISEC&oi=fnd&pg=PP7&dq=dummy+variable&ots=RarRfVtlUy&sig=UBh_XcptQ5n_gLLG4rJs2qBSBq4&redir_esc=y#v=onepage&q=dummy%20variable&f=false [Hämtad 4 januari 2023]

Hilt, D.E. & Seegrist, D.W. (1977). Ridge, a computer program for calculating ridge regression, Upper Darby: Northeastern Forest Experiment

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R, New York: Springer

Kroese, D. P., Brereton, T., Taimre, T. & Botev, Z. I. (2014). Why the Monte Carlo Method Is so Important Today, *Wiley Interdisciplinary Reviews*: Computational Statistics, [e-journal] vol. 6, no. 6, pp.386–392, Tillgänglig Online: https://wires.onlinelibrary.wiley.com/doi/10.1002/wics.1314 [Hämtad 4 januari 2023]

Körner, S. & Wahlgren, L. (2016). Tabeller och Formler för Statistiska Beräkningar, Lund: Studentlitteratur

Mittag, N. (2013). Imputations: Benefits, Risks and a Method for Missing Data, unpublished, Harris School Of Public Policy, Univeristy of Chicago

OECD. (2013). Glossary of Statistical Terms - Cold Deck, Tillgänglig online: https://stats.oecd.org/glossary/detail.asp?ID=3377 [Hämtad 4 januari 2023]

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors, *Quality & Quantity*, [e-journal] vol. 41, no. 5, pp.673–690, Tillgänglig Online: https://link.springer.com/article/10.1007/s11135-006-9018-6 [Hämtad 4 januari 2023]

Parode, I. (2018). Residuals vs Fits Plot, Tillgänglig online https://online.stat.psu.edu/stat462/node/117/ [Hämtad 4 januari 2023]

Picard, R. R. & Cook, R. D. (1984). Cross-Validation of Regression Models, *Journal of the American Statistical Association*, [e-journal] vol. 79, no. 387, pp.575–583, Tillgänglig Online: https://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10478083 [Hämtad 4 januari 2023]

Porta, M. S. (2016). A Dictionary of Epidemiology, New York: Oxford University press

SCB. (2017). Imputera - att ersätta saknade värden, Tillgänglig online: https://www.scb.se/hitta-statistik/artiklar/2017/Imputera--att-ersatta-saknade-varden/ [Hämtad 4 januari 2023]

Sheather, S.J. (2009). A Modern Approach to Regression with R, New York: Springer

Subrahmanya, S. (2018). Missing Data Imputation using Regression. web blog post, Tillgänglig online: https://www.kaggle.com/code/shashankasubrahmanya/missing-data-imputation-using-regression/comments [Hämtad 4 januari 2023]

Taboga, M. (2021). Ridge regression, Lectures on probability theory and mathematical statistics, *Kindle Direct Publishing*, [online appendix], tillgänglig online: https://www.statlect.com/fundamentals-of-statistics/ridge-regression [Hämtad 4 januari 2023]

West, R. M. (2021). Best Practice in Statistics: The Use of Log Transformation, Annals of Clinical Biochemistry: *International Journal of Laboratory Medicine*, [e-journal] vol. 59, no. 3, pp.162–165, Tillgänglig online: https://journals.sagepub.com/doi/full/10.1177/00045632211050531 [Hämtad 4 januari 2023]

Wicklin, R. (2017). Mean imputation in SAS, web blog post, Tillgänglig online: https://blogs.sas.com/content/iml/2017/12/04/mean-imputation-sas.html [Hämtad 4 januari 2023]

Xu, Q.-S. & Liang, Y.-Z. (2001b). Monte Carlo Cross Validation, *Chemometrics and Intelligent Laboratory Systems*, [e-journal] vol. 56, no. 1, pp.1–11, Tillgänglig online: https://www.sciencedirect.com/science/article/pii/S0169743900001222 [Hämtad 4 januari 2023]