Stockholms universitet

Multivariat Multipel Regressionsanalys av försäljning på Preemmackar

Jonathan Jansson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2013:4
Matematisk statistik
Juni 2013

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2013:4**,
http://www.math.su.se/matstat

Multivariat Multipel Regressionsanalys av försäljning på Preemmackar

Jonathan Jansson*

Juni 2013

Sammanfattning

I det här arbetet använder vi oss av multivariat multipel regression för att undersöka vilka variabler som påverkar den totala försäljningen (uppdelat i butiks och bensinförsäljning) hos bemannade Preemmackar. Vi testar speciellt om skillnaden i bensinpriser mellan Preemmackar och konkurrentmackar har någon effekt på Preems försäljning och av vilket slag denna effekt i så fall är. Vi slutar med tre modeller som beskriver den totala försäljningen, där valet av mack, veckodag och tidsperiod finns med i alla modellerna. Analysen visar även på att skillnader mellan mackars bensinpriser har signifikant effekt på den totala försäljningen, där ett lägre pris relativt konkurrenterna medför en högre total omsättning. Resultaten från denna analys kan (och bör) dock ifrågasättas då alla de framtagna modellerna innehåller ett flertal outliers som påverkar deras anpassning.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: jonathan.jansson@outlook.com. Handledare: Martin Sköld.

Abstract

In this report we use *multivariate multiple regression* to check which variables having effect on the total sale in Preem gasoline stations (divided into store and gasoline sale). We test if the difference in gasoline prices between Preem's stations and competitor stations has a significant effect on the total sale and if so, what kind of effect it has. We end up having three models that describes the sale, where the choice of station, weekday and time-period is part of each model. The analysis also shows us that the difference in gasoline price between stations has a significant effect on the total sale and that a lower price with respect to competitors gives a higher all in all sale. The results of this analysis should be interpreted with caution since all the models in the report contains outliers influencing the fit of the model.

Förord

Detta arbete utgör mitt examensarbete om 15 högskolepoäng vid Stockholms Universitet. Jag vill rikta ett stort tack till min handledare Martin Sköld på Stockholms Universitet för den hjälp och det stöd han har givit under arbetets gång. Vidare vill jag även rikta ett stort tack till Marcus Larson på Preem för den tid och energi han lagt ner för att göra arbetet möjligt.

Innehåll

1 Inledning **6**

2 Beskrivning av data **7**

3 Teori **10**

3.1 Multivariat multipel regression 10

3.2 Wilks Lambda 10

3.2.1 Testa alla parametrar i modellen 11

3.2.2 Testa ett urval av parametrar i modellen 11

3.2.3 Wilks lambda som anpassningsmått 12

3.3 B-Spline 12

4 Analys **13**

4.1 Samspel 13

4.2 Sökandet efter modeller 14

4.2.1 Metod 1 15

4.2.2 Metod 2: Forward selection och backward elimination . . 15

4.3 Kontroll och förbättring av modeller 16

4.3.1 Tidsberoende mönster 16

4.3.2 Outliers i modellerna 18

4.4 Signifikanta differensvariabler 20

5 Diskussion **22**

6 Slutsats **23**

7 Appendix **24**

1 Inledning

Preem är Sveriges största drivmedelsbolag och en av aktörerna i det svenska bensinoligopolet. I och med att de är en av få aktörer på marknaden innebär det att skillnader mellan deras och konkurrenternas bensinpriser kan ha stor effekt på drivmedelsförsäljningen och företagets vinster. I det här arbetet studerar vi en mer omfattande försäljning (innefattande både bensinvolym och butiksförsäljning), där målet är att undersöka vilka variabler som påverkar denna försäljning.

Mer specifikt vill vi beskriva försäljningen för bemannade Preemmackar med multivariat multipel regressionsanalys, där försäljningen är uppdelad i butiksomsättning och såld volym 95-oktanig bensin. Slutligen vill vi även testa om någon differensvariabel mellan bensinpriset för Preems mackar och närliggande bemannade konkurrentmackar har effekt på försäljningen.

2 Beskrivning av data

Drivmedelsbolaget Preem har bidragit med all data som används i arbetet. Datamaterialet som analyseras innehåller totalt 2051 observationer utspritt över 4 månader (oktober-januari) och 24 bemannade Preemmackar.

För varje Preemmack finns ett "konkurrensområde" där mackar inom området uppskattas konkurrera om samma kunder. De 24 mackarna har valts ut så att alla mackar har minst en bemannad konkurrent i sitt konkurrensområde. Detta så att det går att skapa, för studien, intressanta variabler baserade på skillnader mellan bensinpriser hos bemannade konkurrentmackar och Preemmackar. Obemannade mackar har i regel ett mycket lägre bensinpris än bemannade mackar. Det medför att den huvudsakliga variationen hos differensvariabler som även inkluderar obemannade mackar skulle förklaras av förekommandet av de obemannade mackarna.

För varje observation finns data för de 13 olika variablerna som listats i tabellerna nedan.

Beroende variabler

	Variabel	Förkortning	Beskrivning
1	Y1	Volym	Total volym 95-oktanig bensin såld per dag
2	Y2	Butik	Totalsumma av försäljning i butik mätt i kronor (exklusive drivmedelsförsäljning)

Oberoende variabler

	Variabel	Förkortning	Beskrivning
3	P1	pris	Medelvärde på det egna bensinpriset
4	P2	prisH	Högsta egna bensinpriset
5	P3	prisL	Lägsta egna bensinpriset
6	X1	dag_nr	Antal dagar efter 1:a Januari
7	X2	vdag	Veckodag som observationen tillhör
8	X3	mack	Mack som observationen tillhör
9	X4	konk	Antalet konkurrenter i närområdet
10	X5	auto	Indikatorvariabel för obemannade konkurrenter i mackens närområde, variabeln antar värdet 1 om det finns någon obemannad konkurrent och 0 annars
11	D1	mindiff	Differens mellan lägsta egna bensinpriset och det lägsta priset observerat bland de bemannade konkurrenterna
12	D2	meanmindiff	Differens mellan medelpriset på den egna bensinen och det lägsta konkurrentpriset bland bemannade konkurrentmackar
13	D3	meandiff	Differens mellan medelpriset på den egna bensinen och medelvärdet för konkurrentpriserna bland bemannade konkurrentmackar

Variablerna X1, X2, X4, X5, D1, D2 och D3 är konstruerade på egen hand medan variablerna P1, P2, P3 och X3 kommer direkt från den ursprungliga datan.

Tidsvariabeln X1 inför vi för att kunna fånga tidsskillnaden mellan observationer i en variabel istället för att betrakta separata datum uppdelat i år, månad och dag. Den andra tidsvariabeln X2 konstrueras istället för att beskriva veckovist återkommande effekter. Variablerna X4 och X5 beskriver båda hur konkurrensen ser ut vid de olika Preemmackarna. Variabeln X4 införs för att kunna beskriva effekten som antalet konkurrenter har på försäljningen och X5 för att beskriva effekten av att obemannade konkurrentmackar finns i området. Båda dessa variabler är givna för varje separat mack och X4 och X5 är alltså linjära kombinationer av variabeln X3. Det är därmed meningslöst att införa någon av variablerna X4 eller X5 i en modell som redan innehåller variabeln X3 då dessa effekter redan är fångade i mack-variabeln.

Datan över konkurrentpriserna är baserat på ett fåtal observationer och är angivet som ett medelvärde per konkurrent och dag. Hur många observationer eller när på dagen observationerna är gjorda framgår dock inte av datan. Därmed konstruerar vi flera differensvariabel för större chans att få en variabel som kan beskriva effekten av prisskillnaderna på försäljningen.

Det ursprungliga datamaterialet bestod egentligen av 2059 observationer men 8 stycken har tagits bort. Att dessa punkter tas bort beror på att de antas vara felaktigt angivna då såld bensin- eller total butiksförsäljning i de fallen är noll (eller mycket nära noll).

3 Teori

3.1 Multivariat multipel regression

Antag att vi har två responsvariabler Y_1 och Y_2 och vi vill beskriva dessa med k förklarande variabler $X_1, ..., X_k$. Vi kan göra detta med en multivariat multipel regressionsmodell $\mathbb{Y} = \mathbb{X}\beta + \Xi$. Givet att vi har n antal observationer kan vi skriva denna formel på matrisform som

$$\begin{pmatrix} y_{11} & y_{12} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \alpha_{01} & \alpha_{02} \\ \beta_{11} & \beta_{12} \\ \vdots & \vdots \\ \beta_{k1} & \beta_{k2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \vdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} \end{pmatrix}$$

där det gäller att:
1. $\mathrm{E}[\mathbb{Y}\,] = \mathbb{X}\beta \Leftrightarrow E[\Xi] = \mathbb{O}$
2. $cov(\mathbb{Y}) \equiv cov(Y_1, Y_2) = \Sigma$
3. Raderna i $\mathbb{Y}$ är oberoende
4. Feltermerna (ε) är normalfördelade

Parametermatrisen β skattas här med maximum likelihood-metoden enligt formeln $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}$. Då det enda som skiljer skattningen av β i den multivariata modellen från den i det multiplafallet är dimensionen på $\mathbb{Y}$-matrisen blir de desamma som om vi hade utfört multipel linjär regression för Y_1 respektive Y_2 separat. Det som skiljer de separata multipla modellerna från den multivariata modellen är att vi i det senare tar hänsyn till korrelationen mellan responsvariablerna när modeller utformas och testas.

För mer om multivariata multipla modeller se Rencher & Christensen (2012 kapitel 10)

3.2 Wilks Lambda

När vi testar parametrar i den multivariata modellen kan vi inte använda oss av F-test och testa modellerna för responsvariablerna separat då de inte tar hänsyn till korrelationer mellan responsvariablerna. F-test kan användas när testet endast omfattar parametrar i en kolumn av betamatrisen. För att testa hypoteser i den multivariata modellen använder vi istället Wilks Λ-test som är modellens likelihood-kvot-test och en allmännare form av F-test.

3.2.1 Testa alla parametrar i modellen

Antag nu att vi vill testa hypotesen H_0: "Alla parametrar i β (förutom interceptet) är noll" mot hypotesen H_1: "Minst en av dessa parametrar är skild från noll".

Vi inför då först matriserna E och $E+H$, där $E = \mathbb{Y}'\mathbb{Y} - \hat{\beta}'\mathbb{X}'\mathbb{Y}$ är här modellens "residual sum of squares and products"-matris och $H + E = (\hat{\beta}'\mathbb{X}'\mathbb{Y} - n\bar{y}\bar{y}') + (\mathbb{Y}'\mathbb{Y} - \hat{\beta}'\mathbb{X}'\mathbb{Y}) = \mathbb{Y}'\mathbb{Y} - n\bar{y}\bar{y}'$ är modellens "total sum of squares and products"-matris, $\bar{y}$ är en 1 x 2-matris innehållande medelvärdena för Y_1 respektive Y_2.

Teststatistikan kan då skrivas som $\Lambda_{tot} = \dfrac{|E|}{|E+H|} \sim \Lambda_{2,k,n-k-1}$ som under H_0 har en Wilks lambda-fördelning med parametrarna $(2, k, n-k-1)$ där 2 står för antalet responsvariabler, k för antalet förklarande variabler och n för antalet observationer.

3.2.2 Testa ett urval av parametrar i modellen

Om vi istället vill testa hypotesen att en delmängd av parametrarna i β är noll delar vi först upp matrisen i två delar, $\beta = \begin{pmatrix} \mathbb{B}_r \\ \mathbb{B}_d \end{pmatrix}$, där $\mathbb{B}_d$ är den h x 2-matris innehållande de parametrar vi vill testa. De vi vill testa är alltså nollhypotesen H_0 : "alla parametrar i $\mathbb{B}_d$ är noll ($\mathbb{B}_d = \mathbb{O}$)" mot alternativhypotesen H_1: "minst en parameter i $\mathbb{B}_d$ är skild från noll ($\mathbb{B}_d \neq \mathbb{O}$)".

Givet att $\mathbb{B}_d$ är en h x 2-matris blir $\mathbb{B}_r$ en $(k-h+1)$ x 2 -matris innehållande de parametrar som inte testas. Definierar vi $\mathbb{X}_r$ som datamatrisen tillhörande $\mathbb{B}_r$ så kan vi bilda $\Lambda^* = \dfrac{|E^*|}{|E_* + H_*|}$ där $E^* = \mathbb{Y}'\mathbb{Y} - \hat{\mathbb{B}}_r'\mathbb{X}_r'\mathbb{Y}$, $H^* = \hat{\mathbb{B}}_r'\mathbb{X}_r'\mathbb{Y} - n\bar{y}\bar{y}'$ och $\hat{\mathbb{B}}_r$ är likelihood-skattningarna av $\mathbb{B}_r$ under H_0. En teststatistika med Wilks Lambda-fördelning kan då skrivas som $\Lambda_{red} = \dfrac{\Lambda_{tot}}{\Lambda^*} \sim \Lambda_{2,h,n-k-1}$. Ju mindre teststatistika vi får desto större anledning har vi att förkasta H_0.

För att lättare kunna bedöma om det är lämpligt att förkasta H_0 kan Λ_{red} konverteras till en statistika som är F-fördelad. Då antalet responsvariabler (p) är 2 går detta att göras exakt för $\Lambda \sim \Lambda_{2,h,n-k-1}$ med formeln $F = \dfrac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \times \dfrac{n-k-2}{h}$ där F är F-fördelad med frihetsgraderna $2 \times h$ och $2 \times (n-k-2)$.

3.2.3 Wilks lambda som anpassningsmått

Wilks lambda för den totala modellen kan även används som ett anpassningsmått för multivariata modeller och i det här arbetet används måttet uteslutande i detta syfte. Ett lägre värde på Λ visar här på en bättre modellanpassning.

För mer om Wilks Lambda se Rencher & Christensen (2012 kapitel 6 10).

3.3 B-Spline

Om feltermerna i en regressionsmodell har tydliga mönster då de plottas mot en viss variabel x tyder det på att effekten av den variabeln inte fångats på ett bra sätt av modellen. För att rätta till det här problemet kan man införa olika polynom av variabeln x i modellen. Ibland kan denna metod dock medföra ganska omständiga modeller med många olika polynom av hög grad. Ett alternativ till införandet av en mängd olika polynom av x är att istället införa en funktion av segmentsvis sammanlänkade polynomfunktioner för olika intervall på variabeln x. En spline är just en sådan funktion som även är kontinuerlig och deriverbar i alla sammanlänkningspunkter (noder).

En fördel med att införa en spline framför olika polynom av x är att de olika polynomen inte lappar över varandra i modellen. Detta gör att en spline lättare kan anpassas till modeller där de är olika mönster för olika intervall på x. En B-Spline är särskilt lämpad för sådana modeller då, till skillnad från splines i allmänhet, ändringar i anpassningen i en del av funktionen inte påverkar hela kurvans anpassning.

En B-spline i en regressionsmodell kan allmänt skrivas på formen $y = b_0B_0(x) + b_1B_1(x) + ... + b_kB_k(x) + \epsilon$ där b_i är en konstant och B_i är ett polynom i variabeln x som båda anpassats efter modellen.

För mer om splines se Maindonald Braun (2010, kapitel 7.5) eller Seber & Lee (2003, kapitel 7.22). .

4 Analys

För att undvika problem orsakade av stark korrelation mellan modellens förklarande variabler så inleder vi med att undersöka dessa korrelationer. Vi skapar därför en korrelationsmatris med de icke-kategoriska variablerna X1, P1, P2, P3, X5, D1, D2 och D3, se Figur 1 nedan.

Figur 1: Korrelationsmatris för de förklarande variablerna

I matrisen ser vi att det finns en stark korrelation mellan prisvariablerna P1, P2 och P3 respektive mellan prisdifferensvariablerna D1, D2 och D3. Således bör vi undvika modeller innehållande två eller fler P- eller D-variabler, då de förklarar samma sak.

4.1 Samspel

Ett naturligt och teoretiskt mycket troligt antagande är att veckodagsvariabeln X2 och mackvariabeln X3 förklarar det mesta av variationen i Y. Vi kan se att så är fallet i separata ANOVA-tabeller för responsvariablerna nedan.

```
 Response butik (Y1) :
             Df     Sum Sq    Mean Sq F value    Pr(>F)
x2            6 9.4325e+09 1572088304  112.96 < 2.2e-16 ***
x3           23 1.0664e+11 4636482931  333.16 < 2.2e-16 ***
Residuals  2021 2.8126e+10   13916649
```

```
------------------------------------------------------------

 Response volym (Y2) :
              Df     Sum Sq   Mean Sq F value    Pr(>F)
x2             6  140688517  23448086   70.11 < 2.2e-16 ***
x3            23 2666612736 115939684  346.66 < 2.2e-16 ***
Residuals   2021  675919046    334448
```

X2 och X3 förklarar över 80% av variationen i både butiks- och volymvariabeln (där X3 beskriver den största delen av variansen i båda fallen). Det kan därför vara av intresse att undersöka om det existerar någon samspelseffekt mellan dessa som bör tas med i fortsatt analys.

För att testa om vi har någon signifikant samspelseffekt testar vi nollhypotesen: "alla samspelsparametrar är noll" mot alternativhypotesen: "minst en av parametrarna är skild från noll" med ett Wilks lambda-test, vi får då ett p-värde som är mindre än 2.2×10^{-16}. Det finns alltså någon samspelseffekt som är intressant för modellen, men då det totala antalet samspelsvariabler mellan X2 och X3 är $24 \times 7 = 168$ stycken försöker vi krympa ner antalet en aning. Vi gör två försök till detta genom att skapa variablerna:

1) "Mackvdag" som innehåller alla kombinationer av den ursprungliga samspelsvariabeln, vilkas skattade koefficienter signifikant skiljer sig från noll på en 95% nivå.

2) "Mackhelg" som innehåller alla signifikanta kombinationer av samspelet mellan mack och helg på en 95% nivå.

Båda är signifikanta för modellen (se Appendix för MANOVA-tabeller) med ett p-värde under 2.2×10^{-16} och med Wilks lambda-värden på 0.76903 respektive 0.80935. Men då Mackvdag innehåller 32 olika nivåer och Mackhelg endast innehåller 5 väljer vi att använda variabeln Mackhelg i den fortsatta analysen.

4.2 Sökandet efter modeller

För att finna lämpliga modeller att beskriva vår responsvariabel med använder vi oss av två tillvägagångssätt. I det första utgår vi från modeller med Y_1 respektive Y_2 som univariata responsvariabler och går igenom ett större antal modeller för att finna bra gemensamma modeller. Den andra metoden vi tillämpar är att använda forward selection och backward elimination direkt på de multivariata modellerna.

4.2.1 Metod 1

Då prisvariablerna P={P1, P2, P3} och differensvariablerna D={D1, D2, D3} båda är grupper innehållande starkt korrelerade variabler är det inte intressant att undersöka modeller med fler än en variabel från respektive grupp. Då blir antalet beskrivande variabler som kan kombineras fritt med varandra $12-2\times3=6$ och därmed antalet modeller som datorn går igenom $4\times4\times\sum_{i=0}^{6}\binom{6}{i}=1024$ med Y_1 respektive Y_2 som univariata responsvariabler. Observera att vi här också tar med den uttunnade samspelsvariabeln "Mackhelg".

De två modeller med bäst anpassningsmått (R^2) med samma antal förklarande variabler tas fram för respektive responsvariabel och intressant kombination av P- och D-variabler. Unionen av dessa modellers variabeluppsättningar jämförs sedan som multivariata modeller med Wilks Lambda som anpassningsmått. De modeller där alla variabler har parametrar signifikant skilda från noll (på en 95% nivå) och där inga variabler är en linjär kombination av andra är listade i tabellen nedan:

Variabler i modell	(Wilks Λ)
X1,X5	0.942890
X1,X2,X5	0.870286
X2,X5,P3	0.863380
X3,P3	0.077193
X3,D1,P3	0.0770974
X2,X3,P3	0.0513526

Vi ser att både X3 och P3 är åter förekommande i de modeller med bäst anpassningsmått. Även variabeln X5 förekommer i hälften av modellerna, dock i de modeller med sämst anpassningsgrad.

4.2.2 Metod 2: Forward selection och backward elimination

Vid forward selection börjar vi med en tom modell och tar sedan vid varje steg in den variabel som bidrar till modellens Wilks Λ minst. Processen fortgår tills ingen ny variabel är signifikant på en 95% nivå. Backward elimination utförs på motsvarande sätt genom att anta en modell med alla variabler och i varje steg utesluts den variabel som sänker modellens lambda mest. Det slutar då alla variabler i modellen är signifikanta på en 95% nivå. Dessa utförs med alla förklarande variabler vilket resulterar i följande två modeller:

Metod	Variabler i modell	(Wilks Λ)
Backward	X1, X2, X3, P2, D3, Mackhelg	0.0406466
Forward	X1, X2, X3, P1, D3, Mackhelg	0.0405461

Båda processerna utfördes även utan variablerna D1, D2 och D3 då dessa modeller skulle kunna vara intressanta vid senare signifikanstest av differensvariablerna. Körningar resulterade dock endast i två modeller identiska till modellerna ovan men utan variabeln D3.

Vi ser att Metod 2 är den metod som givit modeller med lägst lambdavärden. En förklaringen till denna skillnad ligger i att modellerna från Metod 1 får färre variabler än de från Metod 2. Wilks lambda ökar nämligen inte när en ny variabel införs utan kan endast bli lägre eller förbli oförändrad. Anledningen till att Metod 1 ger modeller med färre antal variabler och utan vissa variabler som är starkt sginifikanta för den multivariata är att dessa variabler kan vara starkt signifikanta i den ena, men ej signifikanta i den andra univariata modellen. I Metod 1 skulle en sådan variabel inte komma med medan den skulle göra detta i Metod 2.

4.3 Kontroll och förbättring av modeller

Spridningen på Wilks Λ-värdena bland de framtagna modellerna är väldigt stor. Då ett lägre värde pekar på högre anpassningsgrad för den multivariata modellen väljer vi att fortsätta analysen med de modeller som har lägst Λ-värden givet antalet variabler. Det vill säga framöver kollar vi på modellerna:

Modell	Variabler i modell	(Wilks Λ)
1	X2, X3	0.077193
2	X2, X3, P3	0.0513526
3	X1, X2, X3, P1, D3, Mackhelg	0.0405461

4.3.1 Tidsberoende mönster

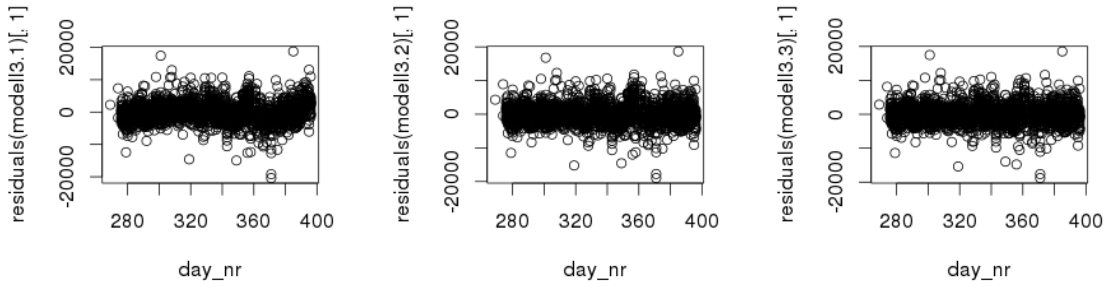
Nu plottar vi respektive modells residualer mot tidsvariabeln X1 (dag_nr). Då residualerna antas vara oberoende med väntevärde noll bör de vara slumpmässigt utspridda kring noll i dessa plottar. Vi ser dock hur residualerna för de tre modellerna har tydliga mönster kring nollinjen (se Figur 2a och 3a). Residualerna för Y_1 (butiksförsäljningen) är S-formade medan de för Y_2 (bensinförsäljningen) går i ett aningen hackigare mönster. Vidare kan vi även se en kraftig avvikelse från noll för alla plottar på intervallet [353,357] vilket motsvarar dagarna 19 - 23 december.

För att åtgärda de huvudsakliga mönstren i dessa residualer måste det införas en ny tidsberoende variabel. Vi inför därmed en B-spline ($SP(X1)$) baserad på

tidsvariabeln X1 i de tre modellerna. Den B-spline som införs är sekvensvist anpassade tredjegradspolynom med 4 interna noder,vilket förbättrar modellernas residualerna och förklaringsgrad markant (jämför bild (a) och (b) i Figur 2 och Figur 3 nedan och se appendix).
Avvikelsen i intervallet [353,357] för X1-variabeln kvarstår även efter införandet av vår spline. Tillåter vi oss att införa en ny variabel för dessa "innan jul"-dagar förbättras residualernas utseende ytterligare.

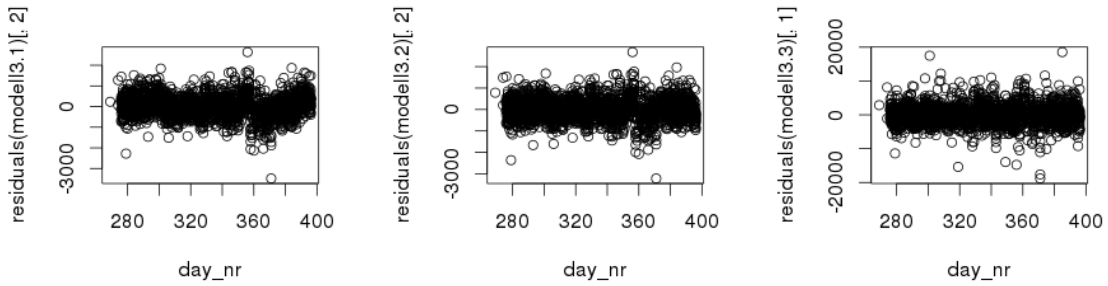
Modell	Variabler i modell	(Wilks Λ)
1	X2,X3,$SP(X1)$,JUL	0.04200921
2	X2,X3,P3,$SP(X1)$,JUL	0.04151731
3	X2,X3,P1,D3,Mackhelg,$SP(X1)$,JUL	0.03201948

Figur 2, 3 nedan visar den korrigering av residualerna som har skett för modell 3, plottar för de två andra modellerna listas i appendix. De skattade parametrarna för alla tre modeller är även de listade i appendix. Parametrarna för de variabler som är gemensamma för alla modeller är mycket liknande skattningar. Modellerna visar alla på en genomsnittligt lägre försäljning under helger och högre försäljning under julen. För alla modeller kan vi också se att de skattade parametrarna bland mackarna varierar väldigt mycket, vilket återigen visar på att Mack är den variabel som förklarar den största variationen i datan.



(a) Ursprunglig modell (b) Spline införd (c) Spline och Jul införd

Figur 2: Modell 3: Residualer för butik (Y_1) plottade mot X1

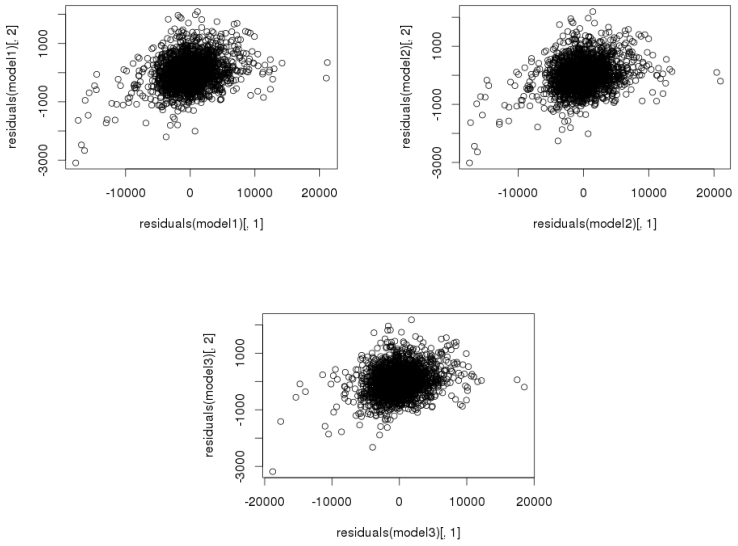


(a) Ursprunglig modell (b) Spline införd (c) Spline och Jul införd

Figur 3: Modell 3: Residualer för volym (Y_2) plottade mot X1

4.3.2 Outliers i modellerna

Om vi istället plottar residualerna för Y_1 mot residualerna för Y_2 ser vi att det finns ett stort antal outliers i alla de tre modellerna (Figur 4).

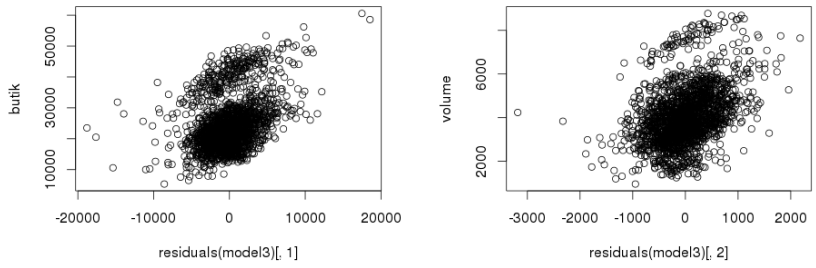


Figur 4: Residualer för Y_1 plottade mot residualerna för Y_2 för Modell 1,2 & 3

Vid närmare undersökning av dessa observationer hittas ingen gemensam avvikelse i någon variabel. De tillhör ett flertal olika mackar och veckodagar, de är utspridda över hela den tidsperiod som analysen avser och har till synes inga extrema värden på någon variabel. Det medför att vi inte har någon anledning

att utesluta observationerna och modellerna förblir med dessa brister.

Vi kan se att de utstickande observationerna har stort inflytande på modellen genom att plotta residualerna mot motsvarande responsvariabel. Här ser vi i Figur 5 (nedan) hur residualerna för modell 3 visar en tendens att gå från negativa till positiva värden när försäljningen av butiksvaror och bensin ökar. Detta är ett resultat av att ett fåtal avvikande observationer har dragit med sig regressionslinjen, vilket medför att modellen blir sämre anpassad för den övriga datan.



(a) Y_1 mot tillhörande residualer (b) Y_2 mot tillhörande residualer

Figur 5: Residualplottar för modell 3

4.4 Signifikanta differensvariabler

Avslutningsvis vill vi undersöka om någon av differensvariablerna D1, D2 eller D3 har signifikant effekt på försäljningen för Preemmackarna. Då Modell 3 redan innehåller variabeln D3 ersätts den variabeln med D1 respektive D2 när dessa variabler testas för modellen. Det visar sig att D3 (differensen mellan medelpriser) har störst signifikans i alla de tre modellerna och är signifikant med 5% felrisk.

Modell	Differensvariabel	Wilks Λ för test av differensvariabel	Approximativt P-värde
1	D1	0.99671	0.03631
1	D2	0.99672	0.03679
1	**D3**	**0.99508**	**0.006999**
2	D1	0.99729	0.06492
2	D2	0.99756	0.08589
2	**D3**	**0.99621**	**0.02200**
3	D1	0.99712	0.05489
3	D2	0.99659	0.03224
3	**D3**	**0.99544**	**0.01011**

Vi gör därför valet att undersöka inverkan av variabeln D3 vidare. För att kolla om D3 har inverkan på butiksförsäljningen (Y_1), bensinförsäljningen (Y_2) eller på båda undersöker vi de separata ANOVA-tabellerna för dessa. Sammanfattningsvis föll resultatet ut som i tabellen nedan (fullständiga tabeller finns i appendix).

Modell	D3 signifikant för Y_1 (*med 5% felrisk*)	D3 signifikant för Y_2 (*med 5% felrisk*)
1	Falskt	Sant
2	Sant	Falskt
3	Sant	Falskt

Resultatet för modell 2 och 3 skiljer sig alltså från resultatet för modell 1, vilket gör det svårt att säga huruvida D3 har en signifikant inverkan på butiks- eller bensinförsäljningen separat.

Det är dock inte nödvändigt att D3-variabeln ska ha signifikant effekt på någon av responsvariablerna separat för att den ska kunna ha en signifikant effekt i det multivariata fallet. I vårt fall är det svårt att säga om variabeln har effekt på Y_1 eller Y_2 separat, men det påverkar inte vårt tidigare resultat om att D3 har en signifikant effekt på kombinationen av butiks- och bensinförsäljning. Vi sammanfattar parameterskattningarna för D3 i de tre modellerna i nästa tabell.

Modell	Parameterskattning för D3 med Y_1 som responsvariabel	Parameterskattning för D3 med Y_2 som responsvariabel
1	-866.4	207.36
2	-1329.5	97.81
3	-1649.2	58.07

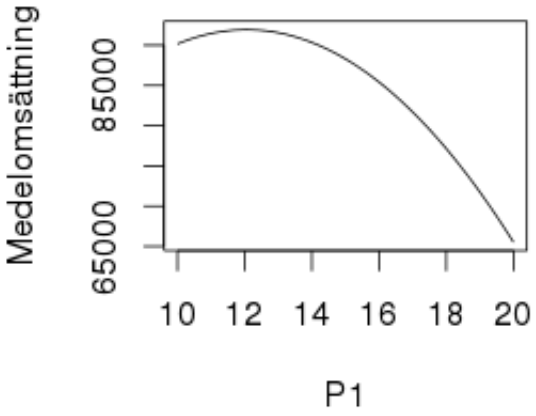
Parameterskattningarna är alltså konsekventa angående tecken genom alla tre modellerna. Det visar på att en ökning som görs i såld bensinvolym som en effekt av D3 sker på bekostnad av butiksförsäljning. En sådan effekt skulle kunna ha en delförklaring i att trängsel och köbildning, som ett resultat av prisskillnaden, i sin tur har negativ effekt på butiksförsäljningen.

Det skulle nu kunna vara av intresse att undersöka hur den totala omsättningen (i kronor) påverkas när bensinpriset ändras. Vi utgår därför från en multivariat modell innehållande variablerna D2 och P1 (som båda beror på det egna medelbensinpriset P1) och kollar på en medelomsättningsfunktion för denna. Modellen vi undersöker innehåller de förklarande variablerna X2, X3, $SP(X1)$, JUL, P1 och D2 där vi fixerar alla variabler som inte beror på P1 till några representativa värden i modellen. Vi sätter dessa variabler till X2=Tisdag, X3=Mack1, $SP(X1)$=0, jul=0, K=14.5 där K står för konkurrent priset och D2 = K-P1. Med skattningarna

```
                    butik      volym
(Intercept)      44437.6274  8940.60768
vdagTuesday      -4173.1009  -428.81681
Mack1             8595.2757  1650.45390
P1               -1636.1538  -395.03634
D2               -1221.0654    23.24090
```

kan vi skriva medelomsättningsfunktionen som $M(P1) = \textit{butiksförsäljning} + P1 \times \textit{bensinvolym} = (44438 - 4173 + 8595 - 1636 \times P1 - 1221 \times (14.5 - P1)) + P1 \times (8941 - 429 + 1650 - 395 \times P1 + 23 \times (14.5 - P1))$ som kan förenklas till $M(P1) = -418 \times P1^2 + 10084 \times P1 + 31154$. Plottar vi denna funktion för ett intervall på P1 (kring konkurrentpriset) får vi grafen i Figur 6 nedan.

Kollar vi på ändringar i P1 i närheten av konkurrentpriset 14.5 ser vi att ett lägre bensinpris medför en ökning i den totala omsättningen. Först under ett bensinpris kring 12 ser det ut som vi förlorar på att sänka priset ytterligare givet att de andra variablerna i modellen hålls konstant. Att ligga på ett bensinpris på 12 så att omsättningen blir så hög som möjligt ser här ut att vara optimalt. Det är dock orimligt att anta att konkurrentpriset förblir 14.5 vid en så stor ändring i de egna bensinpriserna, så effekterna av en sådan sänkning blir troligen bara tillfälliga.



Figur 6: Medelomsättning som funktion av P1

5 Diskussion

Försäljningen för Preemmackar är väldigt komplext att modellera, då en mängd olika variabler och samspel dem emellan inverkar. De tre slutliga modellerna beskriver en stor del av variansen i datan, där variabeln mack står för den huvudsakliga förklaringen. Alla de tre slutliga modellerna har dock outliers vilket är problematiskt då den multivariata modellen bygger på antagandet om att residualerna är normalfördelade. Modellen är alltså känslig för extremvärden vilket medför att alla resultat och slutsatser dragna ur dessa modeller bör behandlas och tolkas varsamt.

Analysen resulterade i att differensvariabel D3 ses ha signifikant inverkan på den allmänna försäljningen i alla de tre modellerna, där en ökning i bensinförsäljningen tillsynes sker på bekostnad av butiksförsäljningen. Ett lägre bensinpris givet konkurrentpriserna ser ut att kortsiktigt medföra en ökning i den totala omsättningen. Vi kan dock från den här analysen inte avgöra huruvida det görs en långsiktig vinst när priset sänks. Det kan finnas långtidseffekter i form av förändringar i marknadsandelar och därmed konkurrenskraft som inte fångas av modellen.

Resultaten tyder alltså på att det kan finnas en effekt mellan mackars bensinprisskillnader och deras totala försäljning. För att försäkra sig om något sådant resultat krävs dock vidare undersökningar. För vidare undersökningar bör differensvariablerna göras mer exakta genom fler kontroller av bensinpriser hos konkurrerande mackar. Vidare kan det vara idé att undersöka mackarna individuellt då många variabler kan antas ha olika effekt beroende på vilket område macken befinner sig i.

6 Slutsats

De modeller vi fått fram i vår analys som bäst beskriver datan i förhållande till antalet förklarande variabler är

Modell	Variabler i modell	(Wilks Λ)
1	X2,X3,$SP(X1)$,JUL	0.04200921
2	X2,X3,P3,$SP(X1)$,JUL	0.04151731
3	X2,X3,P1,D3,Mackhelg,$SP(X1)$,JUL	0.03201948

Där X3 (Mack) är den variabel som förklarar den huvudsakliga variationen och modell 3 är den modell som bäst beskriver datan.

Vid test av differensvariablerna var D3 signifikant på en 95% nivå för alla tre modellerna. Men då modellerna inte uppfyller antagandena för den multivariata modellen gällande normalfördelade residualer, och Wilks lambda-testerna bygger på att detta antagande uppfylls, bör man se skeptiskt till resultaten. För att kunna säga något säkert om effekten av differensvariablerna föreslås en vidare studie där fler kontroller per dag görs på konkurrenternas bensinpriser.

7 Appendix

Kovariansmatris Σ för de multivariata modellerna:

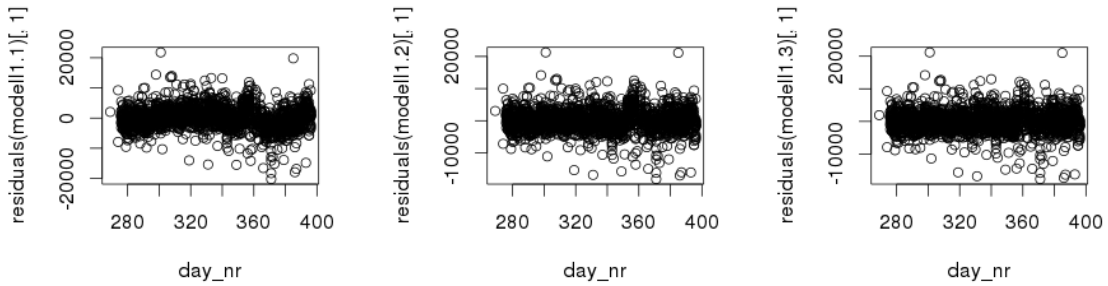
$$\text{cov}\ (\mathbb{Y}) = \Sigma = \begin{pmatrix} 70340090 & 6873744 \\ 6873744 & 1699132 \end{pmatrix}$$

MANOVA-tabeller för test av uttunnade samspelsvariabeler:

```
Type II MANOVA Tests: Wilks test statistic
         Df test stat approx F num Df den Df    Pr(>F)
x2        6   0.72119   58.853     12   3978 < 2.2e-16 ***
x3       23   0.06697  247.690     46   3978 < 2.2e-16 ***
Mackvdag 31   0.76903    9.003     62   3978 < 2.2e-16 ***
```

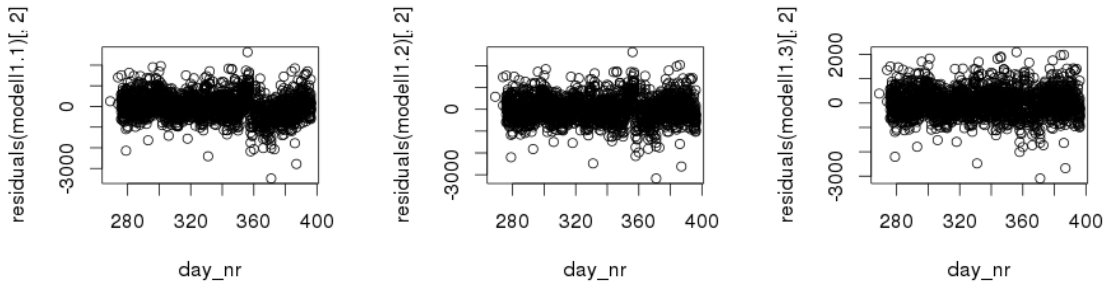
```
Type II MANOVA Tests: Wilks test statistic
         Df test stat approx F num Df den Df    Pr(>F)
x2        6   0.70539   64.059     12   4032 < 2.2e-16 ***
x3       23   0.04921  307.482     46   4032 < 2.2e-16 ***
Mackhelg  4   0.80935   56.224      8   4032 < 2.2e-16 ***
```

Övriga bilder för tidsberoende där spline och JUL variablerna införs:



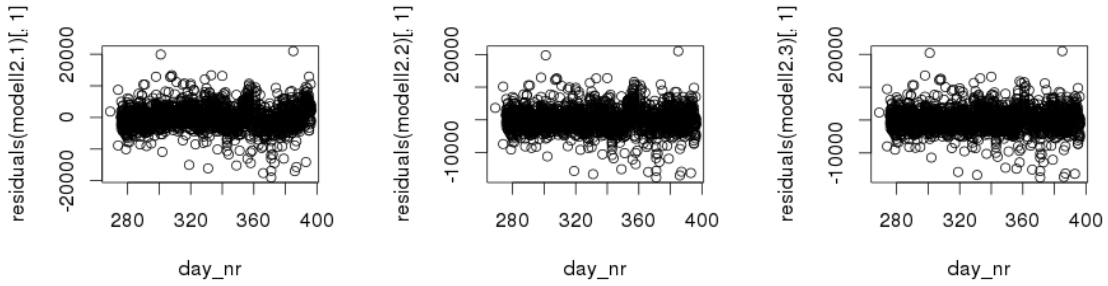
(a) Ursprunglig modell (b) Spline införd (c) Spline och Jul införd

Figur 7: Modell 1: Residualer för butik (Y_1) mot X1



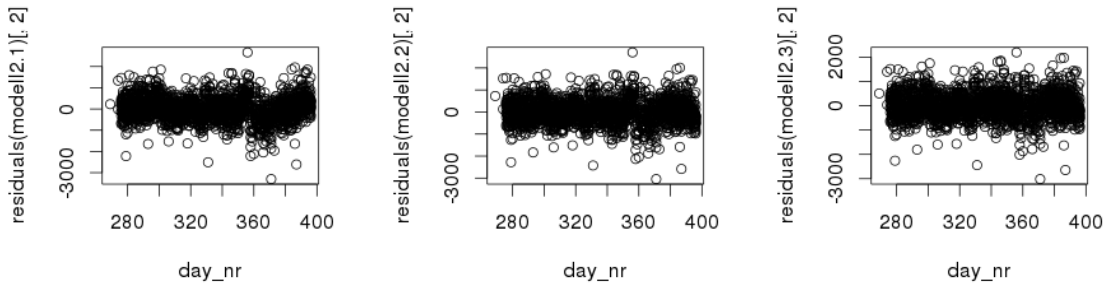
(a) Ursprunglig modell (b) Spline införd (c) Spline och Jul införd

Figur 8: Modell 1: Residualer för volym (Y_2) mot X1



(a) Ursprunglig modell (b) Spline införd (c) Spline och Jul införd

Figur 9: Modell 2: Residualer för butik (Y_1) mot X1



(a) Ursprunglig modell (b) Spline införd (c) Spline och Jul införd

Figur 10: Modell 2: Residualer för volym (Y_2) mot X1

ANOVA-tabeller för undersökning av effekten av D3 på de univariata responsvariablerna för respektive modell

Modell 1:

```
Anova Table (Type II tests)

Response: butik
                       Sum Sq   Df  F value Pr(>F)
x2                 1.0441e+10    6 146.8916 <2e-16 ***
x3                 1.0483e+11   23 384.7388 <2e-16 ***
jul                1.4325e+09    1 120.9139 <2e-16 ***
bs(day_nr, df = 5) 3.4264e+09    5  57.8449 <2e-16 ***
d3                 2.5703e+07    1   2.1696 0.1409
Residuals          2.3860e+10 2014

---------

Anova Table (Type II tests)

Response: volym
                       Sum Sq   Df  F value  Pr(>F)
x2                  142639606    6  82.9366 < 2e-16 ***
x3                 2673822822   23 405.5666 < 2e-16 ***
jul                  38967034    1 135.9423 < 2e-16 ***
bs(day_nr, df = 5)   63131451    5  44.0487 < 2e-16 ***
d3                    1472294    1   5.1363 0.02354 *
Residuals           577300708 2014
```

Modell 2:

```
Anova Table (Type II tests)

Response: butik
                       Sum Sq   Df  F value    Pr(>F)
x2                 1.0459e+10    6 147.6518 < 2.2e-16 ***
x3                 1.0462e+11   23 385.2797 < 2.2e-16 ***
p3                 9.4572e+07    1   8.0106  0.004697 **
jul                1.2307e+09    1 104.2443 < 2.2e-16 ***
bs(day_nr, df = 5) 2.5517e+09    5  43.2283 < 2.2e-16 ***
d3                 5.6164e+07    1   4.7573  0.029290 *
Residuals          2.3765e+10 2013

---------

Anova Table (Type II tests)
```

```
Response: volym
                      Sum Sq   Df  F value    Pr(>F)
x2                 144483642    6  84.7439 < 2.2e-16 ***
x3                2676271691   23 409.4902 < 2.2e-16 ***
p3                   5291997    1  18.6235  1.67e-05 ***
jul                 31860220    1 112.1218 < 2.2e-16 ***
bs(day_nr, df = 5)  53547871    5  37.6889 < 2.2e-16 ***
d3                    303964    1   1.0697    0.3011
Residuals          572008712 2013
```

Modell 3:

```
Anova Table (Type II tests)

Response: butik
                      Sum Sq   Df  F value    Pr(>F)
x2                 7.0803e+09    6 121.3073 < 2.2e-16 ***
x3                 1.0856e+11   23 485.1894 < 2.2e-16 ***
p1                 1.4089e+08    1  14.4831 0.0001457 ***
Mackhelg           4.1937e+09    4 107.7775 < 2.2e-16 ***
jul                1.2131e+09    1 124.7089 < 2.2e-16 ***
bs(day_nr, df = 5) 2.4936e+09    5  51.2674 < 2.2e-16 ***
d3                 8.7692e+07    1   9.0146 0.0027113 **
Residuals          1.9543e+10 2009

---------

Anova Table (Type II tests)

Response: volym
                      Sum Sq   Df  F value    Pr(>F)
x2                 119169433    6  75.6930 < 2.2e-16 ***
x3                2652490308   23 439.5095 < 2.2e-16 ***
p1                   6452607    1  24.5911 7.683e-07 ***
Mackhelg            44135295    4  42.0502 < 2.2e-16 ***
jul                 31343522    1 119.4511 < 2.2e-16 ***
bs(day_nr, df = 5)  54319262    5  41.4025 < 2.2e-16 ***
d3                     12388    1   0.0472     0.828
Residuals          527154139 2009
```

Parameterskattningar för de tre slutliga modellerna

Modell 1:

	butik	volym
(Intercept)	21995.7128	3522.23508
vdagMonday	-4809.7980	-374.17255
vdagSaturday	-7450.1815	-996.65382
vdagSunday	-9581.2382	-957.26824
vdagThursday	-3582.4487	-306.74624
vdagTuesday	-4231.6594	-451.51123
vdagWednesday	-4030.6209	-394.31050
Mack1	8676.3774	1669.63277
Mack2	17086.5490	1013.49428
Mack3	3184.5656	-89.82796
Mack4	2491.8483	204.46279
Mack5	551.3193	-731.24267
Mack6	7077.5589	1393.80599
Mack7	10053.0074	356.93916
Mack8	6159.9975	1681.70441
Mack9	17820.7054	1333.02568
Mack10	7216.5921	241.09580
Mack11	-2749.9201	-264.81367
Mack12	6030.1660	-34.94547
Mack13	2441.1607	1120.10190
Mack14	-812.5430	865.63615
Mack15	22019.0108	4079.40244
Mack16	3850.5929	920.21793
Mack17	-2339.8770	-1478.46769
Mack18	5111.7346	2628.68042
Mack19	12111.8343	900.94956
Mack20	5769.3960	528.80629
Mack21	25479.3689	2108.68962
Mack22	5694.3373	-64.91286
Mack23	3747.8848	216.60882
bs(day_nr, df = 5)1	-2821.9354	421.55822
bs(day_nr, df = 5)2	3984.6246	85.75970
bs(day_nr, df = 5)3	-2210.7626	200.19576
bs(day_nr, df = 5)4	-4200.7801	-685.93363
bs(day_nr, df = 5)5	1956.9470	527.09230
julTRUE	4234.8303	709.58706

Modell 2:

	butik	volym
(Intercept)	36718.5923	8195.84636
vdagMonday	-4747.5582	-354.41523
vdagSaturday	-7481.8016	-1006.69128
vdagSunday	-9617.8482	-968.88966
vdagThursday	-3525.9485	-288.81089
vdagTuesday	-4161.2617	-429.16429
vdagWednesday	-4018.2952	-390.39786
Mack1	8616.0121	1650.47047
Mack2	16982.8022	980.56105
Mack3	2965.2630	-159.44309
Mack4	2347.0052	158.48397
Mack5	586.3058	-720.13661
Mack6	7138.0787	1413.01732
Mack7	10109.7793	374.96074
Mack8	6014.1922	1635.42018
Mack9	17766.2362	1315.73507
Mack10	7181.0549	229.81492
Mack11	-2812.5401	-284.69168
Mack12	6001.3672	-44.08732
Mack13	2397.3518	1106.19527
Mack14	-966.7849	816.67379
Mack15	21969.9281	4063.82169
Mack16	3717.6163	878.00600
Mack17	-2376.0609	-1489.95386
Mack18	5151.0133	2641.14899
Mack19	12141.3013	910.30353
Mack20	5637.1601	486.82952
Mack21	25332.2947	2062.00256
Mack22	5604.7225	-93.36006
Mack23	3652.8578	186.44358
priceL	-1076.6812	-341.78024
bs(day_nr, df = 5)1	-2041.5667	669.27743
bs(day_nr, df = 5)2	3995.3153	89.15336
bs(day_nr, df = 5)3	-2029.5506	257.71947
bs(day_nr, df = 5)4	-3101.8519	-337.09130
bs(day_nr, df = 5)5	2923.6637	833.96558
julTRUE	4042.8474	648.64425

Modell 3:

	butik	volym
(Intercept)	47504.6901	8994.80212
vdagMonday	-4810.0920	-357.42444
vdagSaturday	-5896.8625	-970.04701
vdagSunday	-7813.1033	-1002.84623
vdagThursday	-3528.7005	-286.42687
vdagTuesday	-4165.0039	-429.04521
vdagWednesday	-4048.6682	-399.66028
Mack1	8529.7369	1651.85042
Mack2	19510.2657	1019.21271
Mack3	2770.8867	-151.79555
Mack4	2447.8231	157.61635
Mack5	814.7805	-727.04640
Mack6	7182.3331	1223.41640
Mack7	10345.3580	371.51547
Mack8	5936.2104	1642.62862
Mack9	17865.3330	1312.23221
Mack10	7005.0129	229.78477
Mack11	-2570.2910	-296.32140
Mack12	5872.9057	-43.43267
Mack13	2334.9881	1106.84584
Mack14	-1039.7838	817.47661
Mack15	23623.5045	4228.05390
Mack16	3621.6140	884.71065
Mack17	-2308.7549	-1488.79756
Mack18	5321.5100	2693.78071
Mack19	12072.8009	910.49248
Mack20	5472.5844	489.95394
Mack21	25158.2303	2066.38062
Mack22	5614.2630	-91.90097
Mack23	3793.7486	190.68691
price	-1870.3999	-400.28093
meandiff	-1707.8680	20.29924
Mack2Helg	-16588.3191	-1703.25096
Mack6Helg	-10942.9660	-162.93309
Mack15Helg	567.3338	-1053.67521
Mack18Helg	-256.9350	1138.71244
bs(day_nr, df = 5)1	-1671.4058	730.90976
bs(day_nr, df = 5)2	4124.8680	98.71208
bs(day_nr, df = 5)3	-2210.0301	279.75766
bs(day_nr, df = 5)4	-2270.3280	-258.68000
bs(day_nr, df = 5)5	3423.8487	908.14609
julTRUE	4011.8492	644.85777

Referenslitteratur

1) Methods of Multivariate Analysis (2012), Alvin C. Rencher & William F. Christensen

2) Data Analysis and Graphics Using R – an Example-Based Approach (2010), John Maindonald & W. John Braun

3) Linear Regression Analysis (2003), George A. F. Seber & Alan J. Lee