Modeling cross-selling success in the insurance industry

- Using Generalized Linear Models

Pontus Montgomery

Master's thesis 2013

Abstract

In recent years a shift in sales focus, towards cross-selling, has occurred in the insurance industry. In a saturated market, growth is easier achieved through cross-selling to the customer base, rather than new customer acquisition.

Data containing 50 000 cross-sale attempts, associated customer and company engagement, and the outcome of the cross-sale attempt has been provided by a large Danish insurance company. Four models using different types of Generalized Linear Models (GLM) are examined with the aim to model cross-selling outcome. Each model focus on optimizing a specific sales measurement, such as hit rate, total premium sold and number of products sold. This is achieved by treating the data concerning the customer and their associated products as explanatory variables and the outcome of the cross-sale attempt (e.g. sale/not sale) as response variable in the GLM.

In-sample and out-of-sample analysis shows a strong predictive power in all four models in general and in models using logistic regression in particular.

A real life case study was conducted at a large Danish insurance company where one of the models was tested against a random group, each containing 1300 customers. The model preformed approximately 80 % better in cross-sale hit rate, 90% better in total premium sold and 90% better in total products sold. These results were statistically significant at a 90% level but not at a 95% level.

Tack till

Jag vill tacka min handledare Nader Tajvidi på Institutionen för Matematisk Statistik och min kontaktperson på företaget Fredrik Thuring för all ovärderlig support och kunskap jag fått ta del av.

Förkortningar

LTV	Lifetime Value
GLM	Generalized Linear Model
LR	Logistic Regression
PR	Poisson Regression
NBR	Negative Binomial Regression
ZINBR	Zero-Inflated Negative Binomial Regression
KPI	Key Performance Indicator

Innehållsförteckning

1 Inledning

Försäkringar är inte som andra produkter, om de ens kan definieras som det vi i dagligt tal kallar produkter. En försäkring som säljs inbringar en fast premie varje år, så länge som kunden väljer att ha kvar sitt engagemang i försäkringsbolaget. Kostnaderna för företaget uppstår först när kunden får en skada på sitt försäkrade objekt, eller person. Detta sker vid en obestämd tid och till en obestämd kostnad. För att företaget ska kunna hantera dessa kostnader modelleras kunders och objekts skadefrekvens och kostnader, också känt som riskpremie, genom att använda historiska data. Kunderna och de objekt de försäkrar får sedan en premie baserat på vilken riskkategori de tillhör. Detta är grunderna i prissättning och reservsättning och tillämpas i alla försäkringsbolag i mer eller mindre komplexa modeller (Trowbridge, 1989).

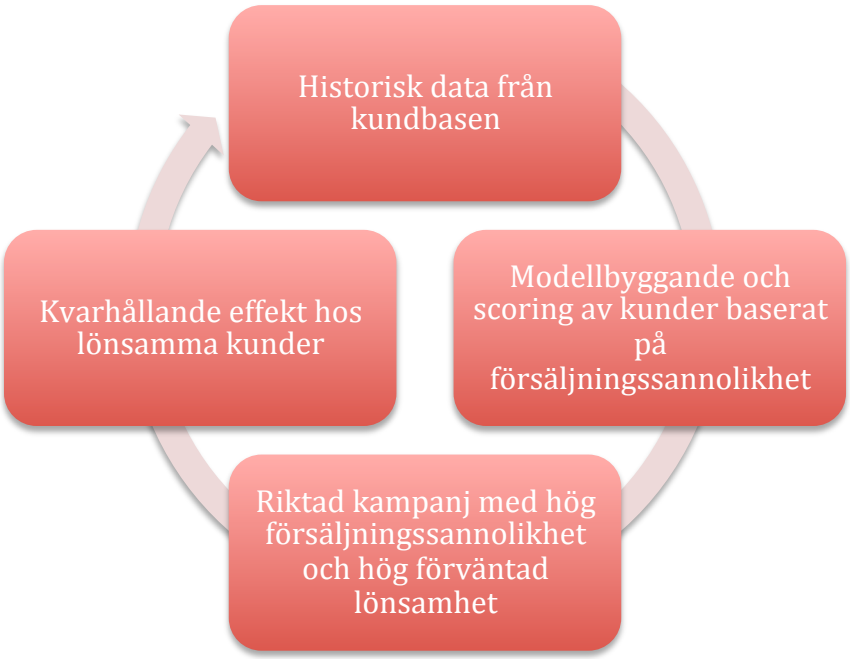
När det kommer till försäljning är modellerna, när det finns sådana, inte alls lika sofistikerade. Detta är förståeligt när de kommer till nyteckning av kunder. Försäkringsbolag kan inte innan de väljer att kontakta en kund veta något dess demografi eller sakinformation om objektet, så som kan göras vid offertgivning. Däremot finns alla dessa uppgifter lagrade i databasen när det kommer till befintliga kunder. Det finns också data lagrat över alla korsförsäljningsförsök och deras utfall. Genom att kombinera dessa data går det att bygga en modell som beräknar sannolikheten för att en kund ska teckna en ny försäkring vid ett korsförsäljningsförsök.

1.1 Bakgrund

Korsförsäljning definieras som att sälja fler produkter eller tjänster till redan existerande kunder. Detta för att stärka relationen mellan förtaget och kunden och på så sätt få kunden att stanna längre. En såld försäkring bringar in en årlig premie så länge kunden väljer att ha kvar sitt engagemang hos företaget och skiljer sig på så sätt från andra branscher. Incitamenten att få kunden att stanna kvar så länge som möjligt är därför större inom försäkringsbranschen, då det höjer kundens ”lifetime value” (LTV)(Akura & Srinivasan, 2005).

Nyteckning av en kund är även förknippat med stora anförskaffningskostnader (Kamakura et al., 2003). Att sälja samma produkt till samma värde med samma risk till en existerande kund är därför mycket mer lönsamt än att göra det till en ny kund.

En annan anledning till att man inom försäkringsbranschen är intresserad av att ha långlivade kunder är att data över deras karakteristik, demografi, beteende, skadehistorik och betalningsförmåga sparas i databaser och används dels för prissättning av andra kunder och dels för att utvärdera kunden i sig. Det går att med hjälp av dessa data rikta korsförsäljningskampanjer mot lönsamma och riskaversa kunder och avhålla sig från att aktivt kontakta olönsamma. Detta skapar alltså en positiv loop. (Kamakura et al., 2003; Larivi`ere & Van den Poel, 2004; Ahn et al., 2011)



1.2 Syfte, upplägg och frågeställning

Denna studie syftar till att föreslå konkreta och lättimplementerade metoder för att optimera korsförsäljning, något som saknas i dagen forskning.

I studien undersöker jag olika korsförsäljningsoptimeringsmodeller för fyra olika försäljningsdefinitioner. Det går sen att välja den modell som bäst passar de försäljningsmål som är aktuella.

Modellerna testas och undersöks för var för sig så att de bäst förklarar det de är optimerade för. Då vissa modeller består av ett flertal delmodeller måste resultaten av dessa viktas samman till ett tal, som rangordnar kunden utefter vad modellerna optimerats för t.ex försäljningssannolikhet, förväntat antal sålda produkter osv. Denna sammanviktning kan göras utefter subjektiva bedömningar kallas därför för poängsättning eller score framöver.

Jag jämför sedan modellerna med varandra, mot ett antal fördefinierade nyckeltal (KPI) som är vanliga att använda vid försäljningsutvärdering inom företaget.

Den modell som anses bäst, utifrån de fördefinierade KPI:erna implementeras sedan, på det företag vars data används i analysen, och testas mot en slumpgrupp.

De frågeställningar jag i mitt arbete framförallt har utgått ifrån är följande:

- Hur optimeras korsförsäljningen i ett försäkringsbolag?

- Vilka statistiska modeller bör användas för att åstadkomma detta?

- Går det att på ett säkerställt sätt mäta modellernas effektivitet?

- Hur implementeras modellerna i en verklig situation?

2 Data – Teori – Modellbyggande

Data till analysen kommer från ett stort danskt försäkringsbolag. Det består av en stor mängd korsförsäljningsförsök gjorda av den utgående telefonförsäljningskanalen i företaget, utförda 2010-2012. Varje rad i datasetet består av ett korsförsäljningsförsök, med tillhörande försäljningsdata, anonymiserade demografiska uppgifter om den kund som kontaktades samt dess engagemang i företaget. Se mer i avsnitt 2.1. Nedan visas summerad information om korsförsäljningsförsöken.

Summary statistics	
Number of rows	≈ 40 000
Number of sales	1 850
Conversion rate	5%
Total Sales	8,2 MDKK
Average premium per sale	4 400 DKK
Average number sold products given sale	1,6
Average premium per attempt	210 DKK

Samtliga modeller använder generaliserade linjära regressionsmodeller som estimeringsmetod. Det innebär att data hanteras som förklarande variabler och responsvariabler. Vilken regressionsmodell som används baseras på vilken som bäst passar till rådande responsvariabeln. Vissa modeller byggs upp av flera delmodeller, som i sin tur enbart använder en responsvariabel, och sen viktas ihop till en scoreingmodell. Detta på grund av att en regression enbart går att utföra med en responsvariabel (Sen & Srivastava, 2011).

2.1 Data

Datat är indelat i förklarande variabler och responsvariabler. I en regression används en eller flera förklarande variabler för att förklara variationen i en responsvariabel. Förklarande variabler är information om kunden och dess

engagemang i företaget. Responsvariabler är försäljningsinformation, d.v.s. om/vad/hur mycket kunden köpte vid försäljningstillfället.

Nedanstående tabell visar ett urval av de viktigaste kolumnerna ur hela datamängden. Datat är fabricerad för att förtydliga i detta exempel. Antalet rader i den riktiga datamängden är 48695.

Id	Age	Etage	Union	Post	House	Home	Car	Boat	Accident	Other
1	56	4A		1040	0	2	0	0	1	0
2	32	3C	KOF	3900	0	1	0	0	0	0
4	21		DSU	3720	1	0	3	0	0	1
5	68	1D		7840	0	0	1	1	1	0

Tot-Prem	Tot-Prod	House-S	Home-S	Car-S	Boat-S	Accident-S	Other-S
4370	2	0	0	0	0	0	0
1685	1	0	0	1	0	0	0
9160	6	0	0	0	0	0	0
6124	3	0	1	0	0	0	2

Tot-Prem-Sale	Tot-Prod-Sale	Sale
0	0	0
2448	1	1
0	0	0
5293	3	1

Information om kund och dess engagemang (förklarande variabler):

Id:	Unikt ID-nummer för varje kontakt.
Age:	Kundens ålder.
Etage:	Vilket våningsplan kunden bor på. Tom ruta innebär hus.
Union:	Vilken fackförening/affinityavtal kunden är ansluten till.
Post:	Postnummer.
House - Other:	Antal policys av respektive försäkring. Även respektive

	policys premie finns med i datat.
Tot-Prem:	Nuvarande total årlig premie.
Tot-Prod	Nuvarande totalt antal försäkringar.

Försäljningsinformation (responsvariabler):

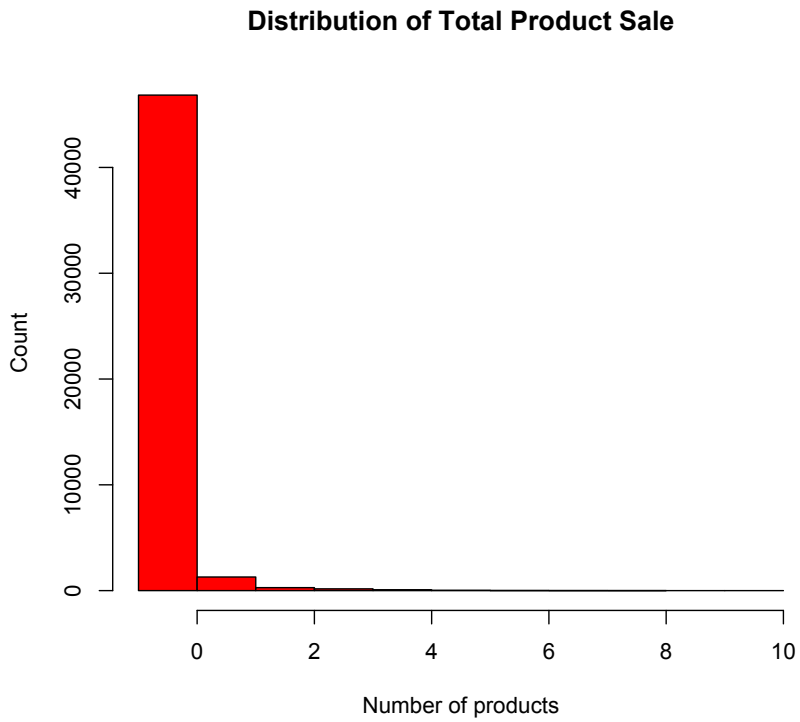
House-S – Other-S:	Antal sålda försäkringar av respektive policy (9 st).
Tot-Prem-Sale:	Totalt årlig premie för sålda försäkringar.
Tot-Prod-Sale:	Totalt antal sålda försäkringar.
Sale:	1 vid försäljning, 0 annars.

Datamängden innehåller även motsvarande säljkolumner för delförsäkringar. Eftersom arbetet i huvudsak är inriktat på korsförsäljning lämnas dessa åt sidan.

2.1.1 Responsvariabler

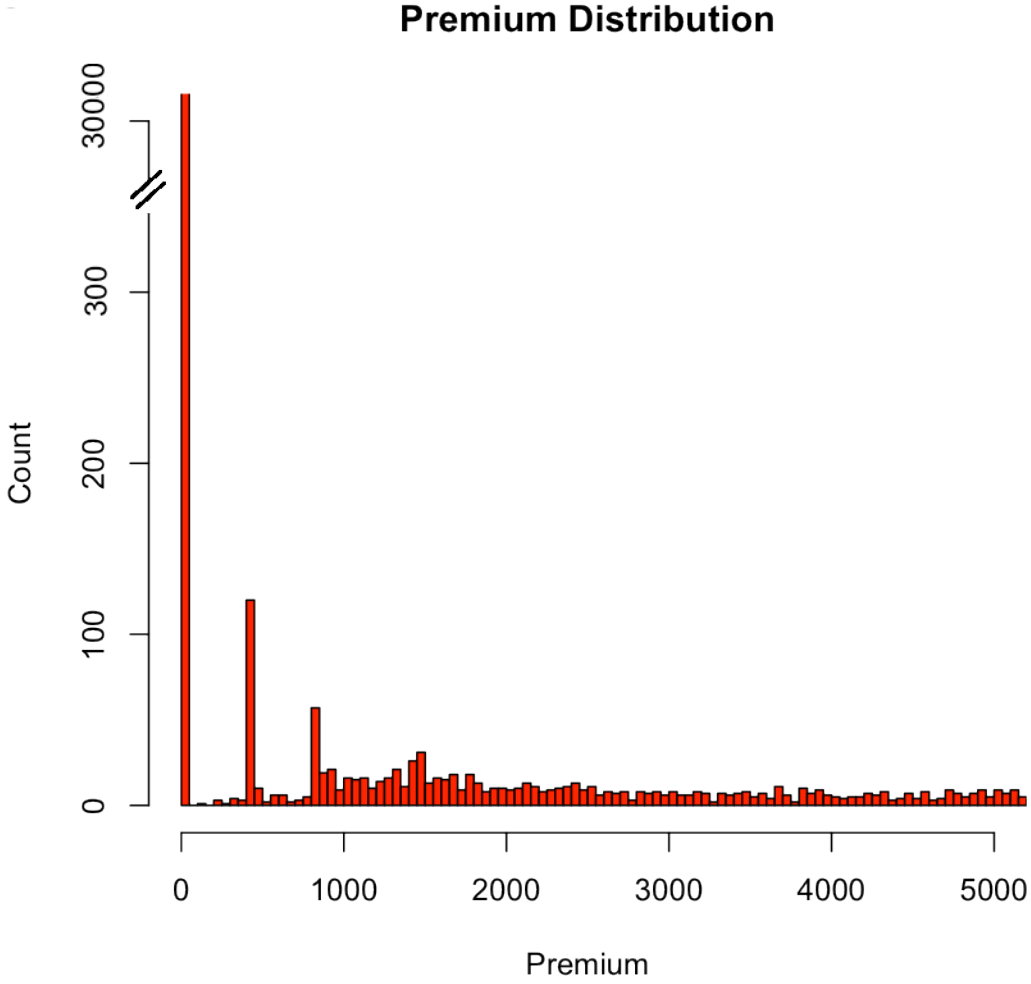
För att välja rätt regressionsmodell till respektive responsvariabel måste en grundlig analys av dess fördelning genomföras. I fallen Sale och House-S – Other-S är den Bernoullifördelad och valet faller naturligt på logistisk regression.

Variabeln Tot-Prod-Sale som anger hur många produkter som såldes vid försäljningstillfället är inte binär då det kan ha sålts mer än en försäkring vid ett försäljningstillfälle. Vid en inspektion av fördelning går det att se en kraftig överrepresentation av nollor.



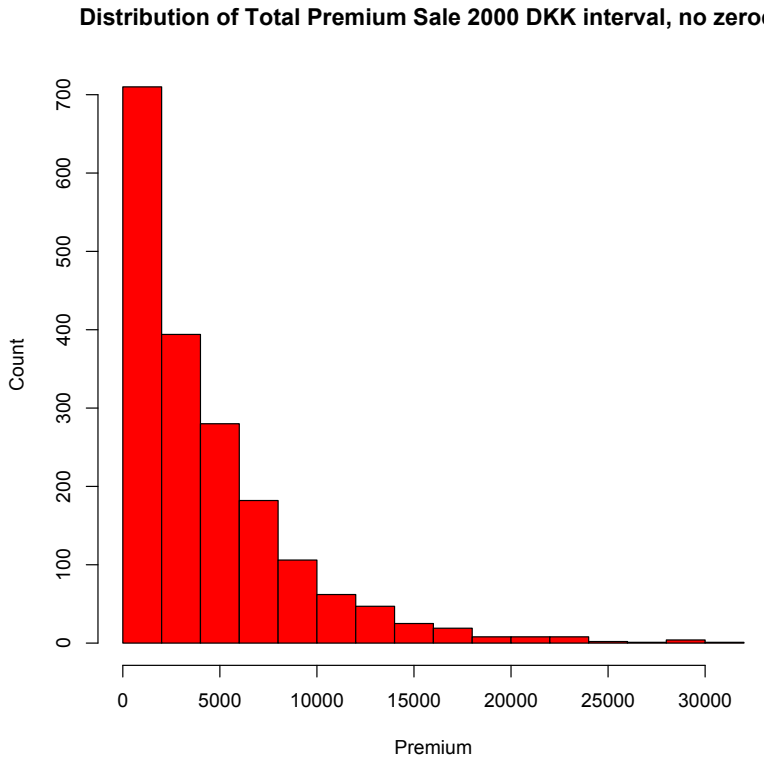
Tas nollorna bort ser fördelning ut som en poissonfördelning. Valet av regressionsmodell bör därför vara inom poissionfamiljen, men där överflödet av nollor hanteras på något sätt.

Variabeln Tot-Prem-Sale har en fördelning som inte liknar någon matematisk beskriven fördelning. Den är inte uppräkningsbar på det sättet som antal sålda försäkringar är, då den börjar på noll och sedan hoppar till ca 500 DKK som nästa värde. Irregulariteten fortsätter sedan över hela intervallet. Därmed finns det inte någon regressionsmodell som passar responsvariabeln.



För att ändå kunna använda Tot-Prem-Sale som responsvariabel i analysen grupperas såld premie på två olika sätt för att kunna användas i två olika typer av regressionsmodeller.

Premien grupperas i intervall om 2000 DKK för att efterlikna en poissonfördelning som i exemplet ovan. Därmed går det nu att använda regressionsmodeller inom poissonfamiljen. Även här måste överflödet av nollor hanteras korrekt. Nedan visas fördelningen då nollorna tagits bort.



Premien delas även in i tre lika stora delar enligt nedan.

TotPremLow, där försäkringstagare som köpt försäkringar för sammanlagt värde av högst 1700 DKK får en etta, övriga en nolla.

TotPremMid, där försäkringstagare som köpt försäkringar för sammanlagt värde av minst 1700 DKK och högst 5000 DKK får en etta, övriga en nolla.

TotPremHigh, där försäkringstagare som köpt försäkringar för sammanlagt värde av minst 5000 DKK får en etta, övriga en nolla.

De tre nya variablerna är nu Bernoullifördelade och det går därmed att använda logistisk regression till var och en av dessa. Se 3.3 hur de tre modellerna viktas ihop till en.

2.1.2 Förklarande variabler

De förklarande variablerna måste analyseras och kategoriseras och framförallt felsökas innan de går att använda i en regression.

Variablerna kan antingen användas som kontinuerliga eller kategoriska i en regression. I de fall en kontinuerlig form användas ska fördelningen helst vara symmetrisk och utan hopp (Sen & Srivastava, 2011). Används kategorisk form bör variablerna grupperas på ett sätt som anses rimligt för att förklara responsvariablens variation (Sen & Srivastava, 2011).

För att finna optimala grupperingar av variablerna använder jag först och främst ett rimlighetsantagande, sedan grafisk analys och slutligen signifikanstest i den enklaste av regressionsmodeller, där Sale används som responsvariabel. Den bästa grupperingen av variablerna har jag sedan använt i alla modeller.

De förklarande variablerna kan även ha samspelseffekt. Ett exempel på detta kan vara att kön och innehav av bilförsäkring i kombination förändrar försäljningssannolikheten utan att de enskilt gör det. Denna egenskap går att utnyttja i regressionsmodellerna genom att kombinationer av variabler hanteras som en variabel och därmed även får ett eget parameterestimat. I de fall där jag använt förklarande variabler i samspel har dessa kombinationer signifikanstestats i respektive modell.

Nedan följer en beskrivning av hur jag hanterat några av de förklarande variablerna. Resonemanget har förts på liknande sätt för andra variabler som ingår i analysen.

Ålder:

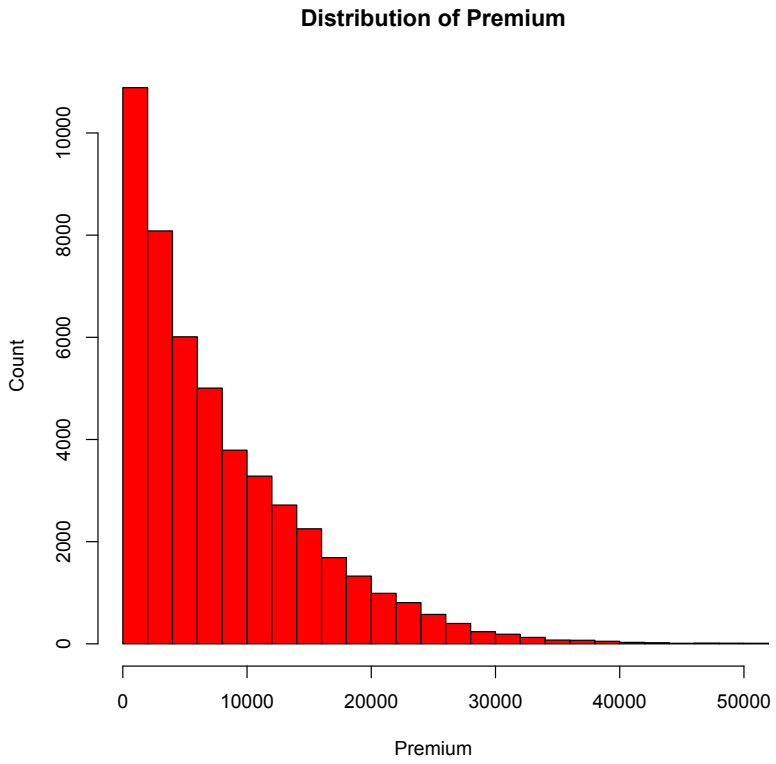
Ålder har en låg förklarandegrad för att estimera försäljningssannolikhet. Indelning i fler olika åldersintervall, såsom 5-års och 10-års intervall, har gett dålig signifikansnivå i regressionerna. Högst signifikansnivå ger uppdelningen i tre grupper + en grupp som saknar ålder. Medianåldern är 55 år.

Total premie

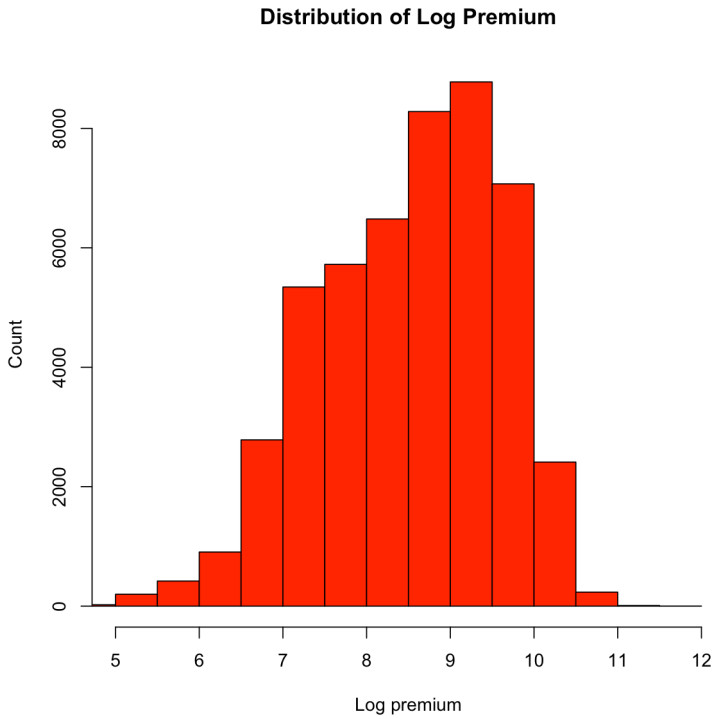
På motsvarande sätt har total premie delats in i ett antal olika percentiler och gett låg signifikansnivå. Bäst har det visat sig att dela in den i två percentiler precis som med ålder.

Denna indelning ger relativt hög signifikansnivå då premie används som ensam förklarande variabel men inte tillsammans med flera då samverkan mellan andra förklarande variabler tar ut dess effekt.

Ett alternativ är låta premien vara en kontinuerlig förklarande variabel, men detta kräver en symmetrisk fördelning. En undersökning av premiens fördelning visar på en kraftig positiv skevhet.

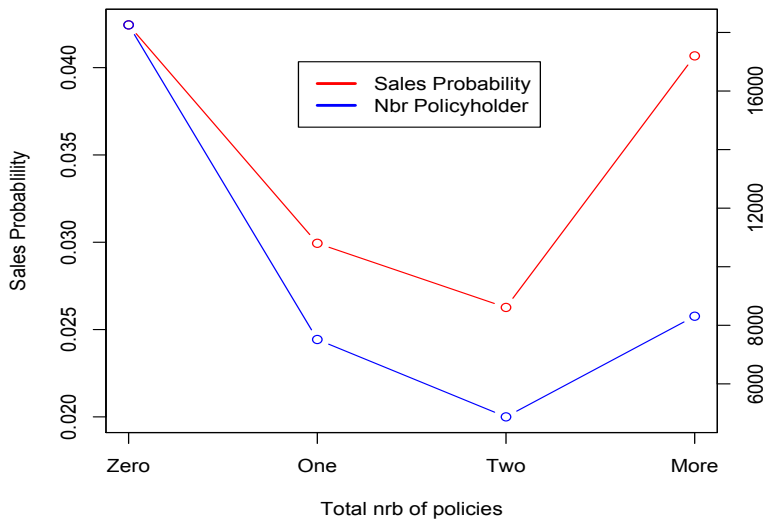


Logaritmen av premien visar däremot på en betydligt mer symmetrisk fördelning och det blir synligt att den är en bättre prediktor än uppdelningen låg-hög premie.



Totalt antal premier

Den grafiska representationen visar tydligt att försäljningssannolikheten är hög vid en försäkring, låg vid två eller tre och ungefär samma som vid en, vid fyra eller fler. Antalet kunder i varje grupp är hög och antyder en hög signifikans.



Den förklarande variabeln ”totalt antal premier” har för att minska antalet faktorer delats in i två grupper. Den första innehåller en + fyra eller fler försäkringar, den andra två eller tre försäkringar. Som den grafiska representationen visar är det en rimlig gruppering då subgrupperna har ungefär samma försäljningssannolikhet. Ur ett rent logiskt perspektiv går det att tänka sig att den som har en försäkring sannolikt aldrig är helförsäkrad oavsett livssituation och därför mer benägen komplettera sin försäkringsportfölj. Den med två eller tre försäkringar har troligtvis en hemförsäkring, en olycksfallsförsäkring och/eller en bilförsäkring. Detta utgör för de flesta en komplett försäkringsportfölj och gör personen mindre benägen att köpa fler försäkringar. Den med fyra eller fler försäkringar har antagligen gått in i en ny fas i livet där hem, bil och kompletteras med barn, sommarhus, båt och/eller en bil till t.ex. och har därför möjlighet att köpa fler försäkringar.

Fackförening

Antalet fackföreningar i datamängden uppgår till ungefär 50 st. Antalet kunder i varje varierar mellan någon enstaka till 5300. De som kunder som inte är anslutna till något fackförbund uppgår till ungefär 13000. Inledningsvis delade jag upp datamängden i faktorerna: icke anslutna, föreningar med 1500 medlemmar eller fler var för sig och övriga fackföreningar. Efter detta utförde jag signifikanstester och de fackföreningar som inte var signifikanta sorterades in under övriga. Eftersom datamängden inte säger något om vilka typer av rabatter/förmåner som ges till de olika fackföreningarna är det svårt att dra en logisk slutsats om varför vissa fackföreningar var signifikanta och andra inte. Kanske är det så att de som var signifikanta erbjuds bättre villkor. Hur som helst blev den bästa uppdelningen enligt följande.

No Union	NÆS	DJØ	Other Union
12966	5259	2441	10006

Postnummer

Uppdelningen av postnummer i faktorer såsom höginkomsttagarområden, låginkomsttagarområden eller storstad, landsbygd eller geografisk uppdelning har gett mycket låg signifikansnivå och används därför inte i analysen.

Antal försäkringar av respektive försäkringsform

Här går det att hitta de förklarande variabler med högst signifikans. De är indelade i faktorerna: noll, en eller flera försäkringar, eller faktorerna noll, en, två eller flera försäkringar beroende på hur de klarat sig i signifikanstesterna.

2.2 Teori

I detta avsnitt beskrivs de regressionsmodeller som används i analysen.

Vanlig linjär regression enligt nedan är användbar när responsvariabeln är normalfördelad.

$$Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + \cdots B_jX_{ji} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma)$$

Responsvariabler i analysen har dock fördelningar som gör att det inte går att använda vanlig linjär regression. GLM är ett ramverk som tillåter att responsvariabeln tillhör exponentialfamiljen. Via en länkfunktion antas responsvariabeln vara linjärt beroende av de förklarande variablerna. Länkfunktionen är i sig beroende av vilken fördelning responsvariabeln har.

$$E(Y) = g^{-1}(BX)$$

Där g^{-1} är en länkfunktion, B är modellens parametrar och X de förklarande variablerna (Madsen & Thyregod, 2011).

2.2.1 Logistisk Regression

Denna regressionsmodell används när responsvariabeln är Bernoullifördelad, alltså endast kan anta värden 1 och 0.

Modellen byggs upp av responsvariabeln Y_i och de förklarande variablerna X_{ji} med tillhörande parametrar B_j. Syftet är att finna sambandet mellan sannolikheten att Y antar värdet 1 och de förklarande variablerna X_i (Madsen & Thyregod, 2011). Sannolikheten att Y antar värdet 1 kan även tolkas som det förväntade värdet av Y, vilket gör att det går att ställa upp modellen med GLM ramverket.

Vänsterledet är ett reellt tal om $Prob(Y = 1)$ ligger mellan 1 och 0.

$$\log\frac{Prob(Y_i=1)}{1-Prob(Y_i=1)} = B_0 + B_1X_{1i} + B_2X_{2i} + \cdots B_jX_{ji} + \varepsilon_i \qquad Y_i \sim Bernuolli(p)$$

Inverteras funktionen fås följande.

$$Prob(Y_i = 1) = \frac{e^{B_0+B_1X_{1i}+B_2X_{2i}+\cdots B_jX_{ji}+\varepsilon_i}}{1 + e^{B_0+B_1X_{1i}+B_2X_{2i}+\cdots B_jX_{ji}+\varepsilon_i}}$$

Länkfunktionen i detta fall är alltså:

$$g^{-1} = \frac{e^{BX}}{1 + e^{BX}}$$

Modellens parametrar estimeras sedan med Maximum Likelihood. Till skillnad från vanlig regression går det inte att finna en sluten lösning utan en iterativ process måste användas, t.ex. Newtons metod (Madsen & Thyregod, 2011).

2.2.2 Poissonregression

Poissonregression används när responsvariabeln är poissonfördelad (Madsen & Thyregod, 2011).

Logaritmen av det förväntade värdet av responsvariabeln Y antas vara linjärt beroende av de förklarande variablerna X enligt följande.

$$\log(\mathrm{E}\{Y_i|X_{ji}\}) = B_0 + B_1X_{1i} + B_2X_{2i} + \cdots B_jX_{ji} \qquad Y_i \sim Pois(u_i)$$

Vilket i inverterad form ger:

$$\mathrm{E}\{Y_i|X_{ji}\} = e^{B_0+B_1X_{1i}+B_2X_{2i}+\cdots B_jX_{ji}}$$

Länkfunktionen är:

$$g^{-1} = e^{BX}$$

Även i detta fall estimeras parametrarna med Maximum Likelihood med numeriska metoder, då en sluten lösning saknas (Madsen & Thyregod, 2011).

2.2.3 Negativ Binomial Regression

Ofta är data inte exakt poissonfördelad utan har en viss överspridning, dvs. variansen är större än medelvärdet. Negativ binomialfördelning är en mer generell variant av poissonfördelning och har en extra parameter κ som kan användas för att justera variansen oberoende av medelvärdet och är därför lämplig att använda då data är överspritt poissonfördelad (Madsen & Thyregod, 2011).

$$\log(\mathrm{E}\{Y_i|X_{ji}\}) = B_0 + B_1X_{1i} + B_2X_{2i} + \cdots B_jX_{ji} + \varepsilon_i \qquad Y_i \sim Negbin(u_i, \kappa)$$

Vilket i inverterad form ger:

$$\mathrm{E}\{Y_i|X_{ji}\} = e^{B_0+B_1X_{1i}+B_2X_{2i}+\cdots B_jX_{ji}+\varepsilon_i}$$

Länkfunktionen är:

$$g^{-1} = e^{BX}$$

Alltså samma som för PR.

2.2.4 Zero Inflated Negativ Binomial Regression

När data innehåller en för stor mängd nollor och dessutom är överspridd är ZINBR en bra regressionsmodell att använda (Green, 1994).

Här modelleras responsvariabeln som två stokastiska processer. En Bernoulliprocess som genererar den stora mängden nollor och en negativ binomialprocess som genererar den negativa binomialdistrubutionen (Green, 1994).

$$P(Y_i = 0) = p$$

$$P(Y_i \sim Negbin(u_i, \kappa)) = 1 - p$$

Där $0 \leq p \leq 1$

2.3 Tester

2.3.1 In-sample tester

Dessa tester sker på samma data som analysen utfördes på och används för att finna en modell som passar tillgänglig data så pass bra som möjligt.

Absoluta tester

Ett absolut test mäter modellens totala kvalitet. Den säger hur mycket bättre en modell passar data kontra en nollmodell, dvs. en modell utan parametrar. Testet är ett mått på om modellen i sig är bättre än ingen modell alls.

Likelihood ratio test

Teststatistiken D definieras som:

$$D = -2\ln\left(\frac{Likelihood\ för\ nollmodellen}{Likelihood\ for\ alternativmodellen}\right)$$

där

$$D \sim X^2(frihetsgrader(alternativmodell) - frihetsgrader(nollmodell)$$

En alternativmodell med fler parametrar än nollmodellen kommer alltid ha minst lika hög likelihood. För att testa om skillnaden är signifikant härleds p-värde av D (Mood & Graybill, 1963).

Relativa tester

Ett relativt test används för att jämföra modeller mellan varandra för att på så sätt utröna vilken modell som är bäst. Den säger inget om modellens faktiska kvalitet. Om alla modeller som jämförs med ett relativt test är dåliga kommer relativa tester inte ge någon varning. Relativa tester är därför bra att använda som en modellurvalsmetod efter att dessa har passerat de absoluta testerna.

Vuongs test

Voungs test är ett likelihood-ratio test som kan användas för att utvärdera modeller som inte är nästlade, dvs. modeller som inte går att återskapa genom att sätta en parameter till noll i en mer komplex modell (Vuong, 1989).

I de fall modeller med olika regressionsmetoder ska jämföras är Vuongs test ett bra val.

Teststatistiken Z definieras enligt:

$$Z = \frac{L_1 - L_2 - \frac{K_1 - K_2}{2} logN}{\sqrt{N}\omega_N}$$

Där L_1och L_2är respektive modells maximala likelihood, K_1och K_2 är antal parametrar i respektive modell och ω_N är variansen av den individuella log-likelihood ration mellan de två modellerna (Vuong, 1989).

Akaike Information Criterion

AIC är ett test som avväger mellan modellens goodness-of-fit och dess komplexitet. Om en variabel läggs till i modellen måste dess förbättring av modellens Likelihood vara större än förlusten av en frihetsgrad för att få ett lägre AIC (Akaike, 1974).

$$AIC = 2 * antal\ parametrar - 2\ln(Modellens\ likelihood)$$

Goodness of fit

Hosmer-Lemeshow test

De predikterade värdena sorteras från minsta till högsta och grupperas sedan i t.ex. 10 st. lika stora grupper och dess värden summeras, så även de tillhörande observerade värdena. Sedan plottas de observerade grupperna mot de predikterade och jämförs med en y=x linje, som anger en fullständig fit. (Hosmer & Lemeshow, 2000).

Detta test går att använda för att jämföra modeller av olika slag. Eftersom responsvariablerna modelleras med både logistisk regression och modeller inom poissonfamiljen är detta test ett bra val.

2.3.1 Out-of-sample tester

Dessa tester sker på data som inte varit med i analysen och visar hur pass bra modellerna predikterar observerade värden.

Quintile plot

De predikterade värdena grupperas i 5 lika stora grupper, från lägst till högst. I grupp 5 finns alltså de 20 % högst scorade kunderna. Sedan plottas de summerade predikterade värden i varje grupp och jämförs med de summerade observerade värden för samma grupp.

En god modell bör ha så stor skillnad som möjligt mellan grupperna, då detta innebär att modellen förklarar mycket av variationen i responsvariabeln. Den bör även ha så liten skillnad som möjligt mellan predikterade och observerade värden. En god modell bör även ha en liknande quintile plot in-sample som out-of-sample.

2.4 Modellbyggande

Datasetet delas slumpmässigt in i en analysdel och en valideringsdel. Uppdelning görs så att analysdelen innehåller 80 % av raderna och valideringsdelen 20 %.

En nyckel sätt för att kunna återskapa samma slumpserie vid andra tillfällen. Slumpvektorn genereras genom dragning utan återläggning och de två delarna skapas med hjälp av slumpvektorn.

Vid urvalet av vilka faktorer som de förklarande variablerna ska delas in i har logistisk regression och tester används på dem separat. I de kompletta modellerna används flera variabler och kombinationsvariabler. Detta kan innebära att en variabel som var signifikant ensamstående inte är det tillsammans med andra.

T.ex. kanske ålder visar sig vara signifikant för sig själv. Används även antal produkter i analysen är inte ålder signifikant längre. Detta beror i så fall på att hög ålder indikerar att kunden har många produkter, som i sin tur ökar försäljningssannolikheten, inte att ålder i sig ökar försäljningssannolikheten.

Nedan visas byggandet och testandet av en av de fyra modellerna, där programspråket R används. Tillvägagångssättet är liknande för alla modeller men testerna kan skilja sig åt. Observera att alla tester nedan är in-sample.

I kapitel 3 visas hur modellerna sätts samman till en försäljningsscore och KPI:er beräknas.

2.4.1 Regression

Här analyseras modellen med responsvariabeln Tot-Prod-Sale, se 2.1 för mer information.

Då responsvariabeln antas vara ungefärligt poissonfördelad testas följande tre regressionsmodeller som är anpassade för poissonfördelade variabler.

1. Vanlig poissonregression (PR), bra om responsvariabeln har korrekt poissonfördelning.
2. Negativ binomial regression (NBR), som kompenserar för överspridd poissonfördelning.
3. Zero Inflated Negative Binomial Regression (ZINBR), som kompenserar för både överspridning och en excess av nollor.

Först används NBR för att välja ut vilka variabler som är signifikanta och därför ska ingå i modellerna. Detta motiveras med att vanlig PR har missvisande signifikanstester för parameterestimaten om responsvariabeln inte har en korrekt poissonfördelning och därför är olämplig att använda för att välja ut variabler (Ismail & Jemain, 2007). Jag anser att inte heller ZINBR är en lämplig modell för att välja ut signifikanta variabler då den som tidigare nämnt består av två delar.

Nedan följer en utskrift från R efter ett förenklat exempel av en NBR där variablerna Ålder, Fackförening, Bilförsäkring, Bostadsform, Premie, Hemförsäkring, Husförsäkring, Husförsäkringspremie och Premie*Husförsäkring ingår.

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.58527    0.42133  -8.509  < 2e-16 ***
AgeYoungYoung  -0.19364    0.08824  -2.195  0.02819 *
AgeOldOld       0.17907    0.08587   2.085  0.03703 *
UnionNÆS        0.48616    0.10671   4.556 5.22e-06 ***
UnionDJØ        0.36331    0.12504   2.906  0.00367 **
UnionUnion      0.52597    0.07316   7.189 6.53e-13 ***
CarOne          0.23460    0.10031   2.339  0.01935 *
CarMore         0.32765    0.16958   1.932  0.05335 .
EtageApartment  0.02140    0.07419   0.289  0.77295
Prem            0.09879    0.05362   1.843  0.06540 .
HomeMore       -0.23501    0.85631  -0.274  0.78374
HouseMore       0.27131    1.27279   0.213  0.83120
HousePrem      -0.03358    0.15153  -0.222  0.82461
Prem:HomeMore  -0.05749    0.09650  -0.596  0.55139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.0697) family
taken to be 1)

    Null deviance: 5343.4  on 30671  degrees of freedom
Residual deviance: 5175.8  on 30658  degrees of freedom
AIC: 14752
```

Först visas parameterestimaten och signifikansnivåerna för de olika variablerna. Väljs en 90% signifikansnivå så är variablerna Ålder, Fackförening, Bilförsäkring och Premie signifikanta i exemplet ovan.

Dispersion parameter anger den estimerade överspridningsparametern. Är detta värde noll så är responsvariabeln poissonfördelade utan överspridning. Dock kan den fortfarande innehålla för många nollor för att vara poissonfördelad.

Under visas ett mått på hur pass bra modellen presterar i sin helhet. Null deviance visar nollmodells (enbart intercept) ”deviance” och antalet frihetsgraden den har. Residual deviance visar alternativmodellens (modellen med alla ovanstående parametrar) ”deviance” och antalet frihetsgrader. Ett Log Likelihood Test visar om modellen som helhet är bättre än ingen modell alls.

> pchisq(null.deviance-deviance,df.null-df.residual)

P-value = 5.679818e-29

P-värde < 0.01 visar att modellen är signifikant bättre än en tom modell.

Slutligen visas även AIC-värdet som mäter modellens relativa prestation. Det säger alltså huruvida en modell med ett visst antal variabler är bättre än en med ett annat antal variabler. Val av variabler i modellerna optimeras därför med hjälp av AIC-värdet, se 2.3.

Tas kontinuerligt den minst signifikanta variabeln bort får jag till slut en modell med enbart signifikanta variabler. Nedan visas den kompletta modellen där alla variabler är signifikanta.

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.43034    0.44783  -9.893  < 2e-16 ***
AgeYoungYoung    0.60942    0.16799   3.628 0.000286 ***
AgeMidMid        0.79987    0.15428   5.185 2.16e-07 ***
```

```
AgeOldOld          0.98904    0.16701   5.922 3.18e-09 ***
CarOneOne          0.15266    0.08582   1.779 0.075257 .
UnionNÆS           0.69600    0.11346   6.134 8.55e-10 ***
UnionDJØ           0.46708    0.12868   3.630 0.000284 ***
UnionUnion         0.55378    0.08027   6.899 5.23e-12 ***
Prem               0.12175    0.05373   2.266 0.023460 *
HomeMore          -2.78605    1.17362  -2.374 0.017601 *
TotP1More         -0.94993    0.31044  -3.060 0.002214 **
TotP23Two-Three    1.07732    0.32050   3.361 0.000776 ***
SummerHouseMore    2.42924    0.85062   2.856 0.004292 **
AccidentOneOne    -0.41960    0.08915  -4.706 2.52e-06 ***
AccidentMoreMore  -0.29623    0.13872  -2.135 0.032725 *
HomeOtherMore     -1.06289    0.26732  -3.976 7.01e-05 ***
HouseAccidentMore -0.46774    0.14749  -3.171 0.001517 **
CarAccidentMore   -0.52652    0.14975  -3.516 0.000438 ***
CarAgeMore        -0.39209    0.15238  -2.573 0.010081 *
HomePrem          -0.30516    0.14299  -2.134 0.032831 *
SummerHousePrem   -0.31686    0.11287  -2.807 0.004996 **
Prem:HomeMore      0.63084    0.13623   4.631 3.64e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.0748) family
taken to be 1)

    Null deviance: 5523.4  on 30671  degrees of freedom
Residual deviance: 5215.3  on 30650  degrees of freedom
AIC: 14636
```

Samma variabler använder jag sedan i de två andra regressionsmodellerna, PR och ZINBR. Vuongs test används för att testa modellerna mot varandra, se 2.3.

```
> vuong(poisson,nb)
model2 > model1, with p-value 6.282678e-50
```

Först testas PR mot NBR där NBR tydligt är signifikant bättre än PR eftersom p-värdet är <0.001.

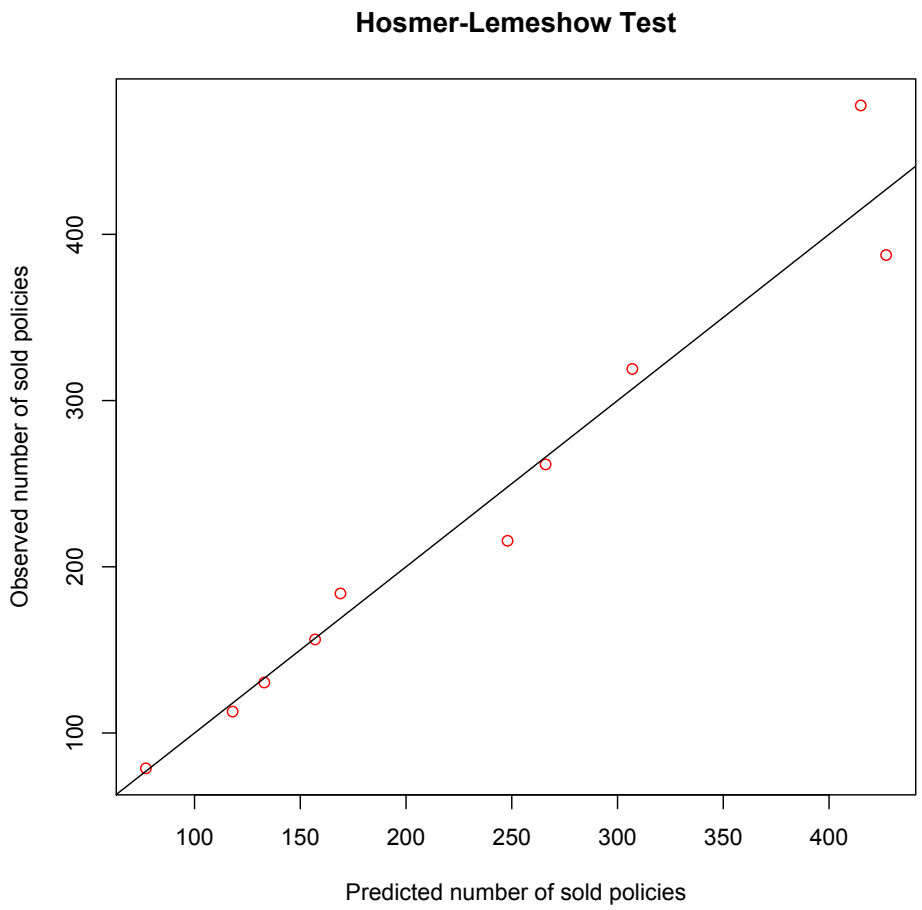
```
> vuong(poisson,zinb)
model2 > model1, with p-value 8.203201e-59
```

PR testas sen mot ZINBR som är än mer signifikant bättre än PR.

```
> vuong(nb,zinb)
model2 > model1, with p-value 1.56686e-21
```

Slutligen testas NBR mot ZINBR, där ZINBR är tydligt bättre än NBR, vilket innebär att ZINBR är den bästa regressionsmodellen, av de tre testade modellerna, till responsvariabeln Tot-Prod-Sale och väljs ut som slutgiltig modell.

ZINBR-modellens goodness-of-fit demonstreras med ett Hosmer-Lemeshow test. Modellen visar på mycket god passning, dock med viss deviation i de högst rankade grupperna.



2.4.2 Övriga modeller

Övriga modeller (ingående variabler och parameterestimat med signifikanstester) till responsvariablerna, Sale, Tot-Prem-Sale och Individual-Insurance-Sale redovisas i Bilagor för respektive modell. Nedan visas dock vilka regressionsmetoder som använts till respektive responsvariabel.

Till responsvariabel Sale används LR som regressionsmetod.

Till responsvariabel Tot-Prem-Sale används LR som regressionsmetod när responsvariabeln är grupperad i tre grupper. När responsvariabeln är grupperad i 2000 DKK intervall används ZINBR.

Till responsvariabel Individual-Insurance-Sale används LR som regressionsmetod.

3 Resultat

Modellerna testas mot varandra genom att först använda parameterestimaten från tidigare respektive regressionsanalys till att scora valideringssetet. Sedan vägs delmodellerna samman till en totalscore enligt de metoder som visas nedan för respektive modell. För varje modell visas sedan Hosmer-Lemeshow test och två quantileplottar, dels in-sample och dels out-of-sample.

Slutligen väljs de 20 % högst scorade kunderna ut från valideringssetet och de fördefinierade KPI:erna beräknas. Dessa KPI:er jämförs sedan med en slumpgrupps KPI:er för att mäta förbättringsgraden.

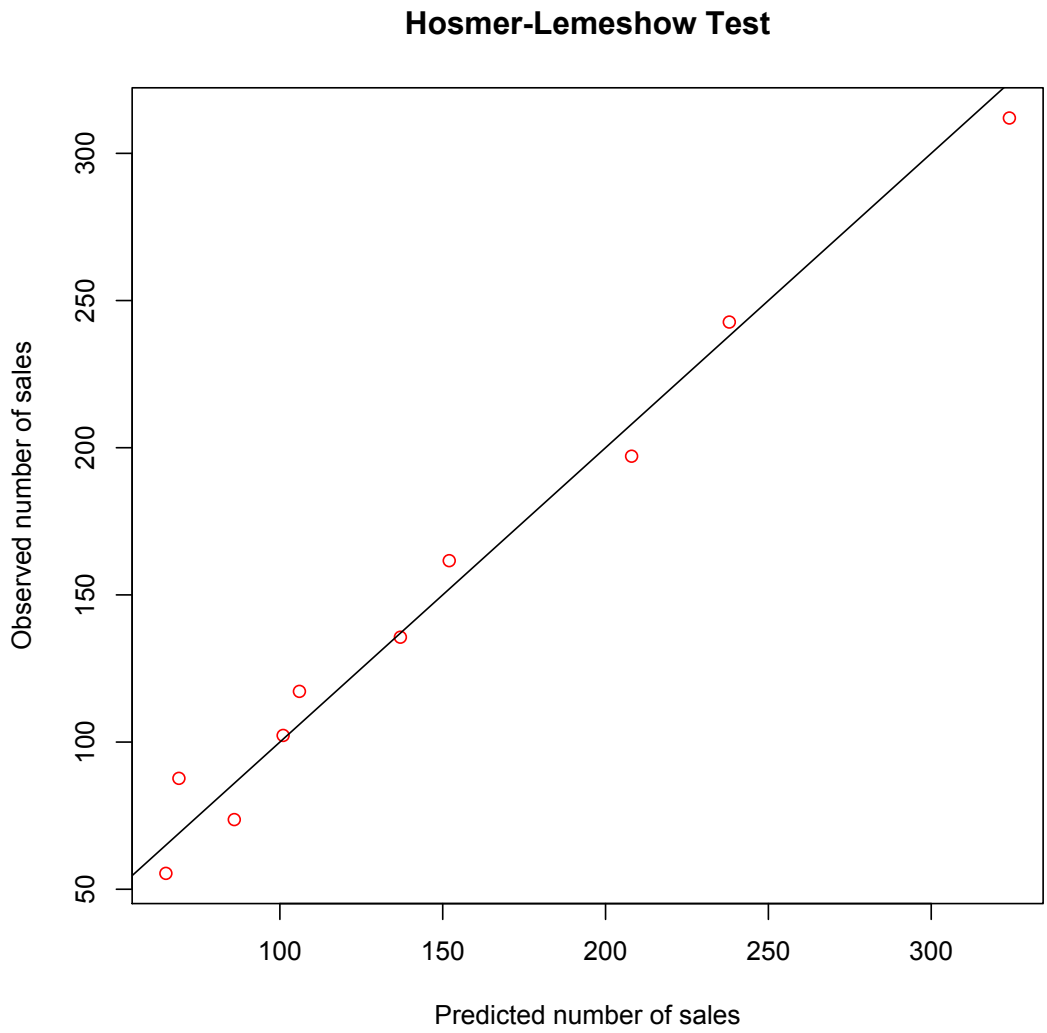
De 20 % högst scorade kunderna väljs ut för att det ansågs vara en rimlig andel av kundbasen att kontakta under ett år. KPI jämförelserna görs för att kunna jämföra modellerna mot varandra, då de inte är konstruerade för att optimera samma sak, för att sedan kunna välja en modell som ska användas i fallstudien.

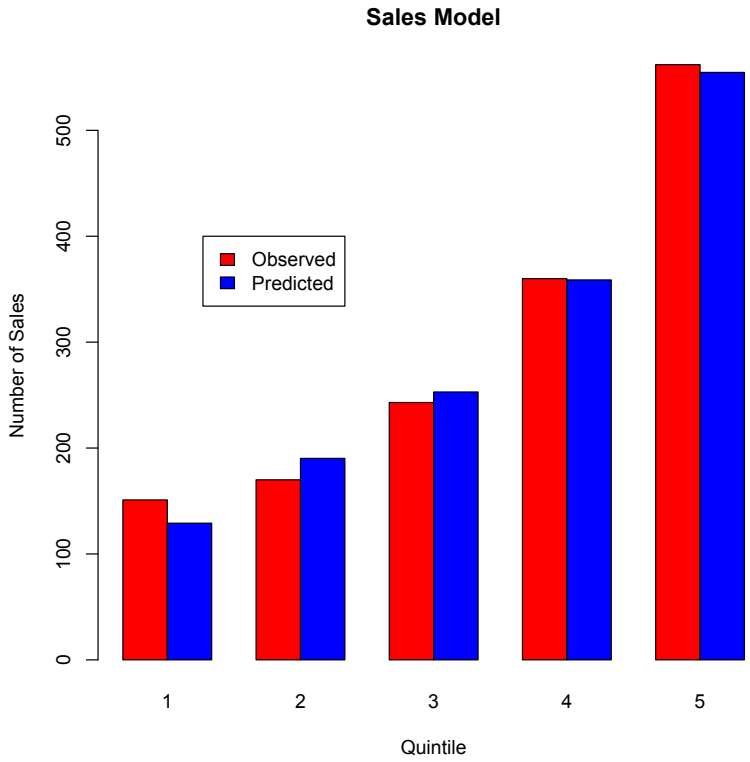
Valideringssetet innehåller 7667 kunder och de 20 % högst scorade av dessa utgör 1533 kunder i alla KPI beräkningar.

3.1 Modell 1 – Responsvariabel Total Sales

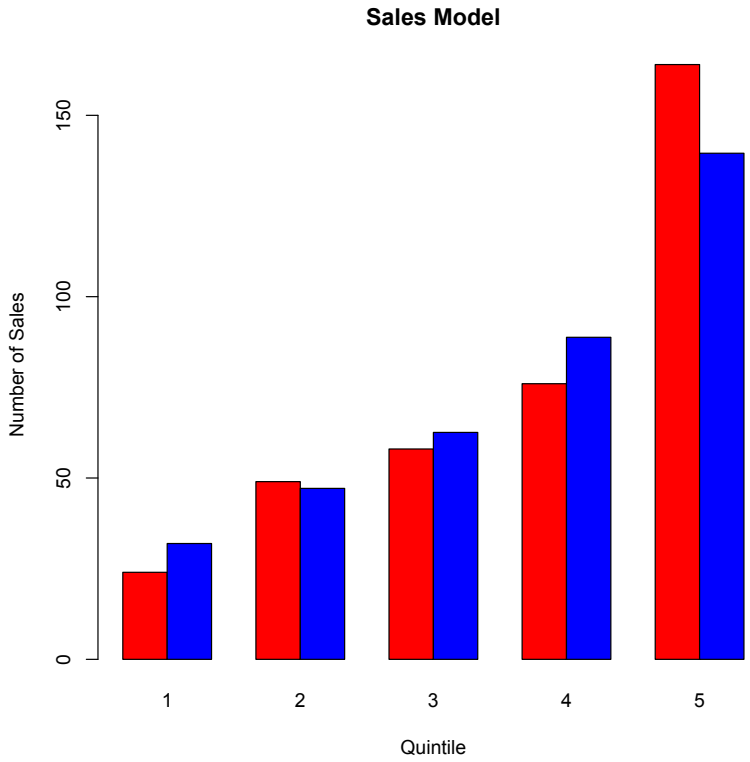
Denna modell har enbart en responsvariabel och det behövs alltså inte viktas ihop flera scorer till en. Modellen optimerar försäljningssannolikheten.

In-Sample


Hosmer-Lemeshow Test
Observed number of sales
Predicted number of sales
50
100
150
200
250
300
100
150
200
250
300


Sales Model
Number of Sales
0
100
200
300
400
500
Observed
Predicted
1
2
3
4
5
Quintile

Out-of-sample



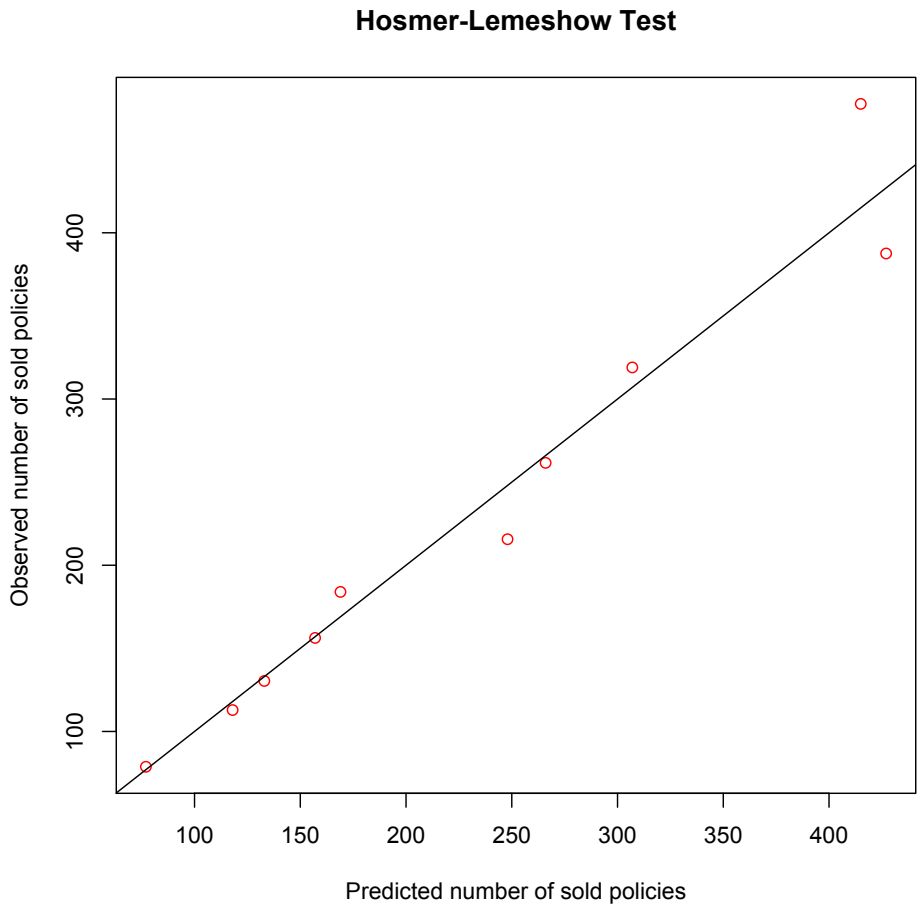
KPI:er – Out-of-sample

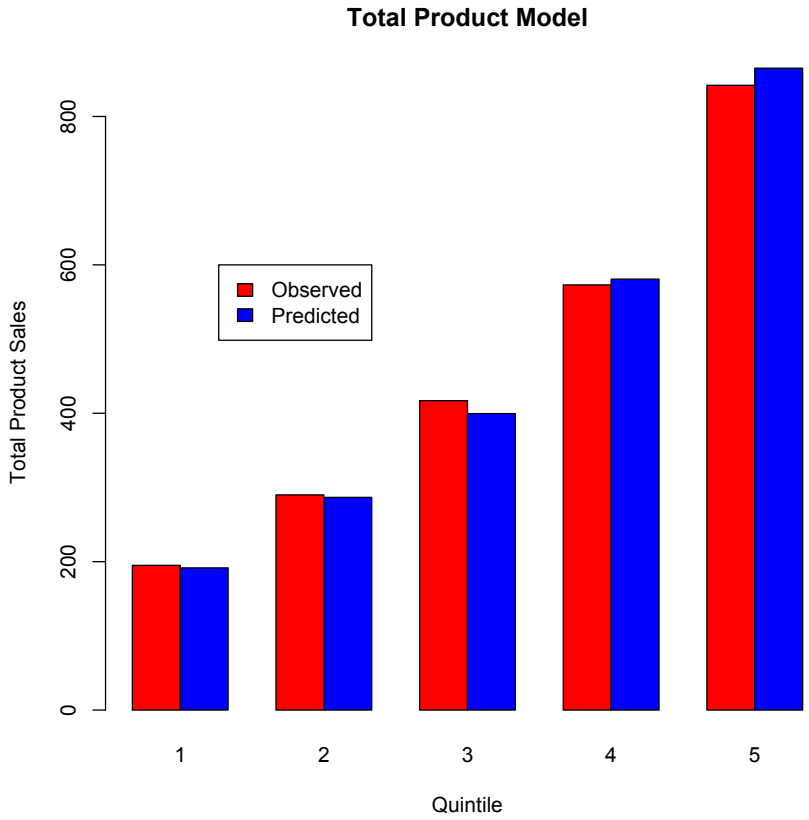
	Random group	Model group	X-times Improvment
Customer	1533	1533	-
Sales	74	164	222%
Products Sold	117	240	205%
Premium	322 600 DKK	583 000 DKK	181%
Sale/Customer	4,8%	10,7%	222%
Product Sold/Sale	1,58	1,46	93%
Premium/Customer	210 DKK	380 DKK	181%
Premium/Sale	4 359 DKK	3 555 DKK	82%

3.2 Modell 2 – Responsvariabel Product Sales

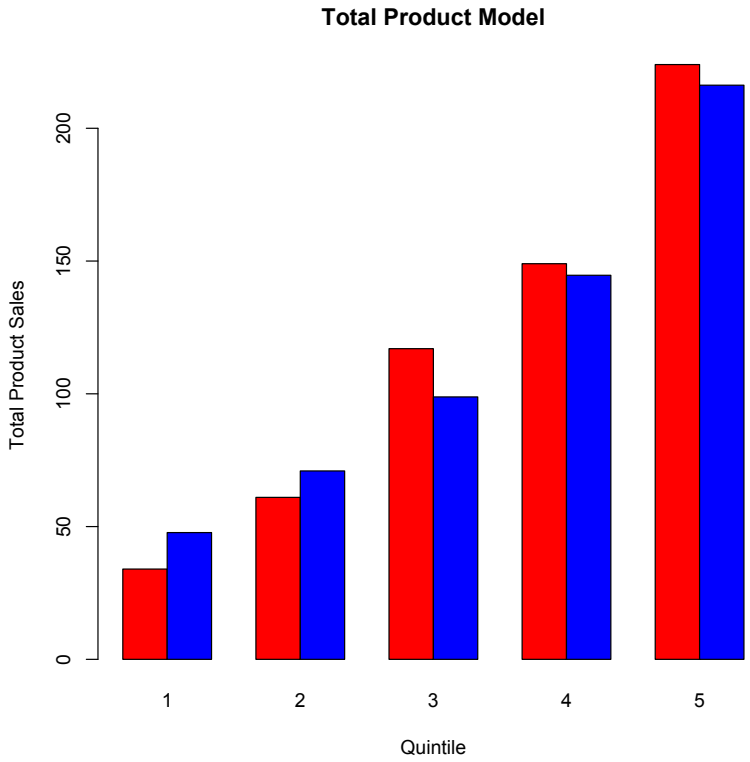
Motivet bakom att använda Total Product Sale som responsvariabel är att skapa en modell som är optimerad för att sälja så många försäkringar som möjligt. Även denna modell har enbart en responsvariabel och behöver inte viktas ihop till en score. Modellens predikterar förväntat antal sålda försäkringar för varje kund.

In-sample




Total Product Model
Total Product Sales
0
200
400
600
800
Observed
Predicted
1
2
3
4
5
Quintile

Out-of-sample



KPI:er – Out-of-sample

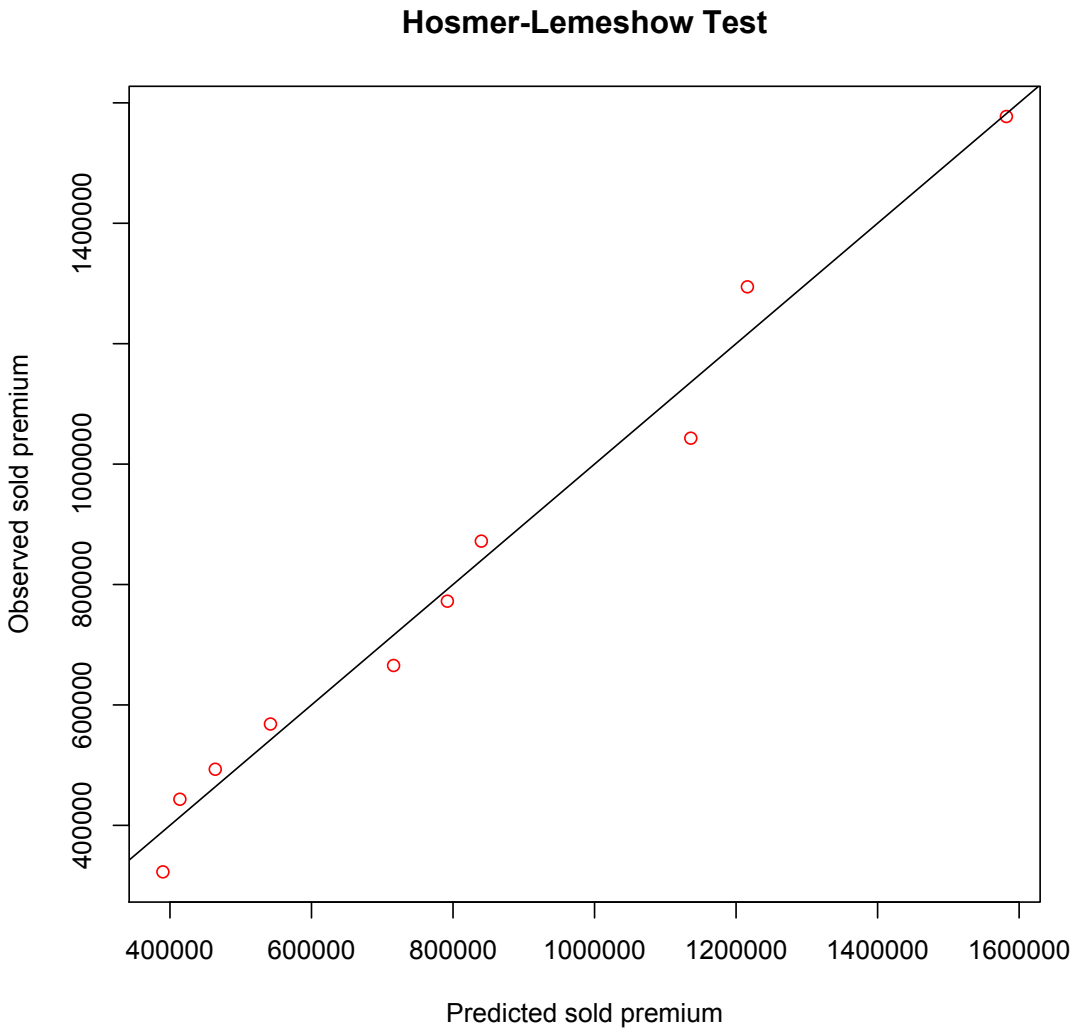
	Mean group	Model group	X-times Improvment
Customer	1533	1533	-
Sales	74	129	174%
Products Sold	117	224	191%
Premium	322 600 DKK	577 000 DKK	179%
Sale/Customer	4,8%	8,4%	174%
Product Sold/Sale	1,58	1,74	110%
Premium/Customer	210 DKK	376 DKK	179%
Premium/Sale	4 359 DKK	4 473 DKK	103%

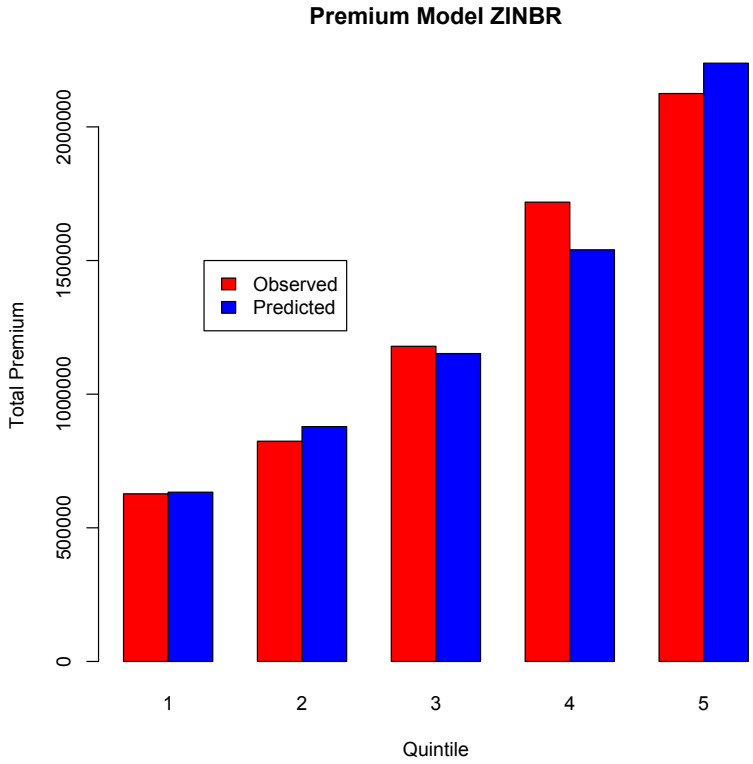
3.3 Modell 3 – Responsvariabel Total Premium Sales

Responsvariablen Total Premium Sales modelleras med två olika typer av regressionsmodeller. En med ZINBR och en med tre logistiska regressioner som viktas samman till en score.

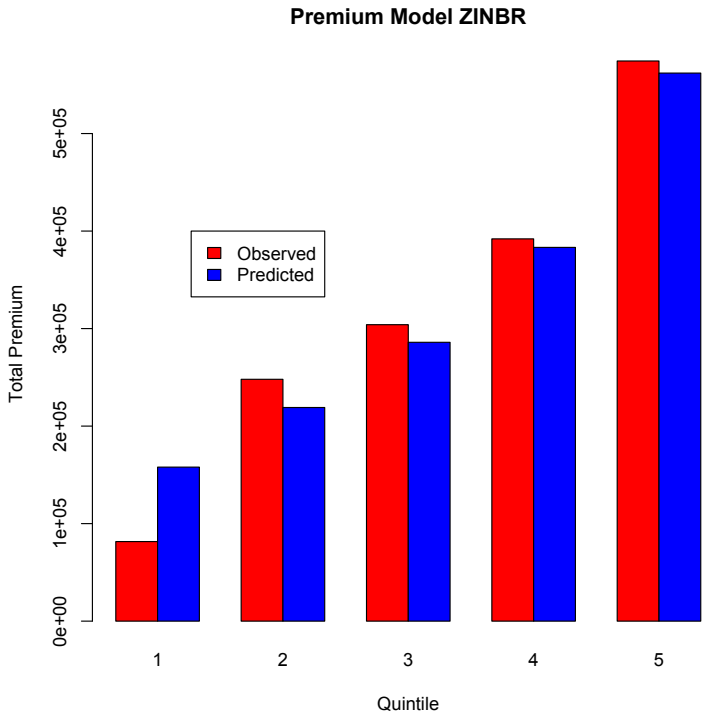
3.3.1 ZINBR Modellen

In-sample




Premium Model ZINBR
Total Premium
0
500000
1000000
1500000
2000000
Observed
Predicted
1
2
3
4
5
Quintile

Out-of-sample



KPI:er - Out-of-sample

	Mean group	Model group	X-times Improvment
Customer	1533	1533	-
Sales	74	125	169%
Products Sold	117	213	182%
Premium	322 600 DKK	584 000 DKK	181%
Sale/Customer	4,8%	8,2%	169%
Product Sold/Sale	1,58	1,70	108%
Premium/Customer	210 DKK	381 DKK	181%
Premium/Sale	4 359 DKK	4 672 DKK	107%

3.3.2 Logistisk Regressions Modellen

Resultatet av regressionen är tre estimerade sannolikheter för varje kund. En, kallad P(A), som anger sannolikheten att de köper en försäkring för mindre än 1700 DKK, en kallad P(B), som anger sannolikheten att de köper en försäkring för 1700-5000 DKK och en P(C) som anger sannolikheten att de köper en försäkring för mer än 5000 DKK.

Eftersom de tre händelserna är ömsesidigt uteslutande kombineras de enkelt genom:

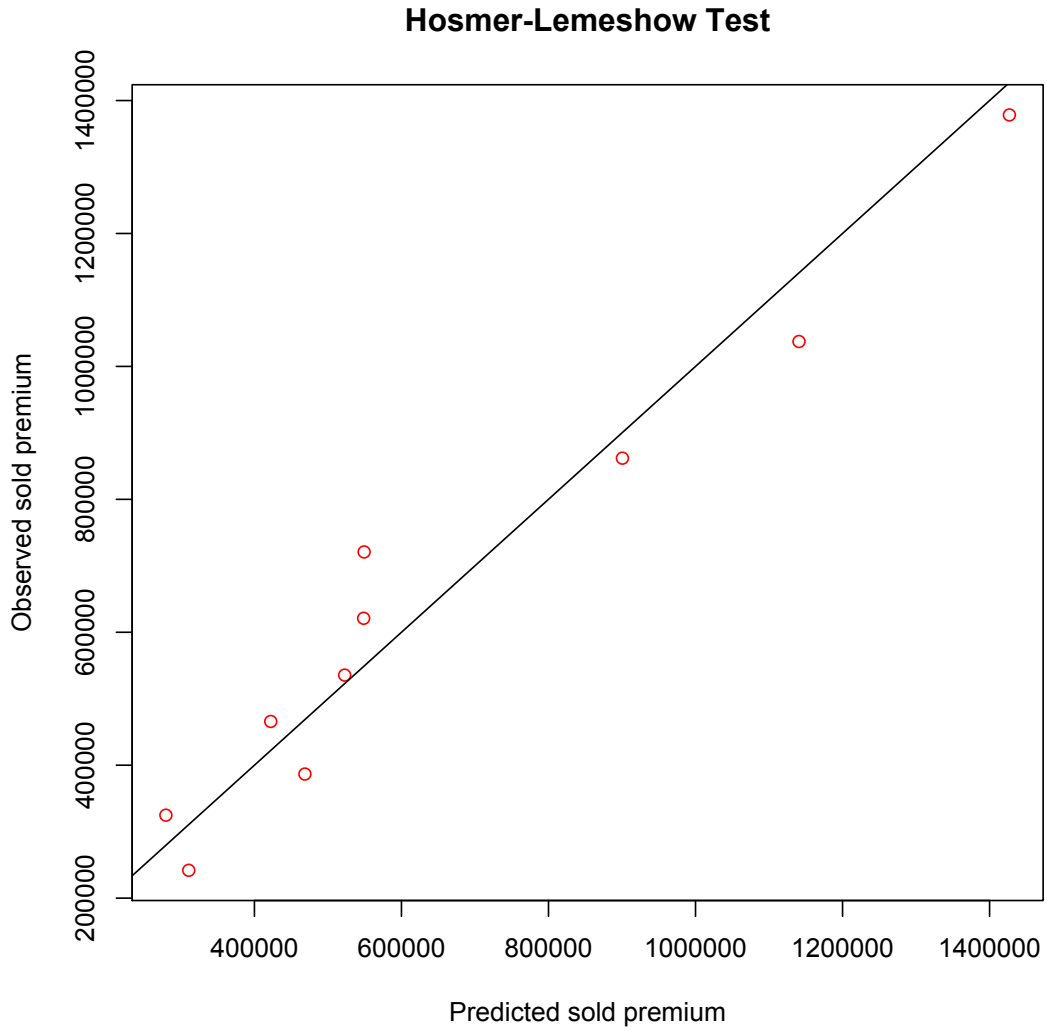
P(A eller B eller C) = $P(A \cup B \cup C) = P(A) + P(B) + P(C)$

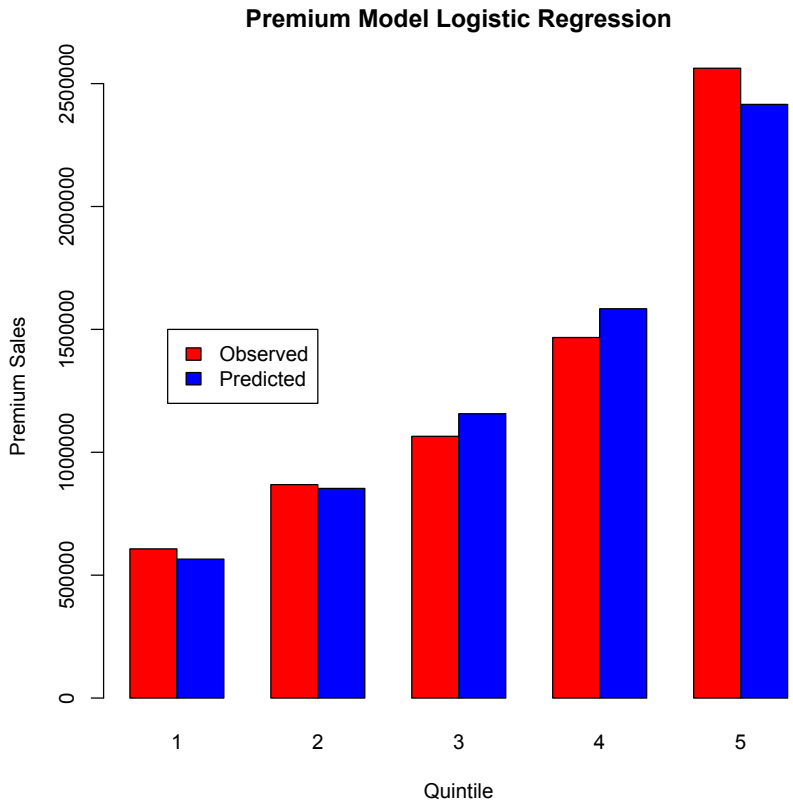
Eftersom P(C) anger sannolikheten att en kund köper en dyr försäkring bör denna sannolikhet viktas högre än de andra då denna modell ska maximera premieintäkten. Modellen byggs därför upp av sannolikheterna multiplicerat med snittpremien i motsvarande intervall.

Rank = P(A)*965+P(B)*3138+P(C)*9353.

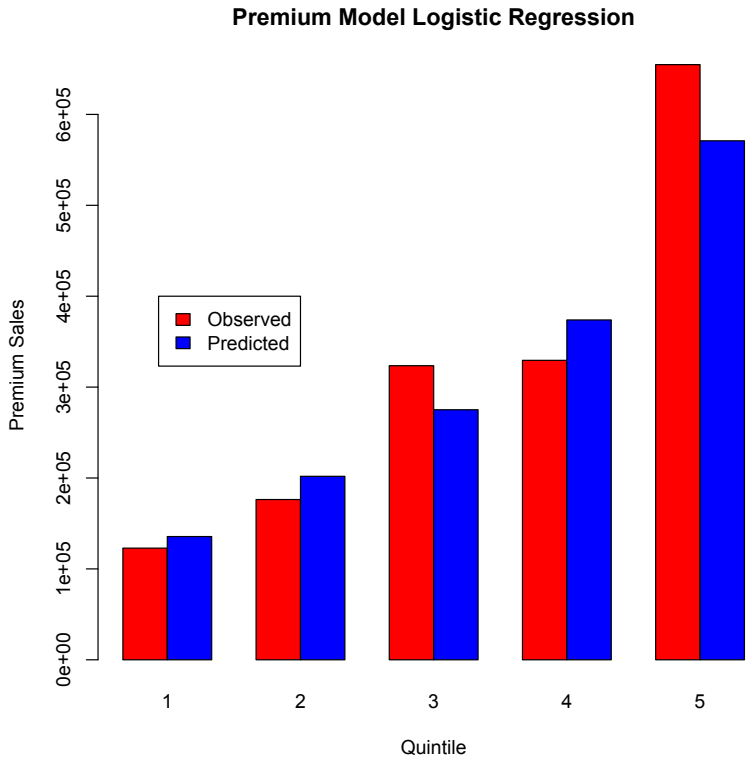
Denna modell beräknar alltså förväntad såld premie för respektive kund.

In-sample


Hosmer-Lemeshow Test
Observed sold premium
Predicted sold premium
200000
400000
600000
800000
1000000
1200000
1400000


Premium Model Logistic Regression
Premium Sales
0
500000
1000000
1500000
2000000
2500000
Observed
Predicted
1
2
3
4
5
Quintile

Out-of-sample



KPI:er - Out-of-sample

	Mean group	Model group	X-times Improvment
Customer	1533	1533	-
Sales	74	135	182%
Products Sold	117	228	195%
Premium	322 600 DKK	655 000 DKK	203%
Sale/Customer	4,8%	8,8%	182%
Product Sold/Sale	1,58	1,69	107%
Premium/Customer	210 DKK	427 DKK	203%
Premium/Sale	4 359 DKK	4 852 DKK	111%

3.4 Modell 4 – Responsvariabel Individual Insurance Sales

Totalt 9 modeller som har en specifik försäkringsförsäljning som responsvariabel.

Eftersom de predikterade sannolikheterna inte är ömsesidigt uteslutande (en bil och en hemförsäkring kan säljas vid samma tillfälle) eller oberoende så går det inte att kombinera ihop dem som tidigare. Istället ges den sammansatta sannolikheten av principen om inklusion/exklusion (Böiers, 2003).

I ett fall med tre mängder ges den av:

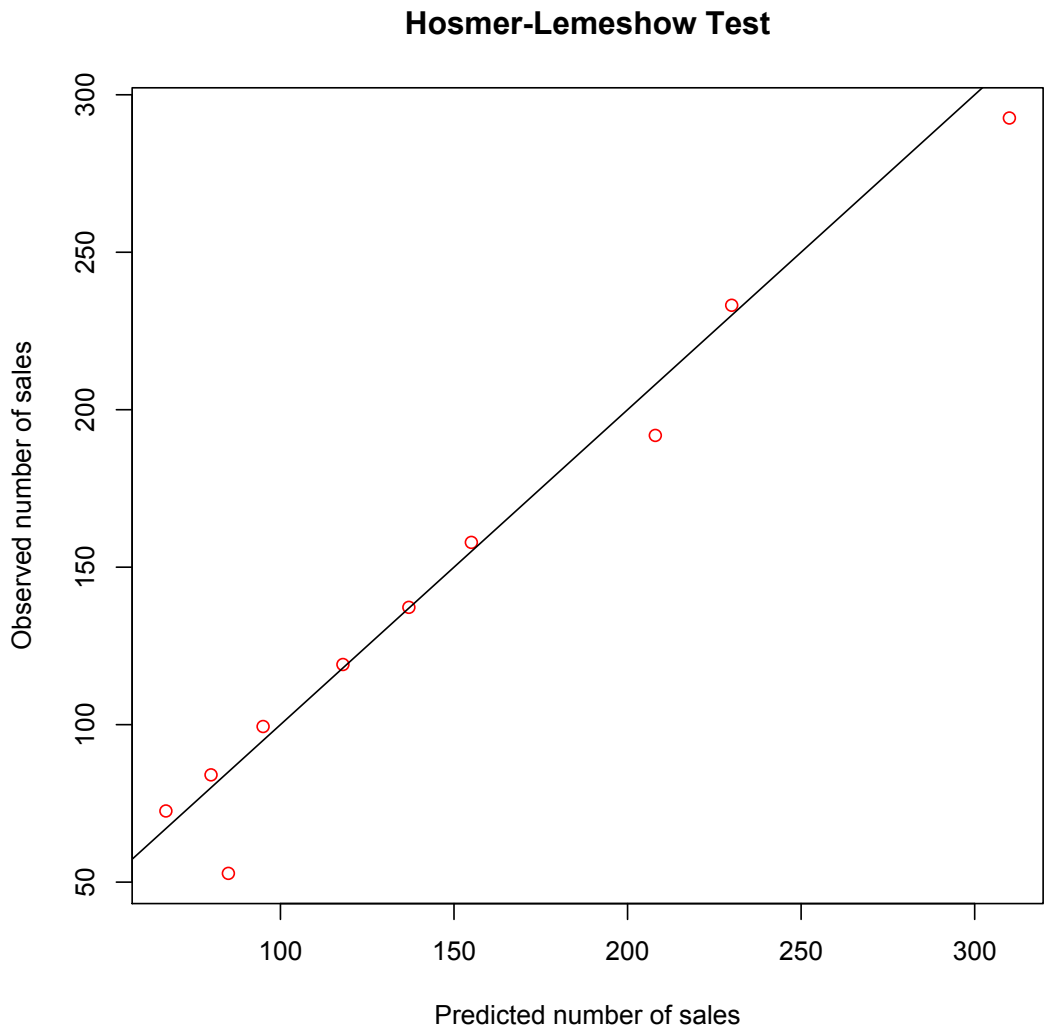
$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

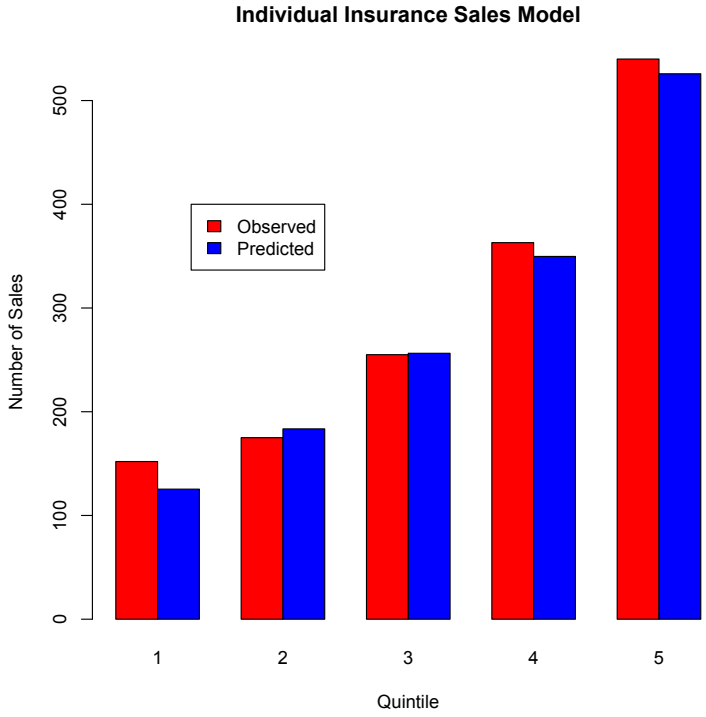
I det generella fallet:

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{1 \le i_1 < \cdots < i_k \le n} |A_{i_1} \cap \cdots \cap A_{i_k}| \right).$$

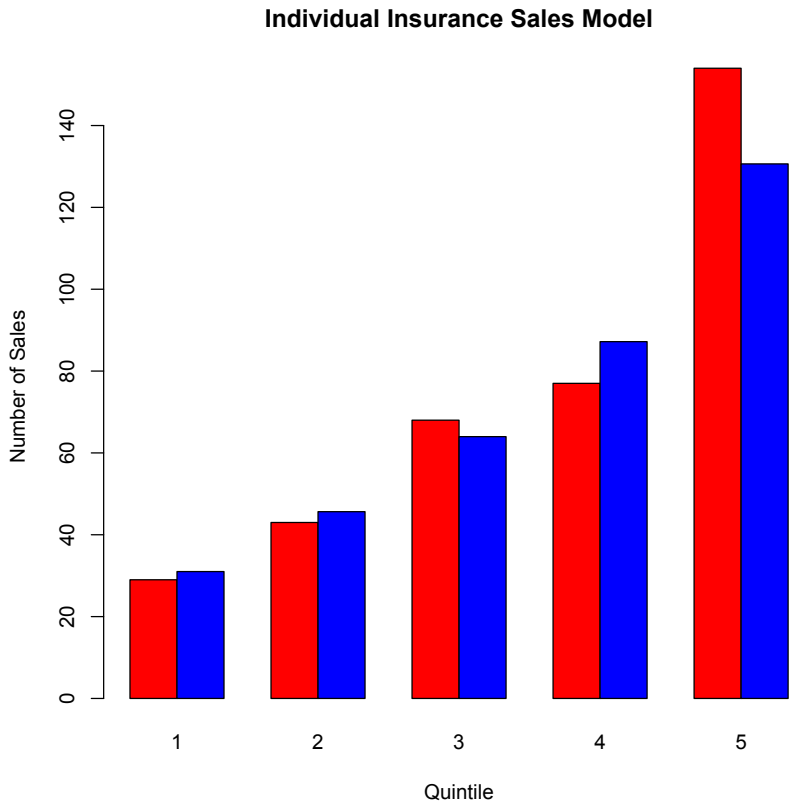
Den kombinerade sannolikheten anger sannolikheten att en försäljning kommer ske av en eller flera specifika försäkringar. Man kan lätt modifiera denna modell och enbart ta med de försäkringar man är intresserad av att sälja vid ett specifikt tillfälle.

In-sample


Hosmer-Lemeshow Test
Observed number of sales
Predicted number of sales
50
100
150
200
250
300
100
150
200
250
300


Individual Insurance Sales Model
Number of Sales
0
100
200
300
400
500
Observed
Predicted
1
2
3
4
5
Quintile

Out-of-sample



KPI:er - Out-of-sample

	Mean group	Model group	X-times Improvment
Customer	1533	1533	-
Sales	74	154	208%
Products Sold	117	243	208%
Premium	322 600 DKK	599 000 DKK	186%
Sale/Customer	4,8%	10,0%	208%
Product Sold/Sale	1,58	1,58	100%
Premium/Customer	210 DKK	391 DKK	186%
Premium/Sale	4 359 DKK	3 890 DKK	89%

4 Fallstudie

En workshop hölls på företaget, där berörda chefer närvarade. Efter presentation och utvärdering av modellerna valdes modell 3 – Total Premium Model med logistisk regression som implementationsmodell.

En kundbas på 4300 hushåll rankades enligt modellen. 1300 av dessa valdes ut slumpmässigt och 1300 valdes ut baserat på dess ranking. Eftersom slumpgruppen väljs slumpmässigt över hela den rankade kundbasen innebär det att ett antal kunder ingår i båda grupperna, och därmed dubbelräknas. Totalt sett innebär det att färre än 2600 hushåll kontaktades i studien.

De hushåll som ingick i kampanjen lades in anonymt i säljsystemet så att den personal som hanterade försäljningen inte hade möjlighet att se vilken grupp kunderna tillhörde. Hushållen kontaktades av säljpersonalen under tidsperioden oktober 2012 – januari 2013.

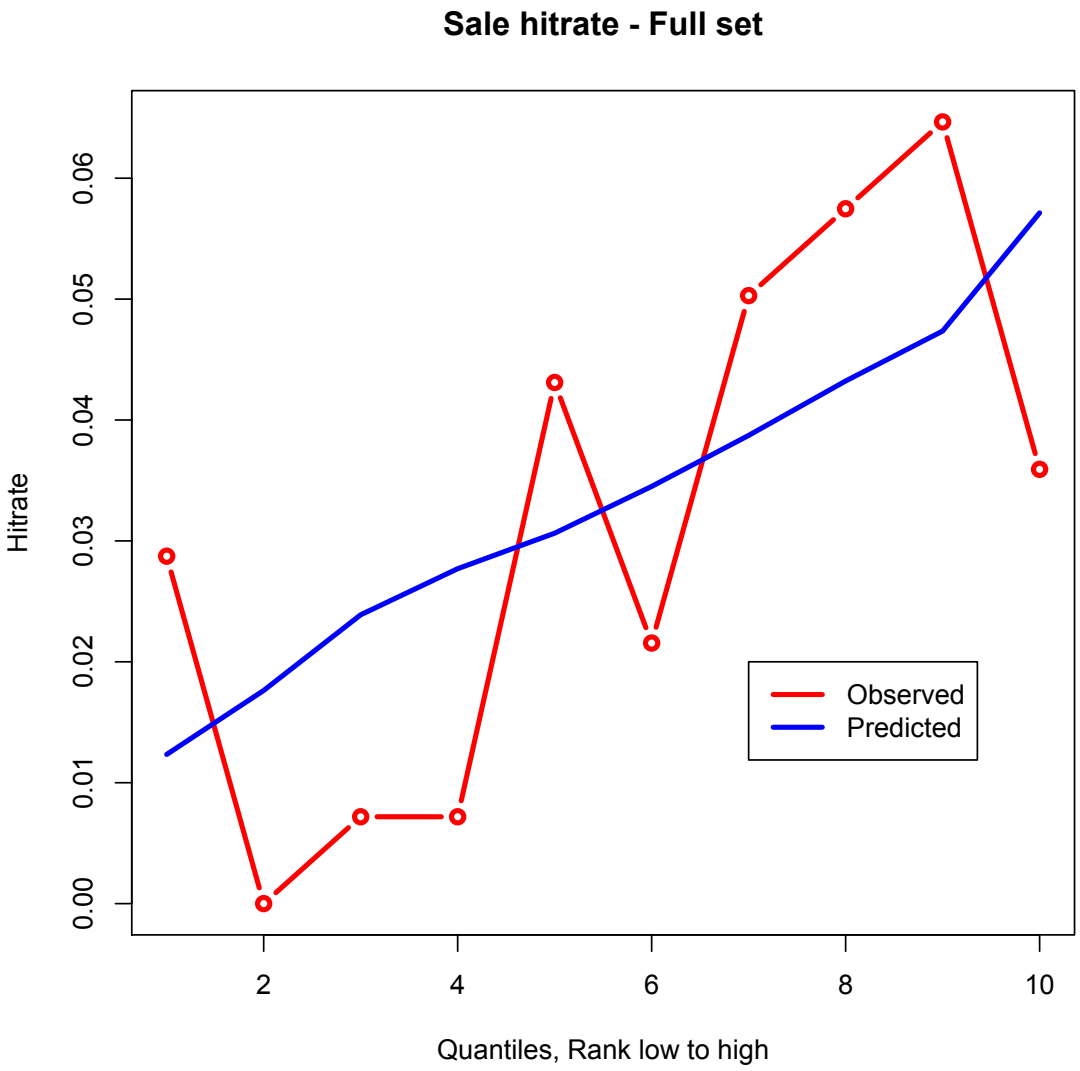
	Random Group	Model Group	X-times Improvment
Customer	1300	1300	
Closure	682	711	104%
Sales	18	32	178%
Sales/Closure	2,6%	4,5%	173%
Products Sold	28	52	186%
Product Sold/Closure	4,1%	7,3%	178%
Premium	77 000,00 DKK	144 000,00 DKK	187%
Premium/Closure	133,00 DKK	203,00 DKK	153%
Premium/Sale	4 280,00 DKK	4 500,00 DKK	105%
Product Sold/Sale	1,6	1,6	100%

För att se om det är någon statistisk säkerställd skillnad mellan de två grupperna används tvåsidigt Fishers exakta test på Sale.

	Random Set	Model Set
Sale	18	32
No Sale	1282	1268
P-value	0.0623	

P-värdet antyder att det med en 95 % konfidensnivå inte går att säkerställa att det finns en statistisk skillnad mellan grupperna. Dock är det mycket nära denna gräns.

Nedan visas predikterad försäljningssannolikhet mot observerad. I nedanstående fall ingår både modellgruppen och slumpgruppen.


Sale hitrate - Full set
Hitrate
0.00
0.01
0.02
0.03
0.04
0.05
0.06
2
4
6
8
10
Quantiles, Rank low to high
Observed
Predicted

5 Diskussion och slutsatser

5.1 Modellerna

Rent generellt framgår det att modellerna byggda med LR presterar bättre än ZINBR i den högsta quintilen, både in-sample och out-of-sample. Förvånande nog har även samtliga modeller, byggda med LR, ett högre observerat än predikterat värde i den högsta quintilen. Detta är något som inte upplevs som positivt ur ett teoretiskt kvalitetsperspektiv men positivt ur ett praktiskt perspektiv, då syftet är att uppnå så hög försäljning som möjligt.

KPI:erna ger en god insikt i hur bra modellerna presterar utanför de områden som de är optimerade för. Detta är intressant ur ett försäljningsperspektiv men säger lite om hur pass bra modellerna är på att beskriva det de är tänkta för. Generellt framgår även här att modellerna byggda med LR ger bättre resultat än modellerna byggda med ZINBR.

En förklaring till varför LR-modellerna är bättre än ZINBR-modellerna kan vara att dessa delats upp i delmodeller som var för sig haft egna förklarande variabler. T.ex. förklaras kanske den typ av kunder som köper dyra försäkringar med ett set av variabler och de som köper billiga av ett annat. ZINBR modellerna har inte den möjlighet att fånga upp denna karakteristik som LR-modellerna.

Multinominal logistisk regression har inte omnämnts som alternativ, då det visade sig bättre att använda flera delmodeller med logistisk regression. Anledningen till detta är troligtvis samma som tidigare nämnts, att kunder som köper dyra försäkringar förklaras med ett set av variabler och de som köper billiga av ett annat. Vid Multinominal logistisk regression måste samma förklarande variabler användas för att förklara samtliga utfall av responsvariablen.

5.2 Fallstudie

Eftersom modellen inte var utformad för att optimera försäljning, utan premieintjäning, är det inte riktigt rättvist att jämföra modell och kontrollgrupp med försäljning. Dock är det svårt att utforma ett test som mäter om premieintjäningen skiljer sig signifikant från kontrollgruppen.

Gruppen som rankades var tyvärr inte så pass stor att modellgruppen kunde bestå av de 20 % högst rankade kunderna, då grupperna i så fall skulle bli så små att det var svårt att mäta signifikanta skillnader. En cut-off nivå på 30 % valdes som kompromiss vilket innebar att varje grupp fick 1300 hushåll.

Den rankade gruppen var inte heller så heterogen som analysgruppen, då företaget hade vissa hygien-och engagemangsregler vid detta uttag. Detta innebar inga problem för utvärderingen av studien då analysgruppen var så stor att även dessa mer homogena hushåll fick en valid ranking. Dock innebar det att gruppen i sig hade lägre förväntad försäljningssannolikhet än analysgruppen.

Modellen som användes var en variant på Total Premium Model med logistisk regression med en slutgiltig score som inte har en direkt koppling till ett faktiskt försäljningsvärde. Det gick därför inte utvärdera modellen genom att jämföra predikterade värden med observerade. Istället har modellen utvärderats genom att även använda Sales Model på gruppen och visa på skillnaden mellan prediktera försäljningssannolikhet och observerade försäljningssannolikhet. Dock är det inte med hjälp av denna modells score som hushållen valdes ut på vilket gör detta test en aning missvisande.

5.3 Möjliga utvecklingar

- Modellernas förklarande kraft kan öka ifall det går att identifiera fler förklarande variabler som antyder att kunden är i behov av att köpa försäkringar. Signaler som ofta tyder på detta är t.ex. att kunden köpt ny bil, flyttat eller har utlöpande försäkringar hos ett annat försäkringsbolag.

- Modellerna predikterar vilka kunder som har hög sannolikhet att köpa ytterligare försäkringar, men säger inget om när kunden bör kontaktas. En möjlig utvecklig är att på något sätt ta med tidsaspekten i modellerna.

- Försäljningssannolikheten bör kombineras med en modell för kundvärde, se 1.1 Bakgrund. En utveckling är att beräkna kundens förväntade förändring av LTV vid eventuell försäljning och multiplicera detta värde med förväntad försäljningssannolikhet och sedan välja ut kunder med högst värde till kampanj.

- I detta arbete har olika varianter av GLM används, men det är mycket möjligt att det finns andra modeller som är lämpliga att använda

- Detta arbete har enbart fokuserat på utgående telefonförsäljning. Det finns dock andra säljkanaler som hade haft nytta av en motsvarande modell. Även delförsäljning och uppgraderingar bör ingå i en komplett modell.

6 Referenser

1. Trowbridge, C. L. (1989). *Fundamental Concepts of Actuarial Science, Revised Edition.* Actuarial Education and Research Fund.
2. Akura, M.T., Srinivasan, K. (2005). *Research Note: Customer Intimacy and Cross-Selling Strategy*. Management Science, 51(6), 1007-1012.
3. Kamakura, W.A., Wedel, M., de Rosa, F., Mazzon, J.A. (2003). *Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction.* International Journal of Research in Marketing, 20, 45-65.
4. Larivi`ere, B., Van den Poel, D. (2004). *Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services.* Expert Systems with Applications, 27(2), 277-285.
5. Ahn, H., Ahn, J.J., Oh, K.J., Kim, D.H. (2011). *Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques.* Expert Systems with Applications, 38(5), 5005-5012.
6. Sen, A., Srivastava, M. (2011). *Regression Analysis — Theory, Methods, and Applications*, Springer-Verlag
7. Madsen, H., Thyregod, P. (2011). *Introduction to General and Generalized Linear Models*. Chapman & Hall
8. Green, W.H. (1994). *Accounting for excess zeros and sampling selection in Poisson and negativ binomial regression models.* Working Paper NO 94-10. New York University, Department of Econometrics.
9. Mood, A.M., Graybill, F.A. (1963). *Introduction to the Theory of Statistics*, 2nd edition. McGraw-Hill
10. Vuong, Q.H. (1989). *Likelihood Ratio Tests for Model Selection and non-nested Hypotheses.* Econometrica 57 (2), 307–333.
11. Akaike, H. (1974). *A new look at the statistical model identification.* IEEE Transactions on Automatic Control 19 (6), 716–723.

12. Hosmer, D.W., Lemeshow, S (2000). *Applied Logistic Regression*. New York: Wiley.
13. Ismail, B., Jemain, A.A. (2007). *Handling overdispersion with negative binomial and generalized poisson regression models*. Casualty Actuarial Society Forum, Winter 2007, 103-158.
14. Böiers, L-C. (2003). *Diskret matematik*. Lund: Studentlitteratur.

Bilagor – Kod

Modell 1 – Responsvariabel Total Sales

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8753  -0.3568  -0.2818  -0.2328   3.0189

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.69311    0.42248 -11.109  < 2e-16 ***
AgeYoungYoung      0.53156    0.16228   3.276 0.001054 **
AgeMidMid          0.58490    0.15223   3.842 0.000122 ***
AgeOldOld          0.79742    0.16171   4.931 8.17e-07 ***
EtageApartment     0.24564    0.06515   3.770 0.000163 ***
CarOne             0.58243    0.10099   5.767 8.07e-09 ***
CarMore            0.78472    0.14593   5.378 7.55e-08 ***
UnionNÆS           0.57970    0.09736   5.954 2.61e-09 ***
UnionDJØ           0.32328    0.11169   2.895 0.003798 **
UnionUnion         0.32029    0.07114   4.502 6.72e-06 ***
Prem               0.09532    0.05119   1.862 0.062573 .
MCMore             0.31698    0.09464   3.349 0.000810 ***
SummerHouseMore    1.97797    0.74897   2.641 0.008268 **
AccidentOneOne    -2.19659    0.54440  -4.035 5.46e-05 ***
AccidentMoreMore  -1.89361    0.53935  -3.511 0.000447 ***
HomeOtherMore     -0.15910    0.08880  -1.792 0.073196 .
HouseAccidentMore -0.46850    0.12121  -3.865 0.000111 ***
HomeCarMore        0.37331    0.13491   2.767 0.005655 **
CarAccidentMore   -0.23623    0.10561  -2.237 0.025290 *
CarAgeMore        -0.39708    0.14548  -2.729 0.006344 **
SummerHousePrem   -0.25993    0.09942  -2.615 0.008935 **
AccidentPrem       0.22914    0.07218   3.175 0.001500 **
InhabTwoTwo       -0.20101    0.06731  -2.986 0.002824 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11896  on 30671  degrees of freedom
Residual deviance: 11487  on 30649  degrees of freedom
AIC: 11533
```

Modell 3 – Responsvariabel Total Premium Sales

ZINBR Modellen

```
Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.3030 -0.2058 -0.1719 -0.1393 29.2926

Count model coefficients (negbin with log link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.45069    0.35207   1.280 0.200512
AgeYoungYoung    0.25092    0.18719   1.340 0.180089
AgeMidMid        0.48266    0.17485   2.761 0.005771 **
AgeOldOld        0.17329    0.18043   0.960 0.336855
UnionNÆS         0.23990    0.11388   2.107 0.035148 *
UnionDJØ         0.15052    0.12223   1.231 0.218179
UnionUnion       0.31350    0.07826   4.006 6.18e-05 ***
Prem            -0.03040    0.04448  -0.683 0.494314
HomeMore        -4.54778    0.93140  -4.883 1.05e-06 ***
TotP1More       -0.45297    0.29355  -1.543 0.122808
TotP23Two-Three  0.78477    0.30500   2.573 0.010082 *
AccidentOneOne   0.34216    0.07595   4.505 6.64e-06 ***
HomeOtherMore   -0.41848    0.25785  -1.623 0.104596
CarAccidentMore -0.49403    0.13479  -3.665 0.000247 ***
Prem:HomeMore    0.48046    0.11409   4.211 2.54e-05 ***
Log(theta)       0.38736    0.13429   2.885 0.003920 **

Zero-inflation model coefficients (binomial with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      4.63421    0.36170  12.812  < 2e-16 ***
AgeYoungYoung   -0.62037    0.17439  -3.557 0.000375 ***
AgeMidMid       -0.38486    0.16394  -2.348 0.018892 *
AgeOldOld       -0.59257    0.17273  -3.431 0.000602 ***
UnionNÆS        -0.48933    0.11243  -4.352 1.35e-05 ***
UnionDJØ        -0.30058    0.12491  -2.406 0.016111 *
UnionUnion      -0.30079    0.07773  -3.870 0.000109 ***
Prem            -0.19065    0.04420  -4.313 1.61e-05 ***
HomeMore        -0.90284    0.95798  -0.942 0.345963
TotP1More        0.37451    0.29530   1.268 0.204727
TotP23Two-Three -0.16688    0.30886  -0.540 0.588990
AccidentOneOne   0.81333    0.07705  10.555  < 2e-16 ***
HomeOtherMore    0.25331    0.26386   0.960 0.337058
CarAccidentMore -0.04612    0.13434  -0.343 0.731385
Prem:HomeMore    0.04831    0.11725   0.412 0.680295
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 1.4731
Number of iterations in BFGS optimization: 53
Log-likelihood: -8372 on 31 Df
```

Logistisk Regressions Modellen

Tot Prem Low

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7770  -0.1971  -0.1342  -0.1017   3.4874

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.500423   0.924881  -7.028 2.09e-12 ***
AgeMidMid            0.373301   0.140309   2.661 0.007801 **
AgeOldOld            0.693397   0.170516   4.066 4.77e-05 ***
EtageApartment       4.008526   0.952919   4.207 2.59e-05 ***
Prem                 0.177606   0.117941   1.506 0.132094
HomeMore             1.145849   0.156024   7.344 2.07e-13 ***
UnionNÆS             0.595445   0.137747   4.323 1.54e-05 ***
UnionDJØ            -0.004178   0.213120  -0.020 0.984358
UnionUnion           0.249352   0.127244   1.960 0.050038 .
HouseMore           -0.478032   0.205740  -2.323 0.020153 *
MCMore               3.108316   1.340741   2.318 0.020430 *
AccidentOneOne      -2.480266   1.070861  -2.316 0.020551 *
AccidentMoreMore    -2.758283   1.081770  -2.550 0.010779 *
HomeOtherMore       -0.452703   0.141789  -3.193 0.001409 **
HouseAgeMore         1.026747   0.399209   2.572 0.010113 *
HomeCarMore         -0.723761   0.366658  -1.974 0.048389 *
HomeAccidentMore    -1.377404   0.253515  -5.433 5.53e-08 ***
CarPrem              0.083431   0.021208   3.934 8.35e-05 ***
BoatPrem            -0.074806   0.029018  -2.578 0.009939 **
MCPrem              -0.410440   0.195480  -2.100 0.035759 *
SummerHousePrem     -0.038583   0.017003  -2.269 0.023253 *
AccidentPrem         0.259957   0.139118   1.869 0.061676 .
InhabTwoTwo         -0.391692   0.113946  -3.438 0.000587 ***
EtageApartment:Prem -0.438668   0.113593  -3.862 0.000113 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5190.1  on 30671  degrees of freedom
Residual deviance: 4688.1  on 30648  degrees of freedom
AIC: 4736.1
```

Tot Prem Mid

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4632  -0.1953  -0.1523  -0.1218   3.3336
```

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -5.13149    0.16442 -31.209  < 2e-16 ***
AgeYoungYoung              0.41894    0.15683   2.671  0.00755 **
AgeMidMid                  0.36516    0.13443   2.716  0.00660 **
EtageApartment             0.84216    0.12379   6.803 1.02e-11 ***
HomeMore                   0.48189    0.17324   2.782  0.00541 **
UnionNÆS                   0.11970    0.19023   0.629  0.52921
UnionDJØ                   0.36985    0.17162   2.155  0.03116 *
UnionUnion                 0.25419    0.11858   2.144  0.03207 *
CarOneOne                  0.81647    0.11532   7.080 1.44e-12 ***
CarMoreMore                1.26380    0.18379   6.876 6.15e-12 ***
AccidentOneOne            -0.28671    0.11758  -2.438  0.01475 *
CarAccidentMore           -0.53482    0.19135  -2.795  0.00519 **
MCPrem                     0.07124    0.02092   3.406  0.00066 ***
InhabTwo                  -0.38854    0.13058  -2.976  0.00293 **
InhabMore                 -0.29145    0.15374  -1.896  0.05800 .
EtageApartment:HomeMore   -1.12881    0.22833  -4.944 7.66e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5026.1  on 30671  degrees of freedom
Residual deviance: 4782.5  on 30656  degrees of freedom
AIC: 4814.5
```

Tot Prem High

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4371  -0.2048  -0.1609  -0.1276   3.5171

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -7.63961    0.74616 -10.239  < 2e-16 ***
AgeYoungYoung         0.82424    0.30956   2.663 0.007754 **
AgeMidMid             0.82336    0.29372   2.803 0.005060 **
AgeOldOld             1.09085    0.30815   3.540 0.000400 ***
EtageApartment       -0.28111    0.11791  -2.384 0.017121 *
UnionNÆS              1.07589    0.20386   5.278 1.31e-07 ***
UnionDJØ              1.06486    0.23404   4.550 5.37e-06 ***
UnionUnion            0.83351    0.14812   5.627 1.83e-08 ***
CarOne                0.50596    0.21667   2.335 0.019537 *
CarMore               1.08697    0.34555   3.146 0.001657 **
TotP1More            -1.10637    0.28850  -3.835 0.000126 ***
TotP23Two-Three       1.19228    0.28840   4.134 3.56e-05 ***
Prem                  0.28739    0.08474   3.392 0.000695 ***
SummerHouseMore       2.24078    0.93279   2.402 0.016295 *
AccidentMoreMore      0.34485    0.17254   1.999 0.045648 *
```

```
HomeOtherMore         -0.92721    0.27562  -3.364 0.000768 ***
HouseAccidentMore     -0.45990    0.26326  -1.747 0.080647 .
HomeCarMore            0.80380    0.40177   2.001 0.045429 *
HomeAccidentMore       0.49227    0.26017   1.892 0.058474 .
HomeAgeMore           -0.78358    0.35185  -2.227 0.025946 *
CarAgeMore            -0.73560    0.32410  -2.270 0.023230 *
GenderFemale          -0.21191    0.11323  -1.871 0.061282 .
UnionNÆS:CarOne       -0.87012    0.35362  -2.461 0.013871 *
UnionDJØ:CarOne       -1.04092    0.38664  -2.692 0.007099 **
UnionUnion:CarOne     -0.39000    0.24367  -1.601 0.109478
UnionNÆS:CarMore      -1.16599    0.50812  -2.295 0.021749 *
UnionDJØ:CarMore      -2.17972    0.79858  -2.729 0.006343 **
UnionUnion:CarMore    -1.12325    0.41112  -2.732 0.006292 **
Prem:SummerHouseMore -0.22541    0.11330  -1.989 0.046658 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4993.1  on 30671  degrees of freedom
Residual deviance: 4812.8  on 30643  degrees of freedom
AIC: 4870.8
```

Modell 4 – Responsvariabel Individual Insurance Sales

Responsvariabel – House

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5358  -0.1231  -0.0646  -0.0367   4.1078

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -7.19080    0.83931  -8.568  < 2e-16 ***
AgeMidMid                  0.59014    0.22416   2.633  0.00847 **
AgeOldOld                  1.30339    0.27852   4.680 2.87e-06 ***
EtageApartment            -2.42100    0.33671  -7.190 6.47e-13 ***
TotPMMore                 -1.07911    0.43777  -2.465  0.01370 *
UnionNÆS                   1.20558    0.26150   4.610 4.02e-06 ***
UnionDJØ                   1.99795    0.25404   7.865 3.70e-15 ***
UnionUnion                 1.01076    0.20395   4.956 7.20e-07 ***
Prem                       0.18104    0.09803   1.847  0.06476 .
HouseMore                 -1.68699    0.37191  -4.536 5.73e-06 ***
AccidentOne                0.35640    0.19217   1.855  0.06366 .
AccidentMore               0.50429    0.23282   2.166  0.03031 *
HomeOtherMore             -0.73528    0.35357  -2.080  0.03756 *
HomeAgeMore               -1.17973    0.51176  -2.305  0.02115 *
CarAgeMore                -0.81548    0.42772  -1.907  0.05658 .
EtageApartment:TotPMMore   1.77737    0.86919   2.045  0.04087 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2249.6  on 30671  degrees of freedom
Residual deviance: 1959.6  on 30656  degrees of freedom
AIC: 1991.6
```

Responsvariabel – Home

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6182  -0.2601  -0.1734  -0.0348   4.2202

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.57089    0.17222 -26.541  < 2e-16 ***
AgeMidMid       0.30867    0.09854   3.132 0.001734 **
AgeOldOld       0.80029    0.14560   5.497 3.87e-08 ***
EtageApartment  0.52912    0.09315   5.680 1.35e-08 ***
UnionNÆS        0.42126    0.16838   2.502 0.012357 *
UnionDJØ        0.52959    0.16040   3.302 0.000961 ***
UnionUnion      0.45280    0.09950   4.551 5.35e-06 ***
TotP1More      -1.58962    0.63471  -2.504 0.012263 *
```

```
TotP23Two-Three  1.33777    0.58064   2.304 0.021225 *
HouseMore        0.58787    0.19272   3.050 0.002285 **
HomeMore         4.13135    1.93827   2.131 0.033051 *
CarOne           0.98029    0.14704   6.667 2.62e-11 ***
CarMore          1.53392    0.32917   4.660 3.16e-06 ***
MCMore           0.33990    0.19030   1.786 0.074076 .
SummerHouseMore  0.48573    0.15144   3.207 0.001339 **
AccidentOne      0.36354    0.15927   2.283 0.022460 *
AccidentMore     0.71975    0.30419   2.366 0.017975 *
CarAgeMore      -0.58516    0.24500  -2.388 0.016919 *
HomePrem        -1.05609    0.26222  -4.028 5.64e-05 ***
InhabTwo        -0.39539    0.12207  -3.239 0.001199 **
InhabMore       -0.34416    0.16695  -2.061 0.039259 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6265.8  on 30671  degrees of freedom
Residual deviance: 5477.1  on 30651  degrees of freedom
AIC: 5519.1
```

Responsvariabel – Car

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4378  -0.2206  -0.1772  -0.1479   3.3080

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.70325    0.69642  -9.625  < 2e-16 ***
AgeYoungYoung        0.58353    0.26172   2.230 0.025776 *
AgeMidMid            0.80469    0.23890   3.368 0.000756 ***
AgeOldOld            0.96214    0.24943   3.857 0.000115 ***
EtageApartment      -0.25249    0.11304  -2.234 0.025503 *
UnionNÆS             0.77225    0.16578   4.658 3.19e-06 ***
UnionDJØ             0.47907    0.19833   2.416 0.015711 *
UnionUnion           0.75376    0.12501   6.030 1.64e-09 ***
TotP1More           -0.91287    0.30261  -3.017 0.002556 **
TotP23Two-Three      0.93904    0.24272   3.869 0.000109 ***
HouseMore            0.36759    0.16985   2.164 0.030449 *
CarOneOne            0.41406    0.16935   2.445 0.014484 *
CarMoreMore          0.65743    0.24179   2.719 0.006549 **
MCMore               0.48778    0.16334   2.986 0.002824 **
Prem                 0.14751    0.08361   1.764 0.077684 .
SummerHouseMore      2.58415    0.82793   3.121 0.001801 **
HomeAccidentMore     0.34065    0.18283   1.863 0.062438 .
CarAgeMore          -0.71509    0.21666  -3.300 0.000965 ***
BoatPrem             0.05979    0.02291   2.610 0.009063 **
AccidentOne          0.28859    0.16246   1.776 0.075674 .
```

```
AccidentMore          0.67911    0.20621   3.293 0.000990 ***
HomeUnionMore        -0.71414    0.17015  -4.197 2.70e-05 ***
Prem:SummerHouseMore -0.23065    0.09709  -2.376 0.017516 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5664.8  on 30671  degrees of freedom
Residual deviance: 5512.2  on 30649  degrees of freedom
AIC: 5558.2
```

Responsvariabel – Boat

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2507  -0.0146  -0.0146  -0.0146   4.2765

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.1439     0.5772 -15.843  < 2e-16 ***
BoatMore      5.6999     0.9224   6.179 6.44e-10 ***
BoatPrem     -0.2350     0.1038  -2.264   0.0236 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 318.68  on 30671  degrees of freedom
Residual deviance: 251.92  on 30669  degrees of freedom
AIC: 257.92
```

Responsvariabel – MC

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3003  -0.0708  -0.0555  -0.0331   3.9083

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.1787     1.1293  -4.586 4.52e-06 ***
AgeMidMid        1.7841     0.5283   3.377 0.000732 ***
AgeOldOld        2.1057     0.5806   3.627 0.000287 ***
Prem            -0.3765     0.1355  -2.779 0.005450 **
BoatMore         0.5870     0.3555   1.651 0.098640 .
CarOneOne        0.9252     0.3593   2.575 0.010015 *
MCMore           1.5111     0.2767   5.461 4.73e-08 ***
OtherMotorMore   0.8282     0.4504   1.839 0.065974 .
CarAgeMore      -1.3485     0.6351  -2.123 0.033740 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 954.79  on 30671  degrees of freedom
Residual deviance: 887.98  on 30663  degrees of freedom
AIC: 905.98
```

Responsvariabel - Summer House

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2399  -0.0783  -0.0634  -0.0400   3.9723

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -9.7484     1.3468  -7.238 4.55e-13 ***
AgeMidMid        1.5463     0.4426   3.494 0.000476 ***
AgeOldOld        1.3413     0.5361   2.502 0.012350 *
Prem             0.2735     0.1542   1.774 0.076118 .
TotPMMore       -1.2481     0.4334  -2.880 0.003976 **
UnionNÆS         0.9147     0.3578   2.556 0.010580 *
UnionDJØ         0.7800     0.4371   1.784 0.074343 .
UnionUnion       0.3251     0.2972   1.094 0.274073
OtherMotorMore   1.2410     0.4378   2.835 0.004589 **
CarAgeMore      -1.1538     0.6033  -1.912 0.055823 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1003.6  on 30671  degrees of freedom
Residual deviance:  962.2  on 30662  degrees of freedom
AIC: 982.2
```

Responsvariabel - Other Motor

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1646  -0.0557  -0.0553  -0.0295   3.9343

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.7391     0.5026 -15.397  < 2e-16 ***
AgeMidMid        1.2570     0.5364   2.343 0.019112 *
AgeOldOld        1.2702     0.5698   2.229 0.025812 *
MCMore           0.8124     0.3760   2.161 0.030703 *
OtherMotorMore   1.3622     0.3923   3.472 0.000516 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 716.08  on 30671  degrees of freedom
Residual deviance: 691.68  on 30667  degrees of freedom
AIC: 701.68
```

Responsvariabel – Accident

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6912  -0.2174  -0.1495  -0.0746   3.8885

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -6.70640    0.65478 -10.242  < 2e-16 ***
AgeYoungYoung    0.53328    0.18706   2.851 0.004360 **
AgeMidMid        0.40696    0.16153   2.519 0.011756 *
EtageApartment   0.29251    0.11080   2.640 0.008288 **
UnionNÆS         0.81511    0.14218   5.733 9.86e-09 ***
UnionDJØ         0.42821    0.18735   2.286 0.022276 *
UnionUnion       0.42361    0.12087   3.505 0.000457 ***
Prem             0.30863    0.08387   3.680 0.000233 ***
TotP1More        1.14577    0.22574   5.076 3.86e-07 ***
TotP23Two-Three -1.09163    0.18012  -6.061 1.36e-09 ***
HouseMore       -0.37526    0.15703  -2.390 0.016863 *
CarOneOne       -0.42326    0.14785  -2.863 0.004200 **
CarMoreMore     -0.76785    0.23524  -3.264 0.001098 **
BoatMore        -0.62710    0.18116  -3.462 0.000537 ***
SummerHouseMore -0.47358    0.12959  -3.654 0.000258 ***
RestMore        -0.71280    0.31789  -2.242 0.024940 *
HomeAgeMore      0.96036    0.18970   5.063 4.14e-07 ***
AccidentOneOne  -0.64750    0.25930  -2.497 0.012521 *
AccidentPrem    -0.28715    0.03299  -8.705  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5569.3  on 30671  degrees of freedom
Residual deviance: 4937.8  on 30653  degrees of freedom
AIC: 4975.8
```

Responsvariabel – Rest

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1393  -0.0376  -0.0235  -0.0235   4.1748

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.2559     0.3958 -18.330   <2e-16 ***
AgeMidMid       -0.9368     0.5678  -1.650   0.0989 .
```

```
AgeOldOld       -2.4356     1.1061  -2.202   0.0277 *
HouseMore        0.9771     0.5772   1.693   0.0905 .
OtherMotorMore   1.6485     0.6995   2.357   0.0184 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 273.86  on 30671  degrees of freedom
Residual deviance: 260.50  on 30667  degrees of freedom
AIC: 270.5
```