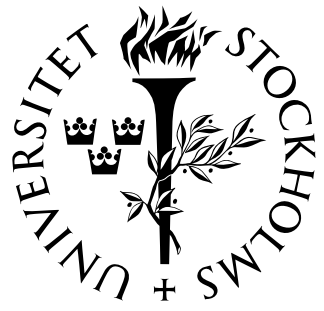

Matematisk statistik
Stockholms universitet

Analys av köpviljan avseende försäkring med logistisk regression och bootstrap

Anna Sandler

Examensarbete 2007:11

Postadress:

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm
Sverige

Internet:

http://www.math.su.se/matstat

Matematisk statistik
Stockholms universitet
Examensarbete **2007:11**,
http://www.math.su.se/matstat

Analys av köpviljan avseende försäkring med logistisk regression och bootstrap

Anna Sandler*

Juni 2007

Sammanfattning

För försäkringsbolag är det vanligt att försäljning av försäkringar sker via telemarketing. Då är det viktigt att veta vilka kunder försäkringsbolag ska kontakta för att få det bästa försäljningsresultatet. Detta har lett till att ett försäkringsbolag vill undersöka vad det är som styr köpviljan för olika försäkringar. I detta arbete tas modeller fram som förklarar köpviljan för bil-, hus-, familj- och olycksförsäkring. Modeller tas fram som dels bygger på ett verkligt datamaterial och dels på ett datamaterial som är framskattat med hjälp av metoden bootstrap. Det framskattade datamaterialet består av flera observationer än det verkliga. I detta arbete vill man undersöka om det framskattade datamaterialet ger trovärdiga modeller i jämförelse med det verkliga datamaterialet. I så fall kan försäkringsbolaget använda sig av ett litet befintligt datamaterial istället för ett stort och sedan på det lilla datamaterialet skatta fram fler observationer med hjälp av bootstrap, vilket försäkringsbolaget skulle tjäna tid och pegnar på.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post: anna_sandler@hotmail.com. Handledare: Joanna Tyrcha.

Abstract

Insurance companies often use telemarketing to sell insurances. When using telemarketing, it is important to know who the insurance company will call in order to get the best selling result. Because of this, it is important to do research to establish the factors that determine the customers' willingness to buy. This study presents statistical models that explain willingness to buy car-, house-, family- and accident-insurances.

The statistical models are based on a database consisting of real observations and on a larger database consisting of estimated observations by using the method bootstrap, which is based on the on the real observations. The estimated database includes more observations than the real database and this study is aimed to see if the estimated database gives reliable models compared with the real database. If this is the case, the insurance company can use a small real database instead of a large real database and then estimate more observations to get a larger database by using bootstrap, which saves time and money.

Förord

Detta examensarbete omfattar 20 akademiska poäng och utgör del av min magisterexamen i matematisk statistik på Matematik-ekonomiprogrammet vid Stockholms universitet. Examensarbetet utfördes till stor del hos Trygg Hansa, Customer & Channel Development Nordic Personal Lines, i Stockholm under vårterminen 2007.

Trygg-Hansa är ett av Sveriges största sakförsäkringsbolag. Trygg Hansa ägs av det danska försäkringsbolaget Codan AS, som i sin tur ägs av Royal & Sun Alliance. Trygg Hansa erbjuder ett heltäckande sortiment av sakförsäkringar till privatpersoner och företag.

Jag vill här passa på att tacka alla som hjälpt mig att genomföra detta arbete. Först och främst ska Paul Andersson på Trygg Hansa ha ett stort tack för att han har givit mig detta uppdrag och låtit mig sitta på företaget och arbetat. Han har också varit en fantastisk bra handledare, som visat intresse för mitt arbete och hjälpt mig vid behov. Dessutom vill jag tacka Paul Andersson och hans kolleger för ett glatt och trevligt sällskap.

Sist men inte minst vill jag rikta ett stort tack till min handledare på institutionen för matematisk statistik vid Stockholms universitet Joanna Tyrcha, som alltid haft svar på mina frågor och som med stort engagemang hjälpt mig.

Innehållsförteckning

1 Bakgrund

De flesta försäkringsbolag eftersträvar ökade försäkringsvolymer. Detta betyder bland annat att de vill sälja så många försäkringar som möjligt. De vill också att deras kunder skall vara nöjda, vilket innebär att försäkringarna skall täcka kundernas behov och önskemål.

Försäkringar kan säljas på olika sätt. Ett sätt är att befintliga och eventuellt blivande kunder kontaktar försäkringsbolagen och köper de försäkringar de vill ha. Ett annat sätt, som är det intressanta i det här examensarbetet, är att försäkringsbolagen kontaktar personer via telefon och på så sätt försöker sälja försäkringar. Detta kallas telemarketing.

Vid telemarketing är det viktigt för försäkringsbolagen att veta vilka personer som kan vara villiga att köpa en viss försäkring och vilka som inte är det. Annars är risken stor att försäljningsinsatsen inte ger ett gott resultat. Med anledning av detta vill försäkringsbolagen undersöka vad det är som styr köpviljan avseende olika försäkringar. Man vill veta om det finns några faktorer som styr köpviljan positivt och om det finns några faktorer som styr negativt. Med kännedom om detta kan försäkringsbolagen vinna fler kunder genom att koncentrera sina försäljningsinsatser mot de köpvilliga personerna.

Att undersöka köpviljan avseende försäkringar är en statistisk uppgift, dvs den går att få fram med statistiska modeller. För att kunna sätta upp en statistisk modell för köpviljan måste man ha ett datamaterial, som visar vilka personer som tidigare har köpt försäkring och vilka som inte har gjort det. Det skall också framgå ur datamaterialet vad som skiljer de olika personerna åt när det gäller de så kallade förklaringsvariablerna, som t ex kan vara ålder och sysselsättning. Därefter kan man sätta upp en statistisk modell för köpviljan.

När det gäller statistiska modeller blir modellerna alltid bättre ju mer datamaterial man har. Om observationerna består av endast tio personer kan man inte dra några säkra slutsatser om vilka personer som köper försäkring och vilka som inte gör det. Däremot om man har t ex 10 000 observationer så kan man få fram modeller som kan stämma bra överens med verkligheten.

Eftersom det är förenat med stora kostnader och mycket arbete samt tidskrävande att få fram ett stort datamaterial genom faktiska observationer vill det försäkringsbolag, som detta examensarbete har utförts hos, undersöka om man kan skatta fram fler dataobservationer med hjälp av metoden bootstrap och genom detta få fram trovärdiga modeller.

2 Syfte och mål

Syftet med detta examensarbete är att för ett försäkringsbolags räkning undersöka köpviljan avseende olika typer av sakförsäkringar, nämligen bil-, hus-, familje- och olycksförsäkring, utifrån ett datamaterial bestående av verkliga och skattade observationer.

Examensarbetets mål är att:

1. ta fram modeller som förklarar köpviljan avseende bil-, hus-, familje- och olycksförsäkring genom att använda logistisk regression på en befintlig mängd verkliga observationer

2. ur en befintlig mängd verkliga observationer skatta fram fler observationer genom att använda metoden bootstrap för att få ett större datamaterial

3. ta fram modeller som förklarar köpviljan avseende bil-, hus-, familje- och olycksförsäkring genom att använda logistisk regression på det datamaterial som tagits fram med bootstrap

4. jämföra modellerna som tagits fram enligt punkt 1 och punkt 3 och sedan avgöra med hjälp av de befintliga observationerna om bootstrap-modellerna är trovärdiga

Med hjälp av ovanstående punkter kan försäkringsbolaget få information om vilka variabler som påverkar köpviljan positivt. Denna information kan vara användbar vid försäljning av försäkringar genom telemarketing och troligtvis bidra till ett bättre försäljningsresultat.

3 Omfattning

Examensarbetet omfattar att ta fram modeller, som visar vilka variabler som påverkar köpviljan avseende bil-, hus-, familje- och olycksförsäkring med och utan bootstrap.

4 Avgränsning

I examensarbetet ingår inte att avgöra hur bra bootstrap-urvalen egentligen är. Det ingår inte heller någon djupare analys av varför signifikanta variabler påverkar köpviljan.

5 Förutsättning för examensarbetet

5.1 Datamaterial

Datamaterialet, som detta examensarbete grundar sig på, består av drygt 5 000 redan befintliga kunder, fysiska personer, hos ett försäkringsbolag. Dessa kunder har bil- och/eller husförsäkring hos försäkringsbolaget. De har blivit uppringda av försäkringsbolaget och då erbjudits köpa en eller flera försäkringar av annan typ, dvs försäkringar som de inte redan har hos försäkringsbolaget. Erbjudandet har gällt bil-, hus-, familje- och olycksförsäkringar.

Kunderna skiljer sig åt genom att de har dels olika så kallade förklaringsvariabler och dels olika så kallade responsvariabler. Förklaringsvariabler är de faktorer, som testas i de statistiska modellerna för att se om de påverkar köpviljan. Exempel på förklaringsvariabler här är ålder och postnummer. Responsvariabler är svar på erbjudandena, dvs om kunderna har köpt eller inte köpt de olika försäkringarna.

6 Datamaterial

6.1 Utökat datamaterial

Från början innehöll datamaterialet ett fåtal intressanta förklaringsvariabler. Fler förklarings-variabler har därför lagts till genom samkörning med andra databaser hos företaget. Alla dessa variabler är av typen sannolikhetsvariabler, t ex sannolikheten att ha barn, sannolikheten att äga en bil. De flesta förklaringsvariablerna är intervallvariabler, dvs sådana variabler där man kan ange numeriskt avstånd mellan värdena på mätskalan, t ex ålder.

För alla förklaringsvariabler har dock dummyvariabler införts och därmed görs ingen skillnad på om förklaringsvariablerna är intervall-, ordinala eller nominala variabler.

De förklaringsvariabler, som används i modellerna för att undersöka köpviljan är följande:

- Postnummer
- Ålder
- Antal år för bilförsäkringsinnehav
- Antal år för husförsäkringsinnehav
- Sannolikheten att man har barn
- Sannolikheten att man har bil
- Sannolikheten att man arbetar
- Sannolikheten att man har en lön som är max 400 063 kronor per år
- Sannolikheten att den äldsta i familjen är mellan 50-64 år

6.2 Gruppering av förklaringsvariabler i Knowledge seeker

Alla förklarande variabler grupperas efter andelen som tecknat försäkring. Det innebär att tex för variabeln ålder kan 20-30-åringar och 60-70-åringar hamna i samma grupp. Det beror på att båda har ungefär samma andel som tecknat försäkring. För att kunna göra denna uppdelning används här ett dataprogram kallat knowledge seeker. Detta dataprogram hittar antingen själv en bra uppdelning mellan olika procentsatser eller så får man göra gruppindelningen själv.
Gruppindelningen går till på följande sätt:

1. Man lägger in först in en responsvariabel och en förklarande variabel från datamaterialet till knowledge seeker.
2. Knowledge seeker delar sedan upp förklaringsvariabeln i tio olika intervall och ger ett procentvärde av hur stor sannolikheten för varje intervall är att de ska köpa den givna försäkringen.
3. Sist lägger man ihop de grupper som har liknande procentvärden och kvar får man ett färre antal grupper. Mellan de kvarvarande grupperna är det en märkbar skillnad mellan procentsatserna.

Ett exempel:
Nytegn_hus, som står för nytecknande av husförsäkring och ålder förs in i knowledge seeker. Knowledge seeker delar in ålder i tio olika grupp som ej behöver vara lika stora, ex 10-20, 20-25, 25-38 osv. Var je intervall innehåller ett antal observationer/personer (ej

samma antal i varje intervall) och för varje intervall redovisas ett medelvärde för sannolikheten att köpa en husförsäkring.
Sedan slås intervall med ungefär samma sannolikheter ihop så att man tex får tre grupper kvar som innehåller olika intervall, tex grupp 1 innehåller åldersintervallen 10-20, 30-35, 78-90 och har sannolikheten 4% att köpa försäkring. Mellan dessa grupper ska det vara en märkbar skillnad mellan sannolikheten att köpa en husförsäkring.

Det är viktigt att observera att hur man slår ihop grupper efter liknande sannolikheter kan påverka den slutliga modellen. Ingen gruppindelning behöver vara fel men det kan resultera i olika bra modeller.

6.3 Gruppering av förklaringsvariabler i SAS

För de statistiska modellerna, logistisk regression och bootstrap, används dataprogrammet SAS. Därmed tar vi här upp hur grupperingarna av förklaringsvariablerna ser ut i SAS.

Alla gruppindelningar av förklaringsvariablerna, som har tagits fram med Knowledge seeker, tilldelas var sitt nummer i SAS. Gruppen med lägst sannolikhet att köpa försäkring får högst nummer och den som har högst sannolikhet får lägst nummer. Om en variabel är gruppindelad i fyra grupper så är grupp 4 referenscellen. Referenscellen är den grupp som har lägst sannolikhet att köpa försäkring. Grupp 1 är den grupp med högst sannolikhet att köpa försäkring.

Efter ovanstående gruppering införs dummyvariabler, som kodas på följande sätt. Om t ex förklaringsvariabeln ålder består av fyra grupper, så kodas grupp 1 till age och grupp 2 till age1 och grupp tre till age2. Grupp 4 kodas inte, eftersom den är en referenscell.

Fördelen med ovanstående gruppering i SAS är att är att alla grupper (utom den med lägsta sannolikheten) har högre sannolikhet att teckna försäkring än referenscellen.

För varje förklaringsvariabel vill man alltså få fram vilka personer, som är mest benägna att köpa försäkringar och inte vilka försäkringsbolaget ska undvika att ringa.

7 Metoder och modeller

7.1 Binomialfördelning

Modellen som används för att undersöka köpviljan är logistisk regression. En logistisk regression innebär att responsvariabeln Y är binär och att sannolikheten att få ett lyckat försök följer en binomialfördelning.

Med en binär fördelad variabel menas en variabel som antar två värden. I detta fall är dessa två värden ja (köper försäkring) eller nej (köper ej försäkring) och kodas 1=ja och 0=nej. Binomialfördelningen ger sannolikheten för att få *n* antal personer som köper försäkring av *N* möjliga försök. Sannolikhetsfördelningen för en binomialfördelningen ser allmänt ut på följande vis:

$$p_X(n) = \binom{N}{n} p^n (1-p)^{N-n}$$

där

N - antalet observationer, *n* - antalet gånger som värdet av binär fördelad variabel ska bli 1, $p_x(n)$ - sannolikheten att få *n* stycken 1:or, *p* - sannolikheten att få en 1:a.

7.2 Logistisk regression

Låt:

1. $N_i \geq 1$ vara antalet observationer i grupp *i*. (*i* står i detta fall för vilken försäkring det är);

2. n_i - antalet observationer med värdet $Y_i = 1$, $N_i - n_i$ antalet observationer med värdet $Y_i = 0$;

3. $p(x_i)$ sannolikheten att $Y_i = 1, \quad 0 < p(x_i) < 1,$ för en individ med värde x_i på den förklarande variabeln x $(x = x_i)$.

Då är definitionen för multipel logistisk regression att:
$n_i, i = 1, 2, \ldots, k,$ är oberoende stokastiska variabler och

$$n_i \sim Bin(N_i, p(x_i)), \quad i = 1, 2, \ldots, k$$

$$\log\left(\frac{p(x_i)}{1-p(x_i)}\right) = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots b_r x_{ri} \quad i = 1, 2, \ldots, k$$

där b_0 en konstant och b_1 till b_r är parametrarna till respektive förklaringsvariabel, *x*:en står för varsin förklaringsvariabel som är *r* stycken.

Löser man ut $p(x_i)$ får man:

$$p(x_i) = \exp(b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_r x_{ri}) / (1 + \exp(b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_r x_{ri}))$$

som är sannolikheten att $Y_i = 1$ ska inträffa.

Alla uträkningar har gjorts i dataprogrammet SAS. Den logistiska regressionen har utförts med procedurerna proc logistic och proc genmod.

Proc logistic tar fram alla ß-skattningar. Den visar också vilka variabler som är signifikanta och vilka som inte är det. I detta arbete används signifikansnivån 5 %, vilket innebär att förklaringsvariabler som har ett *p*-värde under 0,05 är signifikanta och förklaringsvariabler som har *p*-värde över 0,05 är inte signifikanta. Proc logistic visar också hur bra modellens utfall stämmer överens med de observerade utfallen. Detta mäts med diskrepansen (eng. ”deviance”), och ett tillhörande p-värde (se kap.8. Resultat). Diskrepansen mäter hur bra en

given modell kan anpassas till det verkliga utfallen. Till hjälp har man en mättad modell, även kallad grundmodell som består av samtliga förklaringsvariabler. Med hjälp av grundmodellen kan man testa hur bra en specifik modell anpassas till det verkliga utfallen. Om modellen anpassas bra, så bör diskrepansen vara ungefär lika med antalet frihetsgrader, dvs diskrepansen/(antalet frihetsgrader) ska ligga nära 1.

I fallet logistisk regression är diskrepansen definierad som:

$$D = 2\left(l(\hat{p}^{(1)}) - l(\hat{p}^{(0)})\right)$$

där

$l(\hat{p}^{(1)})$ - log likelihood för grundmodell där $\hat{p}^{(1)}$ är maximum likelihood-skattning av p under grundmodellen,

$l(\hat{p}^{(0)})$ - log-likelihood för aktuell modell där $\hat{p}^{(0)}$ är maximum likelihood-skattningen av p under aktuell modell.

Så blir diskrepansen D:

$$D=2((\sum_{i=1}^{k} n_i \log\left(\hat{p}^{(1)}(x_i)/(1-\hat{p}^{(1)}(x_i)\right)+\sum_{i=1}^{k} N_i \log(1-\hat{p}^{(1)}(x_i)) -$$
$$\sum_{i=1}^{k} n_i^{(0)} \log\left(\hat{p}(x_i)^{(0)}/(1-p(\hat{x}_i)^{(0)}\right)+\sum_{i=1}^{k} N_i^{(0)} \log(1-\hat{p}(x_i)^{(0)}) =$$
$$= 2\sum_{i=1}^{k} n_i \log\left(\hat{p}^{(1)}(x_i)/\hat{p}^{(0)}(x_i)\right)+2\sum_{i=1}^{k}(N_i-n_i)\log\left((1-\hat{p}^{(1)}(x_i))/(1-\hat{p}^{(0)}(x_i))\right)$$

Proc genmod ger i stort sätt samma information som proc logistic. Men med proc genmod-proceduren kan man även se vilka variabel-kombinationer som ger den högsta sannolikheten uttryckt i procent att kunderna köper en försäkring.

För att få fram vilka förklaringsvariabler som är signifikanta och inte signifikanta används både backward elimination och stepwise regression.

7.3 Backward elimination

Denna metod går ut på att ta med alla förklaringsvariabler i modellen och sedan plocka bort den förklaringsvariabel som är minst signifikant. Om en förklaringsvariabel är signifikant eller inte avgörs med hjälp av *chi-2*-värdet och *p*-värdet. Den förklaringsvariabel som har lägst *chi-2*-värde och högst *p*-värde tas först bort ur modellen och på så sätt får man en ny modell som testas. Så upprepas metoden tills alla icke-signifikanta förklaringsvariabler har plockats bort och kvar är bara signifikanta. Avgörande gräns är 0,05, dvs p=0,05 är lika med signifikant annars icke signifikant.

7.4 Stepwise regression

Den här metoden anses vara den mest användbara. Den börjar utan förklaringsvariabler och sedan plockas den mest signifikanta förklaringsvariabeln in i modellen. Om en förklaringsvariabel plockas in och en annan förklaringsvariabel som redan är med i modellen blir icke signifikant plockas den icke signifikanta förklaringsvariabeln bort.

7.5 Bootstrap

Bootstrap är en statistisk metod som infördes av Bradley Efron (1979) och fick sitt namn från en historia om en baron von Münchhausen som drog upp sig själv ur ett kärr genom att dra i sina stövelstroppar.

Tanken med bootstrap är att man ska skapa ett nytt urval med observationer från det ursprungliga urvalet. Det ursprungliga urvalet/stickprovet antas vara ett slumpmässigt urval från en större population, där varje observation antas vara oberoende och lika fördelad.

Det finns olika bootstrap-metoder men i detta fall skapar vi nya urval med återläggning. Det går till på följande sätt:

1. Man har ett stickprov, vilket i vårt fall är de ursprungliga observationerna som ingår i datamatrialet. Detta stickprov är ett urval från en större grupp med okänd sannolikhetsfördelning, nämligen alla kunder med bil- och/eller husförsäkring hos försäkringsbolaget.

2. Från detta stickprov drar man ett nytt stickprov med återläggning, dvs en observation kan återkomma flera gånger medan en annan kanske inte förekommer alls. Det nya stickprovet innehåller lika många observationer som det ursprungliga stickprovet. (Det behöver nödvändigtvis inte innehålla lika många observationer som det ursprungliga, men med lika många observationer får man det bästa resultatet.)

3. Steg 2 upprepas b antal gånger, så att man får b olika stickprov.

På varje nytt bootstrap-stickprov kan man sedan göra statistiska beräkningar. I detta examensarbete görs logistisk regression. Det skapas även konfidensintervall för varje parameter för att se om bootstrap-resultaten ger andra signifikanta variabler än resultatet från det ursprungliga stickprovet.

En fördel med bootstrap är att man inte behöver känna till någon fördelning. I vanliga fall när man vill analysera parameterskattningar i en regressionsmodell bildar man konfidensintervall och gör tester som bygger på antaganden om en normalfördelning. Det är dock inte alla gånger man kan anta en normalfördelning. Då kan man istället använda bootstrap som inte förutsätter någon speciell fördelning.

Tanken med bootstrap är att om man t ex ska skatta fram ß i en regressionsmodell, så gäller att $\hat{b}*$, som man skattar fram från bootstrap-stickproven, minus $\hat{b}$, som skattas fram från det ursprungliga stickprovet, ska vara lika med $\hat{b}$ minus b, där b är den skattning vi söker, dvs:

$$\hat{b}* - \hat{b} = \hat{b} - b$$

För att förstå hur detta fungerar följer här ett exempel där man skapat konfidensintervall för ett ß-värde i en regressionsmodell.

Låt

$y_i = bx_i + e_i$

vara den modell vi vill skatta fram.

Antag att (y_i, x_i) är oberoende och lika fördelade.

Dra (y_i*,x_i*) (* symboliserar bootstrap) med återläggning *n* gånger och skatta fram $\hat{b}$ * bootstrap-urvalet och upprepa sedan processen *b* antal gånger.
Välj sedan en konfidensgrad (1-a) och sök därefter reda på värdena för a* b* i bootstrap-fördelningen så att:

P(a*< $\hat{b}$ *- $\hat{b}$ <b*)=1- a

Vilket är samma sak som;

P(a*< $\hat{b}$ - b <b*)=1-a

Löser vi ut ß får vi:

$\hat{b}$ -b*< b < $\hat{b}$ -a*

Här ser via att vi använt relationen:
$\hat{b} * - \hat{b} = \hat{b} - b$.

För ett 95 % konfidensintervall får man fram b* och a* på följande sätt:

Från *b* olika bootstrap-stickprov får man *b* olika $\hat{b}$ * värden. Dessa sorterar man sedan i storleksordning och använder de $\hat{b}$ *som hamnar på den 2,5:e percentilen respektive 97,5:e percentilen.

Då blir;
b*=97,5:e percentilvärdet - $\hat{b}$
a*=2,5:e percentilvärdet - $\hat{b}$

8 Resultat

I alla resultat utan bootstrap kommer ß-skattningar att vara positiva. Det beror på, som tidigare nämnts, att referenscellen är den grupp med minst sannolikhet att köpa försäkring. Gruppindelningen är gjord efter sannolikheten att teckna försäkring (se gruppindelning), där den grupp med lägst sannolikhet fått högst nummer och den med högst sannolikhet fått lägst nummer. Det innebär teoretiskt att alla grupper med låga nummer borde bli mer signifikanta än grupper med högre nummer, vilket man också ser att det blir, se tabellerna nedan. Det är dock inte alltid självklart att det blir så, vilket kan bero på att i vissa grupper är antalet observationer så få att de inte blir signifikanta. Men det går vi inte in på här.

8.1 Bilförsäkring

8.1.1 Resultat utan bootstrap

I bilförsäkringsmodellen ingår följande variabler:

age = åldersgrupper (11-37 och 61-69)
arbetar = sannolikheten att man arbetar (54,5-57,78 och 59,66-62,17 %)
husa = antal år för husförsäkringsinnehav (1-2 och 6-7)
child = sannolikheten att man har barn (22,54-30,19 och 33,8-41,18 %)

I tabell 2 nedan ser vi p-värdet för varje variabel. Här ser vi att alla variabler är signifikanta på 5 % nivån.

Vi har också ett utdrag som visar diskrepans-värdet och *p*-värdet, se tabell 1. Från nu kallar vi diskrepans/(antal frihetsgrader) som diskrepans-värde.

Ett diskrepans-värde på 0,7551 och ett *p*-värde på 0,6857 säger att modellen passar bra till det faktiska utfallet. Som nämnts tidigare ska både diskrepans-värdet och p-värdet ligga så nära värdet 1 som möjligt för att modellen ska vara så perfekt som möjligt. Alltså kan nedanstående värden anses vara bra.

Tabell 1

Kriterium	Värde	DF	Värde/DF	Sh. > Chi2
Diskrepans	8,3061	11	0,7551	0,6857
Pearson	8,4655	11	0,7696	0,6711

Tabell 2

Parameter	DF	Skattning	Standard-fel	Wald Chi2	Sh. > Chi2
intercept	1	-4,4689	0,2668	280,5305	<,0001
age	1	0,6721	0,2561	6,8861	0,0087
arbetar	1	0,5382	0,2557	4,4288	0,0353
husa	1	0,5366	0,2631	4,1603	0,0414
child	1	0,5609	0,2540	4,8773	0,0272

Tittar vi sedan på variabelkombinationerna visar det sig med proceduren proc genmod att för de personer, som faller under kombinationen age, arbetar, husa och child, så är det 10 % chans att de ska köpa en bilförsäkring. Denna variabelkombination ger den största sannolikheten att teckna bilförsäkring. Den näst största sannolikheten är kombinationen age, arbetar, child, (ej husa, dvs husa=0). Denna kombination ger 6,3 % chans att de ska köpa bilförsäkring.

8.1.2 Resultat med bootstrap

Vi testar samma modell som vi testade ovan men med bootstrap, dvs en modell med variablerna age, arbetar, husa och child.

Bootstrap-resultatet i tabell 3 nedan visar att endast husa och age är signifikanta. Tittar man på konfidensintervallen med bootstrap så ser man att för arbetar och för child, som inte är signifikanta på 5 % nivån, ligger den undre gränsen runt -0,03 respektive -0,07. Ett värde på -0,03 ligger så pass nära värdet 0 att vi väljer att ha kvar den i modellen, men däremot tar vi bort variabeln child.

Att child som hade ett *p*-värde på 0,0272, se tabell 2, och var den variabeln som var näst mest signifikant och nu inte blir signifikant längre kan tyckas vara märkligt. Förklaringen till detta kan vara att child, som var signifikant tidigare ändå inte påverkade köpviljan särskilt mycket. Detta märks när vi tar bort child ur modellen. Av de variabler, som ingår i

modellen, är child nämligen den variabel som minst påverkar diskrepans-värdet och *p*-värdet. Därför känns det naturligt att ta bort child ur modellen.

Däremot vill vi testa om variabeln child samspelar med de övriga variablerna i modellen. Det visar sig att variabeln child samspelar med age. Däremot blir modellen med samspel inte lika bra när vi tittar på diskrepans-värdet och *p*-värdet, så vi lämnar denna modell utan att visa några resultat eller utskrifter. De övriga variablerna i modellen samspelar inte.

Tabell 3

	95 % konfidensinter-vall med bootstrap	95 % konfidensinter-vall utan bootstrap
intercept	-5,1071 -3,5351	-4,9918 -3,9460
age	0,1672 1,3616	0,1701 1,1741
arbetar	-0,0326 1,0904	0,0370 1,0394
husa	0,0960 1,0813	0,0209 1,0523
child	-0,0762 1,0092	0,0631 1,0587

8.1.3 ”Bootstrap-modell” på verkliga observationer

Vi testar nu ”bootstrap-modellen”, dvs den modell som vi fick fram med bootstrap, med variablerna age, arbetar och husa och kör den på de verkliga observationerna. Man får då följande resultat från SAS. Här ser vi i tabellerna 4 och 5 att modellen är bra, eftersom diskrepans-värdet och *p*-värdet ligger på 0,6108 respektive 0,6548. Vi ser också att alla variabler i modellen blir signifikanta, vilket vi också ville att de skulle bli.

Tabell 4

Kriterium	Värde	DF	Värde/DF	Sh. > Chi2
Diskrepans	2,4434	4	0,6108	0,6548
Pearson	2,2748	4	0,5687	0,6854

Tabell 5

Parameter	DF	Skattning	Standard-fel	Wald Chi2	Sh. > Chi2
intercept	1	-4,1998	0,2279	339,4803	<,0001
age	1	0,6674	0,2559	6,8047	0,0091
arbetar	1	0,5332	0,2554	4,3573	0,0368
husa	1	0,5543	0,2626	4,4561	0,0348

Tittar man sedan på variabelkombinationerna visar det sig med proceduren proc genmod att för de personer, som faller under kombinationen age, arbetar och husa, så är det 8 % chans att de ska köpa en bilförsäkring.

8.2 Husförsäkring

8.2.1 Resultat utan bootstrap

I husförsäkringsmodellen ingår följande variabler:

post = postnummer (2740-2980 och 4700-8200)
child = sannolikheten att man har barn (41,18-47,06 %)
child1 = sannolikheten att man har barn (11,48-41,18 och 47,06-87,1 %)
arbetar = sannolikheten att man arbetar (46,03-50,13 och 54,5-56 %)
lon1 = sannolikheten att man har en lön på max 400 063 kronor per år (74,76-82 %)
car = sannolikheten att man har bil (74,76-82 %)

De signifikanta variablerna är post, child, arbetar, lon1 och car, se tabell 7. Variabeln child 1 är inte signifikant men tas med i modellen för att den påverkat variabeln child. Utan child1 blir nämligen child icke signifikant och därför behövs child1 med i modellen.

I tabell 6 nedan ser vi också att diskrepans-värdet är 0,7194 och *p*-värdet är 0,8968 vilka båda ligger nära värdet 1 och därför anses modellen stämma bra överens med de riktiga utfallet.

Tabell 6

Kriterium	Värde	DF	Värde/DF	Sh. > Chi2
Diskrepans	26,6168	37	0,7194	0,8968
Pearson	22,6638	37	0,6125	0,9692

Tabell 7

Parameter	DF	Skattning	Standard-fel	Wald Chi2	Sh. > Chi2
intercept	1	-6,4212	0,7828	67,2928	<,0001
post	1	0,7190	0,2905	6,1273	0,0133
child	1	1,6572	0,8302	3,9846	0,0459
child1	1	1,2170	0,7426	2,6854	0,1013
arbetar	1	0,8347	0,2990	7,7942	0,0052
lon1	1	1,3561	0,3736	13,1734	0,0003
car	1	0,7448	0,3164	5,420	0,0186

En variabelkombination av post, child1, arbetar, lon1 och car ger den högsta sannolikheten att köpa en husförsäkring 17,5 %. Det existerar ingen variabelkombination av post, child, arbetar, lon1 och car, vilket egentligen bör ha den högsta sannolikheten enligt ovanstående modell, eftersom child påverkar köpviljan mer positivt än child1.

8.2.2 Resultat med bootstrap

Samma modell testas med bootstrap. I tabell 8 nedan ser man konfidensintervallen både med och utan bootstrap.

De signifikanta variablerna med bootstrap är post, arbetar, lon1 och car, dvs där värdet 0 ej ingår i konfidensintervallen.

De övriga variablerna child och child1 är långt ifrån signifikanta, eftersom värdet 0 ingår i konfidensintervallen och varken den undre eller den övre gränsen i konfidensintervallen ligger nära värdet 0. Vi tar bort dessa två variabler ur modellen och testar den nya modellen på de verkliga observationerna.

Tabell 8

	95 % konfidensinter-vall med bootstrap	95 % konfidensinter-vall utan bootstrap
intercept	-7,6487 -4,6089	-7,9555 -4,8869
post	0,2493 1,3221	0,1496 1,2884
child	-0,7261 2,7721	0,0300 3,2844
child1	-1,7965 2,4273	-0,2385 2,6725
arbetar	0,1373 1,5506	0,2487 1,4207
lon1	0,7350 2,2349	0,6238 2,0884
car	0,0162 1,3897	0,1247 1,3649

8.2.3 ”Bootstrap-modell” på verkliga observationer

I modellen, som vi fått fram genom bootstrap, ingår följande variabler:

post = postnummer (2740-2980 och 4700-8200)
arbetar = sannolikheten att man arbetar (46,03-50,13 och 54,5-56 %)
lon1 = sannolikheten att man har en lön på max 400 063 kronor per år (74,76-82 %)
car = sannolikheten att man har bil (74,76-82 %)

I tabell 10 ser vi att alla variabler är signifikanta.

I tabell 9 ser vi att diskrepans-värdet är 0,9043 och *p*-värdet 0,5351. Diskrepans-värdet är mycket bra och *p*-värdet skulle kunna vara lite högre men är ändå bra.

Tabell 9

Kriterium	Värde	DF	Värde/DF	Sh. > Chi2
Diskrepans	9,9475	11	0,9043	0,5351
Pearson	8,3215	11	0,7565	0,6842

Tabell 10

Parameter	DF	Skattning	Standard-fel	Wald Chi2	Sh. > Chi2
intercept	1	-5,2500	0,3321	249,9501	<,0001
post	1	0,7278	0,2902	6,2896	0,0121
arbetar	1	0,8305	0,2987	7,7281	0,0054
lon1	1	1,1583	0,3640	10,1238	0,0015
car	1	0,8140	0,3141	6,7134	0,0096

Slutligen tittar vi på variabelkombinationen för denna modell. Här är den högsta sannolikheten att köpa försäkring 15,2 %, vilket är en kombination av variablerna post, arbetar, lon1 och car.

8.3 Familjeförsäkring

8.3.1 Resultat utan bootstrap

I familjeförsäkringsmodellen ingår följande variabler:

post = postnummer (1068-2500)
post1 = postnummer (2500-2980)
post2 = postnummer (6760-8200 och 4700-5620)
arbetar = sannolikheten att man arbetar (4,74-46,03 och 54,5-56 %)
arbetar1 = sannolikheten att man arbetar (46,03-50,13 och 52,57-54,5 %)
arbetar2 = sannolikheten att man arbetar (50,13-52,57, 56-57,78 och 59,66-75,79 %)
aldst = sannolikheten att den äldsta i hushållet är 50-64 år (14,29-19,05 %)
aldst1 = sannolikheten att den äldsta i hushållet är 50-64 år (22,86-26,09 %)

I tabell 12 nedan ser vi vilka variabler som är signifikanta och vilka som inte är det. Av tabellen framgår att aldst1, arbetar2 och post2 inte är signifikanta. Anledningen till att dessa

variabler är med i modellen är att man misstänker att de faktiskt kan påverka köpviljan av familjeförsäkring. Detta beror på att *p*-värdena för dessa variabler ligger runt 0,07, vilket är nära 0,05, men också på att dessa variabler påverkar diskrepans-värdet och *p*-värdet ordentligt. Utan dessa variabler sjunker både diskrepans-värdet och *p*-värdet väldigt mycket och därför vill man testa om dessa variabler blir signifikanta när man använder bootstrap.

När det gäller variabeln arbetar2 så påverkar den även arbetar1 så pass mycket att arbetar1 inte blir signifikant om arbetar2 tas bort. Det syns tydligt i tabell 12 att arbetar1 är signifikant och därför måste arbetar2 vara kvar i modellen.

Vi tittar också på diskrepans-värdet och tillhörande p-värde, se tabell 11. Diskrepans-värdet och *p*-värdet blir 0,9587 respektive 0,5435. Diskrepans–värdet är mycket bra och *p*-värdet bra. *P*-värdet skulle kunna vara lite högre för att bli riktigt bra men anses ändå bra.

Tabell 11

Kriterium	Värde	DF	Värde/DF	Sh. > Chi2
Diskrepans	37,3881	39	0,9587	0,5435
Pearson	38,9569	39	0,9989	0,4718

Tabell 12

Parameter	DF	Skattning	Standard-fel	Wald Chi2	Sh. > Chi2
intercept	1	-4,0881	0,2956	191,2185	<,0001
post	1	0,8027	0,2021	15,7801	<,0001
post1	1	0,5550	0,1668	11,0641	0,0009
post2	1	0,3208	0,1786	3,2253	0,0725
arbetar	1	0,9932	0,3015	10,8543	0,0010
arbetar1	1	0,7199	0,3190	5,0944	0,0240
arbetar2	1	0,5478	0,3025	3,2782	0,0702
aldst	1	0,4412	0,1870	5,5648	0,0183
aldst1	1	0,3563	0,2020	3,1101	0,0778

Vid variabelkombinationerna så är den högsta sannolikheten att köpa en familjeförsäkring 13,6 %, vilket är variabelkombination av post, arbetar och aldst.

8.3.2 Resultat med bootstrap

Tabell 13 visar konfidensintervall med och utan bootstrap.

De signifikanta variablerna med bootstrap är post, post1, arbetar, aldst och aldst1. Det har alltså skett en förändring mellan med och utan bootstrap, när vi testar samma modell. Nu kan man också se att aldst, som inte var signifikant utan bootstrap, har blivit signifikant.

Däremot ser vi att arbetar1 och arbetar2 inte blir signifikanta. Variabeln post2 är på gränsen till signifikant så den får vara kvar i modellen, trots att den lutade över till att vara icke-signifikant. Vi tar bort variablerna arbetar 1 och arbetar2 ur modellen och testar den nya modellen på de verkliga observationerna.

Tabell 13

	95 % konfidensinter-vall med bootstrap	95 % konfidensinter-vall utan bootstrap
intercept	-4,6436 -3,3613	-4,6675 -3,5087
post	0,4844 1,2408	0,4066 1,1988
post1	0,2596 0,8798	0,2281 0,8819
post2	-0,0285 0,6740	-0,0293 0,6709
arbetar	0,1356 1,5557	0,4023 1,5841
arbetar1	-0,1515 1,4305	0,0947 1,3451
arbetar2	-0,2207 1,1588	-0,0451 1,1407
aldst	0,1070 0,8221	0,0747 0,8077
aldst1	0,0361 0,7013	-0,0396 0,7522

8.3.3 ”Bootstrap-modell” på verkliga observationer

I modellen, som vi fått fram genom bootstrap, ingår följande variabler:

post = postnummer (1068-2500)
post1 = postnummer (2500-2980)
post2 = postnummer (6760-8200 och 4700-5620)
arbetar = sannolikheten att man arbetar (4,74-46,03 och 54,5-56 %)
aldst = sannolikheten att den äldsta i hushållet är 50-64 år (14,29-19,05 %)
aldst1 = sannolikheten att den äldsta i hushållet är 50-64 år (22,86-26,09 %)

I tabell 15 ser vi att post2 är precis på gränsen till signifikans och även aldst1 är nära signifikansgränsen 0,05.

I tabell 14 ser i att modellen är bra. Diskrepans-värdet är 0,8240 och *p*-värdet är 0,6665.

Tabell 14

Kriterium	Värde	DF	Värde/DF	Sh. > Chi2
Diskrepans	14,0088	17	0,8240	0,6665
Pearson	13,9150	17	0,8185	0,6731

Tabell 15

Parameter	DF	Skattning	Standard-fel	Wald Chi2	Sh. > Chi2
intercept	1	-3,5525	0,1227	838,7346	<,0001
post	1	0,7619	0,2005	14,4377	0,0001
post1	1	0,5613	0,1667	11,3333	0,0008
post2	1	0,3443	0,1783	3,7270	0,0535
arbetar	1	0,4523	0,1344	11,3174	0,0008
aldst	1	0,4475	0,1868	5,7421	0,0166
aldst1	1	0,3737	0,2017	3,4331	0,0639

Eftersom vi bara tagit bort variablerna arbetar1 och arbetar2 så blir fortfarande den bästa variabelkombinationen post, arbetar och aldst, som har en sannolikhet på 13,6 % att kunderna ska köpa en familjeförsäkring.

8.4 Olycksförsäkring

8.4.1 Resultat utan bootstrap

I olycksförsäkringsmodellen ingår följande variabler:

post = postnummer (4700-5620 och 6760-8200)
age = ålder (11-52 och 58-69)
arbetar = sannolikheten att man arbetar (39,49-46,03 %)
arbetar1 = sannolikheten att man arbetar (54,5-56 %)

child = sannolikheten att man har barn (22,54-26,67 och 37,5-41,18 %)
aldst = sannolikhet att den äldsta i hushållet är 50-64 år (14,29-30 %)
lon1 = sannolikheten att man har en lön på max 400 063 kronor per år (2,91-35,51, 42,45-66,95 och 74,76-82 %)
husa= hur länge man har ägt en husförsäkring mätt i år (1-3, 6-7)

I tabell 17 nedan ser vi att alla variablerna i modellen är signifikanta. Diskrepans-värdet och *p*-värdet är 0,7686 respektive 0,9938, se tabell 16, så modellen passar bra till det riktiga utfallet.

Tabell 16

Kriterium	Värde	DF	Värde/DF	Sh. > Chi2
Diskrepans	155,2612	202	0,7686	0,9938
Pearson	246,3165	202	1,2194	0,0181

Tabell 17

Parameter	DF	Skattning	Standard-fel	Wald Chi2	Sh. > Chi2
intercept	1	6,4081	0,4382	213,8508	<,0001
post	1	0,5050	0,2300	4,8199	0,0281
age	1	0,9450	0,3119	9,1782	0,0024
arbetar	1	1,5274	0,3077	24,6388	<,0001
arbetar1	1	1,0677	0,3404	9,8380	0,0017
arbetar2	1	0,5742	0,2724	4,4448	0,0350
child	1	0,5429	0,2273	5,7046	0,0169
aldst	1	0,5372	0,2096	6,5706	0,0104
lon1	1	0,6375	0,2544	6,2790	0,0122
husa	1	0,6603	0,2816	5,4992	0,0190

Vid betraktande av variabelkombinationerna så är den högsta sannolikheten att köpa en olycksförsäkring 26 %. Det är en variabelkombination av post, age, arbetar, child, aldst,

lon1 och husa. Det är dock endast en person, som faller in under alla dessa variabler, så därför redovisas fler variabelkombinationer:

17,4 % för variabelkombination av age, child, aldst och lon1
16,9 % för variabelkombination av post, age, aldst och lon1

8.4.2 Resultat med bootstrap

Vid körning av samma modell som ovan på bootstrap-stickproven så blir även här alla variabler signifikanta. I tabell 18 nedan ser man konfidensintervallen med och utan bootstrap. Inte i något intervall ingår värdet 0, så alla variabler blir signifikanta.

Tabell 18

	95 % konfidensinter-vall med bootstrap	95 % konfidensinter-vall utan bootstrap
intercept	-7,2060 -5,5841	-7,2670 -5,5492
post	0,0478 0,9829	0,0542 0,9558
age	0,3710 1,4651	0,3337 1,5563
arbetar	1,0764 2,1376	0,9243 2,1305
arbetar1	0,4885 2,6626	0,4005 1,7349
arbetar2	0,2910 1,2677	0,0403 1,1081
child	0,0294 1,2316	0,0974 0,9884
aldst	0,1829 0,9589	0,1264 0,9480
lon1	0,0178 1,1515	0,1389 1,1361
husa	0,0725 1,4594	0,1084 1,2122

8.4.3 ”Bootstrap-modell” på verkliga observationer

Tabellerna 19 och 20 visar exakt samma värden som tabellerna 16 och 17, dvs resultat utan bootstrap. Detta beror på att alla variabler blev signifikanta både med och utan bootstrap, så vi får alltså samma modeller.

Tabell 19

Kriterium	Värde	DF	Värde/DF	Sh. > Chi2
Diskrepans	155,2612	202	0,7686	0,9938
Pearson	246,3165	202	1,2194	0,0181

Tabell 20

Parameter	DF	Skattning	Standard-fel	Wald Chi2	Sh. > Chi2
intercept	1	6,4081	0,4382	213,8508	<,0001
post	1	0,5050	0,2300	4,8199	0,0281
age	1	0,9450	0,3119	9,1782	0,0024
arbetar	1	1,5274	0,3077	24,6388	<,0001
arbetar1	1	1,0677	0,3404	9,8380	0,0017
arbetar2	1	0,5742	0,2724	4,4448	0,0350
child	1	0,5429	0,2273	5,7046	0,0169
aldst	1	0,5372	0,2096	6,5706	0,0104
lon1	1	0,6375	0,2544	6,2790	0,0122
husa	1	0,6603	0,2816	5,4992	0,0190

Variabelkombinationerna ger här samma resultat som under resultatet utan bootstrap.

9 Slutsats

Av resultatet i föregående avsnitt kan man dra följande slutsatser:

- att de variabler, som var tveksamma att ta med i modellerna från början, visade sig inte bli signifikanta vid bootstrap

- att de variabler, som inte var signifikanta men låg på gränsen till signifikans och som påverkade diskrepans-värdet och *p*-värdet mycket, blev signifikanta eller att de närmade sig signifikansnivån mer än tidigare

- att vi får enklare modeller, som visar ett bra resultat, när vi testar bootstrap-modeller på det riktiga datamaterialet. Det goda resultat framgår av respektive variabels p-värde, som förändrades i förväntad riktning. Det goda resultatet framgår också av diskrepans-värdet och tillhörande *p*-värde.

- att när man tittar på variabelkombinationerna av de mest signifikanta variablerna med och utan bootstrap, så skiljer inte sannolikheten att man ska köpa en försäkring särskilt mycket. Inte heller *p*-värdet och diskrepans-värdet skilde så otroligt mycket vilket gör att bootstrap-metoden känns trovärdig och bra

Ovanstående fyra punkter ger oss en positiv bild av bootstrap, dvs skattning av fler dataobservationer med hjälp av bootstrap verkar ge trovärdiga modeller.

En mycket stor fördel med bootstrap är att det är en billig och snabb metod i jämförelse med att skaffa fram verkliga observationer.

Kommentarer till signifikanta variabler m m

När man tittar på vilka variabler som är signifikanta, så ser man att postnummer, arbetar och ålder ofta är det. Man kan förvänta sig att dessa tre variabler är signifikanta.

Postnummer är knutet till ett visst bostadsområde och i olika bostadsområden finns olika typer av människor representerade. Brottsligheten varierar, vilket påverkar prissättningen av försäkringar, vilket i sin tur påverkar köpviljan.

Det känns också naturligt att variabeln ålder påverkar köpviljan. Man lever olika slags liv i olika åldrar och är på så sätt i behov av olika saker och därför även av olika försäkringar.

Det som kanske däremot är förvånande är att sannolikheten att ha bil inte påverkar köp av bilförsäkring. Den togs inte med i modellen överhuvudtaget, eftersom *p*-värdet låg runt 0,07 utan bootstrap och den blev inte heller signifikant vid bootstrap. Den påverkade inte heller diskrepans-värdet eller p-värdet så mycket, vilket är lite förvånande.

Nu får man bara hoppas att bootstrap-modellerna och metoden bootstrap fungerar lika bra i praktiken. I så fall är bootstrap lika fantastisk som baronen själv var, när han drog upp sig själv ur kärret genom att dra i sina stövelstroppar.

Referenser

Andersen Aage T, Feilberg Michael, Jakobsen René B, Milhøj Anders (2006): *Statistik med SAS*, Statistikforlaget

Efron Bradley (1982): *The Jackknife, the Bootstrap, and Other Resampling Plans*, CBMS-NSF Regional Conference Series in Applied Mathematics

Efron Bradley and Tibshiran Robert J (1994): *An Introduction to the Bootstrap*, Chapman & Hall

Garthwaite Paul H, Jolliffe Ian T and Jones Byron (2002): *Statistical Inference*, second edition, Oxford Science Publications

Hand D J (1996): *Statistics and Computing*, Volume six number two, Chapman & Hall

Johansson Björn, (2005): *Matematiska modeller inom sakförsäkring*, Kompendium, Matematisk statistik, Stockholms universitet

Ohlsson Esbjörn (2003): *Log-linjära modeller och Logistisk regression*, Kompendium, Matematisk statistik, Stockholms universitet

Ohlsson Esbjörn (2005): *Korthandledning i SAS*, Kompendium, Matematisk statistik, Stockholms universitet

SAS Institute (1999): *SAS OnlineDoc version 8*, http://v8doc.sas.com/sashtml

Weiss Eric: *Bootstrapping With SAS*, http://www.winchendon.com/bootstrap.html