LUND UNIVERSITY
School of Economics and Management

Informationsasymmetri på den svenska bilförsäkringsmarknaden

Gustav Lundström och Nils Shannon

Sammanfattning

Denna uppsats undersöker informationsasymmetri på den svenska bilförsäkringsmarknaden. Genom att använda kunddata från Länsförsäkringar Uppsala analyserar studien korrelationen mellan täckning och risk. Två logitmodeller estimerar effekten av att en ökad grad av försäkring leder till större sannolikhet för skadeanmälan. Tillsammans med dessa estimeringar kontrollerar vi för att olika försäkringstyper inkluderar fler punkter för skadeanmälningar. Samtidigt som denna studie påvisar ett positivt och signifikant ($p \leq 0{,}05$) resultat vid korrelationen mellan täckning och risk, kan inte resultatet tillskrivas informationsasymmetri.

Nyckelord: Moral hazard, Adverse selection, Bilförsäkringsmarknaden

Examensarbete – kandidatnivå
Nationalekonomiska institutionen
Handledare: Jerker Holm
Datum: 26 maj 2021

Tack

Vi vill tacka vår handledare HJ Holm för hans stöd och djupa kunskaper inom ämnet. Sedan vill vi också tacka Ulrica Hedman, Ulrika Götrich, Jenny Wallmark och samtliga från Länsförsäkringar Uppsala.

Innehåll

Tabeller

1 Introduktion

Då olika parter sluter överenskommelser med varandra kan problem gällande asymmetrisk information uppstå (Abbring, Chiappori & Pinquet, 2003). När en avtalspart har mer information jämfört med den andra avtalsparten är informationen relaterad till transaktionen skevt fördelad. Att det föreligger en skev fördelning i information mellan två parter i ett kontrakt står i skarp kontrast till antaganden i neoklassisk teori. Ett huvudantagande inom teorin är just perfekt information (Varian, 2014). Situationer där informationsasymmetri föreligger är därmed intressant att studera.

Gällande försäkringsindustrin förekommer två begrepp kopplade till informationsasymmetri frekvent, nämligen *moral hazard* och *adverse selection*. Denna studie tillämpar begreppen på ett område av den svenska bilförsäkringsmarknaden. *Moral hazard* innebär att en individ förändrar sitt beteende efter att den ingått i ett avtal. *Adverse selection* avser situationer där ena parten har mer information än den andra gällande avtalet ifråga (Akerlof, 1970). Begreppen uppkommer parallellt och deras respektive effekter är ofta svår att skilja från varandra.

Det faktum att informationsasymmetri uppstår konsekvent i många sammanhang medför att forskningen inom området är välstuderat och detta gäller i synnerhet försäkringsmarknader. Datauppgifter från försäkringsbolag utgör tämligen goda underlag för empiriska undersökningar då kontrakten ofta är standardiserade och består av tvärsnittsdata. Stora datamängder finns att tillgå då miljoner av människor i Sverige innehar någon typ av försäkring som justeras årligen. Varje bilägare har en skyldighet att hålla med trafikförsäkring, och en möjlighet att teckna en utökad försäkring, givet att bilen är i bruk. Detta skapar en obligatorisk kontakt mellan bilägare och försäkringsbolag. Då olika individer varierar i vilken grad de vill försäkra sin bil, kan försäkringsbolagen gradera dessa individer med hjälp av olika typer av bilförsäkringar.

Syftet med denna uppsats är att undersöka informationsasymmetri på den svenska bilförsäkringsmarknaden. Den här studien tar sitt teoretiska avstamp genom en litteraturstudie över hur tidigare forskning utvecklat modeller för att undersöka informationsasymmetri. Efter detta görs en empirisk analys av korrelationen

mellan täckning och risk. Täckning definieras som graden av försäkring individen innehar. Risk definieras som sannolikhet för en skadeanmälan. Vi kontrollerar för att olika försäkringar de facto innehåller olika mängd anmälningspunkter genom att dela in analysen av försäkringstyper i två grupper. Dessa grupper jämförs och kontrolleras tillsammans med hur stor andel av samtliga skador de olika anmälningspunkterna utgör på den svenska bilförsäkringsmarknaden. Såvitt vi vet är detta ett unikt angreppsätt. Studien innefattar estimeringar av korrelationen mellan täckning och risk med hjälp av två logit-modeller och tvärsnittsdata från ett regionalt försäkringsbolag. Tillsammans med den procentuella förändringen mellan olika försäkringstyper kontrollerar vi för tidigare nämnda logistiska estimeringar. Med detta redogjort lyder vår övergripande forskningsfråga:

Går det att estimera informationsasymmetri på den svenska bilförsäkringsmarknaden genom korrelationen mellan täckning och risk?

Det finns ett antal begränsningar i denna studie. Vi analyserar data från ett regionalt försäkringsbolag vilket påverkar generaliserbarheten i våra resultat. Detta eftersom forskningsfrågan vi försöker besvara gäller den svenska bilförsäkringsmarknaden. Vidare saknar vi detaljerad data på skadeanmälningarna som utgör grunden i vår analys. Riskprofilen hos individerna är därmed ofullständig.

Denna uppsats är strukturerad på följande vis: I avsnitt 2 presenteras en genomgång av relevant forskning i ämnet. Avsnitt 3 behandlar teori och introducerar modeller som tillsammans med empirin från avsnitt 4 och metoddelen i avsnitt 5 analyseras i resultatdelen i avsnitt 6. Avsnitt 7 är diskussionsdelen vilken avslutas med slutsatser.

2 Forskningsöversikt

Inom kontraktteori har forskning gällande asymmetrisk information blivit allt mer utmanad och analyserad. För att undvika anakronistiska misstag följer en översikt över forskning som försökt estimera informationsasymmetri på försäkringsmarknader. Den tidiga forskningen inom området har bland annat visat att informationsasymmetri i den konkurrensutsatta marknaden leder till ineffektiva resultat

och ett marknadsmisslyckande (Arrow, 1963), (Pauly, 1968, 1978) och (Rothschild & Stiglitz, 1978). En annan infallsvinkel inom samma fält hämtar teori från forskning gällande att högre risk sammankopplas med en allt mer omfattande försäkringstäckning. Slutsatserna från Rothschild & Stiglitz (1978) samt Wilson (1977) ger detta evidens. Kopplingen mellan täckning och risk är den samma som *adverse selection*, där högrisk-individer väljer ett högre försäkringsskydd. Vidare tillkom även forskning kring den andra stora delen av informationsasymmetri, *moral hazard*. Shavell (1979) och Holmström (1979) initierade debatten med sin forskning inom området. Holmström skriver, "... att ursprunget till *moral hazard* och problematiken gällande incitament är att det råder informationsasymmetri mellan individer. Eftersom att individens samtliga handlingar inte går att observera, går det heller inte att kontraktera dem" (Holmström, 1979, s. 150;[vår kursivering]).

Nästa generations forskning var av en annan karaktär då den tog avstamp i vad den tidigare forskningen teoretiskt grundlade för att utveckla modeller för *adverse selection* och *moral hazard*. De första observationsstudierna på bilförsäkringsmarknaden gav visst stöd för informationsasymmetri-hypotesen (Puelz & Snow, 1994), (Dahlby, 1983), (Dahlby, 1992). Dock inte helt utan kontroverser då påföljande studier av Chiappori & Salanié (2000) och Dionne m.fl. (2001) kritiserade Puelz & Snows (1994) artikel. Kritiken grundade sig i att författarna utelämnade variabler och att de inte tog individens ålder och körförmåga i beaktande vid analysen. Till skillnad från de första observationsstudierna använde Chiappori & Salanié (2000) åldersgruppen unga förare för att undkomma problem som uppstår när individen över tid tillförskaffar sig asymmetrisk inlärning. Detta angreppsätt förutsätter att, ju mer erfarenhet en individ har i trafiken, desto mer information har hen om sina egna körvanor och riskbeteende. Försäkringsbolaget saknar denna information vilket skapar informationsasymmetri mellan parterna. Informationen är bristfällig om försäkringsbolagen inte samarbetar. Bolagen delar inte data gällande individers körvanor, samtliga olyckor i trafiken samt kundens historik mellan varandra. Genom dessa antaganden undersöker Dionne m.fl. (2013) *moral hazard* i deras studie. Till skillnad från Chiappori & Salaniés (2000) tidigare forskning kring försäkringsskador använder sig Dionne m.fl. (2013) av paneldata från en treårsperiod. Tillsammans med paneldata analyserar Dionne m.fl. utfallet med ett Granger kau-

salitetstest[1]. Utfallet av testet visar stark empirisk evidens för *moral hazard*. Vid estimering och särskiljning av *moral hazard* och *adverse selection* menar Abbring m.fl. (2003) att det främst handlar den tillgänglig data. Till exempel, handlar premien och i förlängningen försäkringsbolagets riskbedömning om datapunkter gällande föraren och bilen. Försäkringsbolag använder sig av denna tvärsnittsdata för att bestämma priset men frågan om hur det går att empiriskt studera den kausala effekten av informationsasymmetri lämnar författarna obesvarat.

Israel (2004) behandlar också detta problem och konstaterar att villkorlig korrelation med tvärsnittsdata endast är applicerbar vid robust test av informationsasymmetri. Att kunna särskilja på *adverse selection* och *moral hazard* för att sedan mäta effekten kräver analysen longitudinell data. Med tvärsnittsdata saknas en kontrollgrupp till vilken det går att jämföra den behandlade gruppen med. Likt en randomiserad kontrollstudie använder man longitudinell data för att studera effekten av *adverse selection* och *moral hazard* mellan de två grupperna. Från Israels resultat gick det att utvinna en liten men signifikant *moral hazard*-effekt. Skillnaden mellan Abbring m.fl. (2003) och Israel (2004) är att det finns delade uppfattningar mellan studierna gällande ursprunget till den villkorliga korrelationen med tvärsnittsdata. Abbring m.fl. (2003) använder Heckman & Borjas (1980) tillvägagångsätt när de utformar sin modell. De använder en proportionell-modell[2], vilken jämför fördelningen mellan individer som anmäler flertalet skador till försäkringsbolaget. I annan forskning har även Abbring m.fl. (2008) samt Dionne m.fl. (2013) använt sig av paneldata vid analys av *moral hazard*s existens på försäkringsmarknaden. Denna empiriska tillämpning ligger i grunden för Cohen & Siegelmans (2010) tolkning av forskningsområdet. De menar att forskning inom informationsasymmetri på försäkringsmarknader inte skall utgå ifrån att försöka bevisa eller motbevisa huruvida *adverse selection* och korrelationen mellan täckning och risk existerar. Cohen & Siegelman hävdar att det finns en god grund att använda dessa två begrepp som antaganden vid analys av vissa försäkringsmarknader, där försäkringsmarknaden för bilar ingår. Forskningen kring detta ämne är i ett skede

[1]Ett statistiskt hypotestest som avgör om en tidsserie är användbar för att förutsäga en annan tidsserie. För en mer utförlig förklaring se Granger (1969).

[2]Statistisk modell som jämför ett före-efter-scenario vid en isolerad händelse. För vidare läsning se Cox, D.R. (1972).

där det gäller att utröna var *moral hazard* och *adverse selection* förekommer och utforma modeller för att beräkna effekterna av dem.

Ett exempel från den nya skolan av forskning är hämtat från Weisburds (2015) analys av bilförsäkringsmarknaden i Israel. Resultatet av Weisburds forskning var att slutsatser gällande försäkringen var av företags- eller privatförsäkringstyp samt en självkostnadsreducering spelade roll för individens beteende. Från resultatet estimerade Weisburd att ungefär 10 procent av skillnaden gick att tillskrivas *moral hazard*. Ett tillvägagångssätt inom test för *moral hazard* inom försäkringsmarknader är att studera data från enkäter till skillnad från försäkringsbolagets egen statistik. Detta är något Finkelstein & McGarry (2006) utvecklade och som flertalet andra undersökningar, däribland Rowel m.fl. (2017), har studerat. Att enbart använda tvärsnittsdata anses problematiskt för att estimera *moral hazard*, menar Rowel m.fl. (*op.cit.*). Samtliga studier som väljer att studera effekten av informationsasymmetri tillsammans med tillgänglig data graderar dessa i enlighet med vissa principer. Dessa principer säger att paneldata och enkätdata är bättre för att uppskatta effekterna jämfört med vad tvärsnittsdata från försäkringsbolagets egen statistik är. Det råder alltså en skillnad mellan primär- och sekundärdata vid undersökning av *moral hazard* enligt litteraturen.

Slutligen finner vi att i en artikel från Chen & Jiang (2019) behandlas olika faktorer som påverkar premieprissättningen. Detta leder i förlängningen också till en bedömningsgrund från försäkringsbolagets sida gällande individens risk. Författarna nämner bland annat ålder, jobb, hemvistelse och antal körda mil som faktorer vilka tillsammans påverkar risken för en trafikolycka. Chen & Jian hämtar evidens av dessa variabler från Paefgen m.fl. (2014) vilket visade på en signifikant positiv korrelation mellan antal körda mil och risk för en trafikolycka. Vidare behandlar författarna ämnet perfekt information gällande individens körvanor samt i vilken utsträckning föraren utsätter sig och andra i trafiken för risk. Detta är och förblir en viktig aspekt i de problem försäkringsbolagen väljer att prissätta, argumenterar författarna. I och med den individuella skillnaden mellan olika förare samt att det råder informationsasymmetri mellan förare och försäkringsbolag uppstår det heterogenitet när det kommer till aktuariemässiga bedömningar om individer. Enligt författarna går det att undkomma denna problematik genom att installera ”över-

vakningsteknik" i bilarna som mäter individens körvanor (Chen & Jiang, 2019, s. 1969). Detta skulle i förlängningen råda bot på de problem försäkringsbolagen ställs inför när det gäller att basera premieprissättningen på individuella variabler som inte perfekt matchar den egentliga korrelationen av sannolikheten för en olycka (Desyllas & Sako, 2013). Samt problemet för försäkringsbolaget att uppskatta den ena *moral hazard*-problematiken då asymmetrisk inlärning uppstår vid individens ålder och den andra *moral hazard*-problematiken gällande utelämnade av olycksfall (Chiappori, 2000)

3 Teori

Att det råder perfekt information på marknaden är ett implicit antagande från det grundläggande nyttoteoremet (Von Neumann & Morgenstern, 2007). Alltså är samtliga karaktäristiska drag observerade av samtliga deltagare på marknaden. Utan detta villkor kan inte distinkta marknader existera för varor som har olika attribut och därmed faller antagandet om kompletta marknader (Mas-Colell, Whinston, Green m. fl., 1995). I verkligheten är situationer med informationsasymmetri regel snarare än undantag. Ett tydligt exempel hämtat från verkligheten är när ett företag anställer en ny medarbetare och är osäker på hur produktiv individen är. Denna informationsasymmetri definieras i litteraturen som *adverse selection* och *moral hazard*.

3.1 Adverse selection

Adverse selection, alternativt "snedvridet urval" definieras i litteraturen som en situation vilken uppstår när ena parten i ett avtal har mer information än den andra parten gällande avtalet. (Chiappori & Salanié, 2012). I Akerlof (1970) återfinns en utförlig förklaring men vi väljer att illustrera modellen med ett förenklat exempel gällande en försäkringsmarknad från Einav & Finkelstein (2011) samt Autor (2016).

Låt oss anta en homogen grupp av konsumenter där individer är von Neumann-Morgenstern maximerande, VNM, med en nyttofunktion av typen $U(f_i) = \ln(f_i)$ (Von Neumann & Morgenstern, 2007). Varje konsument har en ursprunglig förmö-

genhet på 200 kr samt en femtioprocentig sannolikhet att råka ut för en skada L_i där förlusten L är lika med $L_i \cdot 100$. Konsumenterna är enhetliga och indexerade i intervallet $i \in [0, 1]$ För en konsument med $i = 0{,}7$ gäller följande:

$$\begin{aligned} E[f_i] &= 200 - 0{,}5 \cdot 70 = 165 \text{ kr} \\ U(E[f_i]) &= \ln(165) = 5{,}11 \\ E[U(f_i)] &= 0{,}5 \cdot \ln(200) + 0{,}5 \cdot \ln(130) = 5{,}08 \\ CE(E[U(f_i)]) &= e^{5{,}08} = 160 \text{ kr} \end{aligned}$$

Utifrån ovanstående ekvation går det att se att en individ med $i = 0{,}7$ har en betalningsvilja för riskpremien på 5 kr för att uppnå heltäckande försäkring. Sista ledet står för ”certainty equivalent”, CE, och är nyttan individen får av det säkra valet. Vi generaliserar uttrycket till samtliga konsumenter i och får följande resultat,

$$\begin{aligned} U[E(f_i)] &= 150 - \ln(200 - 0{,}5 \cdot 100 \cdot i) \\ E[U(f_i)] &= \ln(200^{0{,}5} \cdot (200 - 100 \cdot i)^{0{,}5}) \\ CE(E[U(f_i)]) &= e^{(E[U(f_i)])} \end{aligned}$$

Alla individer vet sitt värde i med den distinktionen att värdet är okänt för försäkringsgivaren. Vi målar upp ett scenario där enbart en försäkringstyp går att köpa och ersätter L_i till konsument i vid ett skadeärende. Försäkringsgivaren prissätter försäkringen baserat på förväntade förluster över hela populationen till priset 25 kr då $L_i \sim U[0, 100]$ leder till $E[L_i] = 50$ och sannolikheten för en skada är 50 procent.

Låt oss analysera vilka konsumenter som väljer att köpa en försäkring givet detta villkor. Vilka är de förväntade vinsterna för försäkringsgivaren? Först definerar vi marginalkonsumenten i' som är indifferent mellan att köpa försäkringen och att inte ha någon försäkring. Konsumenter som har större förväntad förlust än i'

köper försäkringen då premien är samma för alla konsumenter. Konsumenter som har lägre förväntad förlust än i' köper inte försäkringen. Vi vill alltså lösa följande med avseende på i' i ekvationens vänsterled.

$$E[U(f_{i'})] = 0{,}5 \cdot \ln(200) + 0{,}5 \cdot \ln(200 - 100 \cdot i') = \ln(200 - 25)$$

Vänsterledet avser förväntad nytta av i' utan försäkring medan högerledet avser förmögenhet av i' med försäkring.

$$\begin{aligned}
0{,}5 \cdot \ln(200) + 0{,}5 \cdot \ln(200 - 100 \cdot i') &= \ln(175) \\
\ln(200 - 100 \cdot i') &= (2 \cdot \ln(175) - \ln(200) \\
200 - 100 \cdot i' &= \exp[2 \cdot \ln(175) - \ln(200)] \\
i' &= [\exp[2 \cdot \ln(175) - \ln(200)] - 200]/(-100) \\
i' &= 0{,}46
\end{aligned}$$

Vi löser ekvationen med avseende på marginalkonsumenten och ser att $i' = 0{,}46$. Därmed är konsumenter i intervallet $i' \in [0{,}46, 1]$ villiga att köpa försäkringen. Notera att konsumenter $\in [0{,}46, 0{,}5)$ också köper försäkringen även om deras förväntade förluster är lägre än 25 kr. Från försäkringsgivarens perspektiv ser vi att försäkringen förlorar pengar i genomsnitt. Förväntad kostnad per försäkrad individ är följande,

$$100 \cdot 0{,}5 \cdot \frac{1 + 0{,}46}{2} = 36{,}50 \text{ kr}$$

Försäkringsgivaren säljer försäkringen för 25 kr vilket framgår i inledningen av exemplet. Alltså, kan inte denna situation leda till marknadsjämvikt. Problemet försäkringsgivaren står inför är *adverse selection*. Marknaden är ineffektiv då allo-

keringen sker genom förhållandet mellan marginalkostnad, MC, och efterfråga, D, medan jämvikts-allokeringen sker mellan genomsnittlig kostnad, AC, och efterfråga (Einav & Finkelstein, 2011).

3.2 Moral hazard

I föregående del beskrev vi informationsasymmetri vid kontraktets början. Skev fördelning i information mellan två parter efter kontraktet har ingåtts kallas i litteraturen för *moral hazard* och är ett principal-agent problem.

I den nationalekonomiska litteraturen definieras *moral hazard* som individens benägenhet att förändra sitt beteende när den är försäkrad, med den viktiga distinktionen att försäkringsgivaren inte kan observera denna beteendeförändring (Boyer & Dionne, 1989). Sannolikheten för en olycka är en endogen variabel och beror på individens val i förhållande till försäkringens täckning. Genomgående i litteraturen görs en viktig distinktion mellan *ex ante* respektive *ex post moral hazard*. Då denna distinktion är viktig väljer vi att grundligt redovisa skillnaden mellan de två. *Ex ante moral hazard* representerar förändringen av individens beteende före en olycka och förklaras utförligt i styckets inledning. Försäkringsbolag är intresserade av olyckor som resulterar i en skadeanmälan. Beslutet om en olycka övergår i en skadeanmälan är delvis beroende av individens val och därmed har kontraktets utförande en signifikant inverkan (Chiappori & Salanié, 2012). I litteraturen kallas detta fenomen *ex post moral hazard*. Vi kan illustrera skillnaden mellan de två genom ett exempel. En individ som uppvisar ett riskabelt beteende efter att den köpt en bilförsäkring definieras som *ex ante moral hazard*. Om ett ärende sedan har skett och individen undanhåller väsentlig information definieras det som *ex post moral hazard*.

Distinktionen mellan *ex ante* samt *ex post* är viktig. Att ta hänsyn till dess möjliga inverkan på empiriska resultat är väsentligt vid analys av ämnet. Inom försäkringsbranschen ses ofta tillägget av en självriskklausul[3] som en enkel men effektiv åtgärd för att minska antalet mindre ärenden som trots dess ringa karaktär utgör fasta kostnader för försäkringsbolaget. Individer som ingår i ett försäkringsavtal med en

[3]Del av skadekostnaden individen betalar själv.

hög självrisk är mindre benägna att rapportera mindre skador och detta kan leda till missvisande resultat. En falsk korrelation kan uppstå mellan valet av kontrakt och antal registrerade ärenden (Chiappori & Salanié, 2012).

3.3 Exklusiva respektive allmänna kontrakt

I litteraturen framkommer en viktig distinktion gällande kontraktens exklusivitet. Exklusiva kontrakt medför att försäkringsgivaren har ett exklusivt avtal med försäkringstagaren och är vedertaget när det gäller bilförsäkringar. Hade kontrakten vi studerade varit allmänna skulle individer köpa ett antal mindre försäkringar från olika bolag för att försäkra sig vilket hade lett till andra empiriska resultat (Chiappori & Salanié, 2012). Alltså, förutsätter denna studie att en individ enbart har ett avtal med ett försäkringsbolag.

3.4 Teorins prediktioner

I forskningsöversikten redovisades studier gällande korrelationen mellan täckning och risk. Vi har tidigare beskrivit att en försäkringsgivare inte har möjlighet att göra en distinktion med avseende på risknivå eftersom att denna är okänd. Därmed erbjuds samma kontrakt till individer som har olika nivå av risk (Cohen & Siegelman, 2010). Teorins prediktion är att individer med högre risk köper mer försäkring än individer med lägre risk. När försäkringsgivare erbjuder avtal med varierande grad av försäkringstäckning är teorins prediktion att korrelationen är positiv och signifikant gällande täckning och risk.

Chiappori & Salanie (2012) ger oss ett exempel på hur denna korrelation skattas. Om korrelation mäts måste risknivå samt försäkringsnivå mätas separat. Då vi är intresserade av *ex post* risk approximeras den med en binär variabel ϑ. Det vill säga, risknivå hos en individ representeras av en 0-1 binär variabel där en etta innebär att en skada har inträffat. Nivån av försäkring, i litteraturen, ofta skrivet som ”coverage” kallar vi φ och även den kan vara binär. Låt säga att den representeras av en 0-1 binär variabel där värde 1 medför att en individ har helförsäkring och värde 0 representerar en individ med bara den obligatoriska trafikförsäkringen. Slutligen skapar vi en radvektor, X som representerar alla egenskaper som försäkringsbolaget

känner till och tillämpas i deras prissättning av kunder. Ett grundläggande sätt att mäta korrelationen ρ är följande vis,

$$\rho = \Pr(\vartheta = 1|\varphi = 1, \mathrm{X}) - \Pr(\vartheta = 1|= 0, \mathrm{X}) \tag{3.1}$$

Standardteorin inom litteraturen är att en signifikant korrelation ρ som funktion av radvektorn X innebär att informationsasymmetri är orsaken (Chiappori & Salanié, 2012). Ett annat tillvägagångsätt vilket tillämpades av samma författarduo var att bygga två separata probit-modeller och utföra en bivariat probit-analys.

$$\begin{aligned} \vartheta &= 1(\mathrm{f}(\mathrm{X}) + \varepsilon > 0) \\ \varphi &= 1(\mathrm{g}(\mathrm{X}) + \gamma > 0) \end{aligned} \tag{3.2}$$

Där ϑ är skadesannolikhet och φ är försäkringstäckning. En signifikant korrelation av residualerna ε och γ ger vikt åt prediktionen mellan täckning och risk (Chiappori, 2000).

Cohen & Siegelman (2010) förklarar, att det är viktigt att betona att prediktionen medför en korrelation mellan täckning och risk vid kontroll av variabler som försäkringsgivaren känner till . Det är just dessa icke-observerbara egenskaper som tilldelas *moral hazard* samt *adverse selection*. Faktumet att residualerna beror på antingen *moral hazard* eller *adverse selection* är diskuteras i litteraturen. Detta eftersom båda begreppen visar sig i ett samband, med omvänd kausalitet. Mellan täckning och risk är det svårt att påvisa *adverse selection* eller *moral hazard* med statistisk säkerhet. Låt oss måla upp ett scenario där en studie kommer fram till en positiv och signifikant korrelation mellan täckning och risk när man kontrollerar för observerade variabler. En slutsats skulle kunna dras att en individ som köper mer försäkring är mer benägen att hamna i en olycka. Men förklaringen är tvetydig. Det går att hävda att individens val av försäkring minskar incitamenten att köra säkrare. Det går också att argumentera för att individen är medveten om sin egen risknivå och väljer därmed en försäkring som kompenserar mer vid en olycka.

4 Empiri

Denna studie hämtar empiri från två olika källor. Den första källan till data är från det svenska försäkringsbolaget Länsförsäkringar Uppsala och avser kunder från året 2018. Efter förfrågan av Länsförsäkringar Uppsala fick vi ett utdrag från deras register. Den andra källan till data är hämtad från branchorganisationen Svensk försäkrings offentliga databas och avser samma tidsperiod.

Data från Länsförsäkringar Uppsala åsyftar försäkringsbolagets kunder. När denna data tillhandhölls från bolaget önskade vi ett slumpmässigt stickprov för att säkerställa autenticitet. Detta då en naturlig fördelning av stickprovet överensstämmer bättre med försäkringsbolagets totala kunder. Totalt finns det tio variabler i datasetet och sammanlagt handlar det om 5 413 kunder varav 775 stycken skadeanmälningar. Datapunkter är hämtade ur den lokala databasen vid Länsförsäkringar Uppsala och motsvarar några av de variabler försäkringbolaget observerar sina kunder med. Att vi får åtkomst till dessa datapunkter, vilket har som syfte att hjälpa försäkringsbolaget med sin analys, gör att vi använder oss av sekundärdata. Vidare är samtlig data från Länsförsäkringar Uppsala och Svensk försäkring utformad som tvärsnittsdata då vi saknar information gällande andra år än 2018. Länsförsäkringar Uppsala är ett lokalt förankrat dotterbolag till moderbolaget Länsförsäkringar AB. Sammanlagt är de lokala dotterbolagen delägare i moderbolaget och med en marknadsandel på 30,6 procent på bilförsäkringsmarknaden 2020 är moderbolaget den största aktören på den svenska marknaden (Svensk försäkring, 2018).

Data från Svensk försäkring åsyftar skadestatistik från samtliga försäkringsbolag på den svenska bilförsäkringsmarknaden året 2018. Denna data utgör en målpopulation av samtliga skadeanmälningar på den svenska bilförsäkringsmarknaden. Totalt finns det nio typer av skadeanmälningar med 1 262 471 observationer i datasetet. Svensk försäkring är en branschorganisation för svenska försäkringsbolag och samlar bland annat statistik gällande samtliga aktörers skadedata (Svensk försäkring, 2021). Detta resulterar, givet samma logik som för data från Länsförsäkringar Uppsala, att vi arbetar med sekundärdata.

Det finns flera anledningar till varför vi valt att arbeta med dessa datakällor. På

grund av att Länsförsäkringar är en såpass stor aktör på den svenska bilförsäkringsmarknaden går vissa generaliserbara antaganden att tillämpas vid analysen. Hade analysen varit centrerad runt ett mer nischat bilförsäkringsbolag minskar de generella slutsatsernas betydelse. En annan aspekt som är positiv med data från Länsförsäkringar Uppsala och Svensk försäkring är att inom ramen för denna studie, möjliggör användningen av sekundärdata att vår analys kan innefatta ett mycket stort och detaljerat dataset under den relativt korta tidshorisonten denna studie utformas vid. (Bryman, 2016). Detta beror också på att vi valt att analysera bilförsäkringsmarknaden och dess aktörer. Försäkringsbolag, branschorganisationer och statliga myndigheter har stora databaser vilka de kan lämna ut till forskningssyfte. Ytterligare en aspekt som tåls att lyftas fram är att valet av data erbjuder oss en möjlighet att utföra undergruppsanalyser (Bryman, 2016). Detta eftersom ett stort urval av individer med olika kategoriseringar återfinns i vårt dataset. Det betyder alltså att vi har en möjlighet att studera olika gruppers inverkan vid analysen. Anledningen till att vi använder oss av statistik från Svensk försäkring är att data från Länsförsäkringar Uppsala avser skadeanmälningar från kunder till försäkringsbolaget, vilket också Svensk försäkrings data avser. Vi betraktar data från Svensk försäkring som målpopulationen samtidigt som data från Länsförsäkringar Uppsala är ett stickprov. Detta på grund av att vi vill analysera den marginella effekten av olika variabler för att sedan jämföra dessa med marginella effekten av att kategorisera om målpopulationens skadeanmälningar.

Bristen på variabler är något vår analys är belastad med och anledningen till detta är mångbottnad. En anledning är att enligt lagen om kompletterande bestämmelser till EU:s dataskyddsförordning (SFS 2018:218), GDPR, uppstår det problem för försäkringsbolaget att lämna ut viss information[4]. Detta är information som gör det möjligt att identifiera individer till specifika datapunkter. I synnerhet uppstår det problem i mängden variabler vi har till vårt förfogande från Länsförsäkringar Uppsala. Ju fler variabler försäkringsbolaget delar med sig av, desto lättare blir det att identifiera en enskild individ. Utöver denna GDPR-lagstiftning föreligger även problem i vilka datapunkter Länsförsäkringar Uppsala vill lämna ut. Detta

[4]GDPR-lagstiftningen ger skydd åt fysiska personer med avseende på behandling av personuppgifter i det fria flödet samt direktivet till bland annat företag SFS 2018:218.

eftersom det är en företagshemlighet hur försäkringsbolagen prissätter sina kunder. Givet dessa två förutsättningar, har vi önskat efter variabler som motsvarar den metod vi vill tillämpa inom ramen för denna uppsats.

4.1 Beroende variabel

I denna studie utforskar vi den beroende variabeln skadeanmälning. I rapporteringen av variabeln redovisas förekomsten av en skadeanmälning som ett datum. Då vi analyserar data från 2018, vilket är en bestämd tidsperiods antal observationer, kodar vi om skadeanmälning som en binär variabel. Vi tillskriver observationen ett värde 0 om ingen skadeanmälning och värde 1 vid inrapporterad skada. Anledningen till detta är att vi vill att den beroende variabeln ska passa vår modell.

4.2 Oberoende variabler

De oberoende variabler vi testar mot den beroende variabeln är följande:

- *Ålder*. Innefattar försäkringstagarnes ålder 2018.
- *Årsmodell*. Detta är en variabel vilken beskriver året och upplagan av bilen.
- *Total årsnettopremie*. Innefattar priset på premien plus övriga avgifter och betalas av individen årsvis.
- *Körsträcka*. I denna klassificering ingår fem stycken kategorier vilka beskriver hur långt bilen kört det gångna året. Intervallen $\leq$1000, 1001–1500, 1501–2000, 2001–2500 samt >2500 inkluderar samtliga datapunkter. Nämnda intervall behandlar antal körda mil under 2018.
- *Trafikförsäkring*. Denna variabel beskriver om individen har trafikförsäkring eller ej. Försäkringen ger täckning för trafikskador. Viktigt att notera är att enligt svensk lag är bilistens skyldig att inneha en trafikförsäkring. Enligt trafikskadelagen är undantagen till detta att ditt fordon inte är registrerat hos vägtrafikregistret[5], fordonet är avställt eller att försäkringen har upphört av

[5]Transportstyrelsens register på samtliga fordon och körkortsinnehavare (Transportstyrelsen, 2021).

någon annan anledning (2 § Trafikskadelag 1975:1410). Vi betraktar samtliga individer utan trafikförsäkring som avställda fordon för att analysen ska vara konsekvent.

– *Delkaskoförsäkring.* Klassificeringen innefattar individer med halvförsäkring, detta utöver trafikförsäkringen. De som har delkaskoförsäkring har alla trafikförsäkring och tillsammans är dessa halvförsäkrade. Försäkringen ger skydd för stöld, allrisk, assistans och räddning. Utöver detta ges också rättsskydd och ersättning vid kris-, brand-, glas- och maskinskador (Länsförsäkringar, 2021).

– *Vagnförsäkring.* Innefattar individer som valt en helförsäkring. Detta inkluderar det skydd som trafik- och delkaskoförsäkringen erbjuder. Utöver detta skydd erbjuder denna försäkring också ersättning vid vagnskador där försäkringsinnehavaren varit vållande, någon form av skadegörelse eller yttre olyckshändelser. Ersättning erhålles även för detaljer på bilen som tar skada, till exempel en monterad takbox (Länsförsäkringar, 2021).

– *Merförsäkring.* Denna försäkring går endast att teckna givet att individen har en hel- eller halvförsäkring. Det går alltså att ha en merförsäkring kombinerat med antingen en delkasko- eller vagnförsäkring. Skillnaden mellan merförsäkringen och de andra försäkringstyperna är att det ingår hyrbil till den som utsätts för någon form av skadegörelse eller stöld av bil. Det går även att välja att få kontant ersättning istället för tillkomst till en hyrbil. Utöver detta ingår en lägre självrisk vid viltolyckor eller skadegörelse av bil (Länsförsäkringar, 2021).

4.2.1 Kontinuerliga variabler

Totalt analyseras tre kontinuerliga variabler i analysen. Från den deskriptiva statistiken (se *tabell 1*) finns det några beröringspunkter som lyfts i denna empiridel. Till att börja med finns det en avvikande frekvenssumma, angående variabeln ålder. Detta beror på att inom denna variabel ingår en ytterligare klassificering, privatperson eller företag. Vi exkluderade samtliga observationer som var av företagstyp. Däremot fanns det två observationer som varken var av privat eller

företagstyp. Genom denna distinktion går det att studera några avvikande datapunkter. Vi ser till exempel att det återfinns individer med en ålder på 103 år till skillnad från medelvärdet som är på 47 år. Från premie i *tabell 1* ser vi också att den högsta premien är på en bil som betalar 42 531 kr per år vilket också avviker kraftigt från medelvärdet på 4 529 kr per år.

Tabell 1: Deskriptiv statistik för kontinuerliga variabler.

Variabel	**Antal**	**Medelvärde**	**Median**	**Max(Min)**	**SD**
Premie (kr)	5 511	4 529	3 886	42 531(668)	2900
Ålder (år)	5 413	47	46	103(22)	14,3
Årsmodell (år)	5 511	2004	2006	2018(1934)	9,4

Källa: Författarnas beräkningar av data från Länsförsäkringar Uppsala (2021).

Not: SD står för respektive variabels standardavvikelse.

4.2.2 Dummyvariabler

För att analysera försäkringstagarnas antal körda mil och vilken typ av försäkring anpassar vi värdena till dummyvariabler. Variabeln körsträcka delas in i fyra nya kolumner. Det lägsta intervallet, $\leq$1000, agerar som den utelämnade variabeln och ingår i referensgrupperingen. Detta för att inte hamna i dummyvariabel-fällan. Justerar vi för samtliga individer som saknar trafikförsäkring, och därmed enligt svensk lag inte tillåts använda bilen i trafiken, bortkommer 1 019 observationer från variabel $\leq$1000. *Tabell 2* visar fördelningen av individer i de olika intervallen av körsträcka. De två lägsta intervallen har betydligt fler observationer i vårt stickprov följt av färre observationer som har de två högsta intervallen av körsträcka. Slutsatsen är en avtagande mängd individer för varje ökning i intervallen av körsträcka.

Gällande försäkringstyp kodar vi om datasetet för att utröna vilka individer som tillhör vilken försäkringstyp. Det finns en viss problematik i att samtliga som har, till exempel vagnförsäkring, per definition också har trafikförsäkring. Denna problematik löser vi genom att justera om typerna i nya kategoriseringar. En annan justering vi tillämpar på datasetet är att vi utesluter individer som inte har någon trafikförsäkring. Tillämpningen av denna justering är nödvändig för att vi skall

Tabell 2: Deskriptiv statistik för variabeln körsträcka.

Variabel	Antal	Procent	Kumulativ procent
Körsträcka (mil/år)			
≤1000	2 507	45,5	45,5
1001–1500	1 986	36,0	81,5
1501–2000	656	11,9	93,4
2001–2500	181	3,3	96,7
>2500	181	3,3	100
Total	5 511	100	–

Källa: Författarnas beräkningar av data från Länsförsäkringar Uppsala (2021).

kunna analysera data på vårt ramverk. Individer utan trafikförsäkring bör inte utgöra någon risk i trafiken. Nästa steg i att anpassa vår data från Länsförsäkringar Uppsala handlar om att särskilja mellan två olika set av täckning.

Första kategoriseringen särskiljer på individer som minst har vagnförsäkring och de som inte har det med en dummyvariabel $d_{i,5}$, $i = 1, \ldots, n$. Detta innefattar individer som har vagnförsäkring och individer som har vagn- och merförsäkring. Vi låter $d_{i,5} = 1$ om individ i har vagnförsäkring. De individer som utesluts från den första grupperingen är individer som enbart har trafikförsäkring, delkaskoförsäkring samt delkasko-, och merförsäkring. Dessa individer ingår i referensgruppen och vi låter $d_{i,5} = 0$ om individ i inte har vagnförsäkring. Kategoriseringen framgår i (4.1).

$$d_{i,5} = \begin{cases} 1 & \text{Vagnförsäkring} \\ 0 & \text{Ingen vagnförsäkring} \end{cases} \tag{4.1}$$

Andra kategoriseringen särskiljer mellan individer som minst har delkaskoförsäkring och de som inte har det med dummyvariabeln $d_{i,5}$, $i = 1, \ldots, n$. Detta inkluderar individer som enbart har delkaskoförsäkring, har delkasko- och merförsäkring, har vagnförsäkring samt de som har vagn- och merförsäkring. Vi låter $d_{i,5} = 1$ om individ i har delkaskoförsäkring. De individer som utesluts från den första grupperingen är individer som enbart har trafikförsäkring. Dessa individer ingår i referensgruppen vid detta test. Vi låter $d_{i,5} = 0$ om individ i inte har delkaskoför-

säkring. Kategoriseringen framgår i (4.2).

$$d_{i,5} = \begin{cases} 1 & \text{Delkaskoförsäkring} \\ 0 & \text{Ingen delkaskoförsäkring} \end{cases} \tag{4.2}$$

Sista kategoriseringen av datasetet från Länsförsäkringar Uppsala är av den kontinuerliga variabeln ålder. Vi kodar om denna variabel till en dummyvariabel. Detta eftersom att vi vill kontrollera för hur ålder påverkar sannolikheten för en skadeanmälning. Likt Chiappori & Salaniés (2000) metodik där författarna kategoriserar om ålder till två binära grupper. Grupperna inkluderar unga förare och icke-unga förare genom dummyvariabeln $d_{k,}$, $k = 1, \ldots, n$. Vi låter $d_{k,l} = 1$ om individ k är en ung förare. Detta betyder att individen är under 25 år. Icke-unga förare ingår i referensgruppen och vi låter $d_{k,1} = 0$ om individ k är 25 år eller äldre. Kategoriseringen framgår i (4.3).

$$d_{k,1} = \begin{cases} 1 & < 25 \text{ år} \\ 0 & \geq 25 \text{ år} \end{cases} \tag{4.3}$$

I *tabell 3* finns tre beröringspunkter vi vill lyfta fram. Deskriptiv statistik från de två seten av försäkringstyper samt deskriptiv statistik från den binära variabeln unga förare. I det första setet är fördelningen mellan referens- och vagnförsäkringsgruppen jämn med 45,5 respektive 54,5 procent. I det andra setet är fördelningen mellan de två grupperna relativt ojämn då vi ser en skillnad på 66,5 procentenheter. Vid kategoriseringen för unga och icke-unga individer framgår det från *tabell 3* att fördelningen är relativt ojämn. Enbart 1,1 procent av stickprovet utgörs av unga förare samtidigt som 98,9 procent är icke-unga förare.

4.3 Kontrollerande variabler

Data från Svensk försäkring är uppdelad i nio variabler med totalt 1 262 471 observationer (se *tabell 4*). För att det skall råda stringens i analysen väljer vi att redovisa samtliga variabler från Svensk försäkring i enlighet med Länsförsäkringar

Tabell 3: Kategorisering av Länsförsäkring Uppsala.

Variabel	$d_{i/k}$	**Antal**	**Procent**
Set 1			
Referens	0	2 506	45,5
Vagn	1	3 005	54,5
Set 2			
Referens	0	924	16,8
Delkasko	1	4 587	83,2
Total	–	5 511	100
Ålderskategori			
Icke-unga	0	5 352	98,9
Unga	1	61	1,1
Total	–	5 513	100

Källa: Författarnas beräkningar av data från Länsförsäkringar Uppsala (2021).

Uppsalas kategorisering. Genom tidigare redogörelse kring vad varje försäkringstyp har för täckning går det att kategorisera om Svensk försäkrings data. Det framgår ur *tabell 4* att andelen skadeanmälningar av glas- och vagnskadekaraktär är kraftigt dominerande i antal anmälda fall.

Likt kategoriseringen av försäkringstyper från föregående kapitel justerar vi data i syfte att särskilja mellan vilken täckning de två kategorierna inkluderar. I det första setet gör vi precis som under kategoriseringen för data från Länsförsäkringar, det vill säga att vi skiljer mellan vagnskador och resterande skadeanmälningar. Detta betyder att vi inkluderar samtliga skador till vilken enbart vagnsförsäkring har täckning för. Genom den nya kategoriseringen erhåller vi två binära grupper som enbart inkluderar skador till vilket vi kan skilja mellan försäkringstyp och skadeart. Dessa separeras genom dummyvariabeln d_i, $i = 1, \ldots, n$. Vi låter $d_i = 1$ om skadeanmälning i uteslutande faller inom vagnförsäkring. I referensgruppen låter vi $d_i = 0$ om skadeart i inte uteslutande faller inom vagnförsäkringens täckning.

Tabell 4: Deskriptiv statistik för Svensk försäkring

Typ av skada	Antal	Procent	Kumulativ procent
Trafikförsäkring			
Trafik	147 358	11,7	11,7
Delkaskoförsäkring			
Ansvar	931	0,1	11,8
Brand	6 298	0,4	12,2
Glas	520 040	41,2	53,4
Maskin	80 713	6,4	59,8
Rättsskydd	1 011	0,1	59,9
Räddning	91 727	7,3	67,2
Stöld	30 872	2,4	69,6
Vagnförsäkring			
Vagn	383 521	30,4	100
Total	1 262 471	100	–

Källa: Författarnas beräkningar av data från Svensk försäkring (2018).
Fetstil beskriver vilken försäkringtyp som ger täckning för skadan.

Kategoriseringen framgår i (4.4).

$$d_i = \begin{cases} 1 & \text{Uteslutande vagnförsäkring} \\ 0 & \text{Inte uteslutande vagnförsäkring} \end{cases} \tag{4.4}$$

Det andra setet skiljer mellan skadeart som faller uteslutande inom trafikförsäkring eller inte med dummyvariabeln d_i, $i = 1, \ldots, n$. Vi låter $d_i = 1$ om skadeanmälning i uteslutande faller inom delkasko- eller vagnförsäkring. I referensgruppen låter vi $d_i = 0$ om skadeart i inte uteslutande faller inom delkasko- eller vagnförsäkrings täckning. Kategoriseringen framgår i (4.5).

$$d_i = \begin{cases} 1 & \text{Uteslutande delkasko- eller vagnförsäkring} \\ 0 & \text{Inte uteslutande delkasko- eller vagnförsäkring} \end{cases} \tag{4.5}$$

I *tabell 5* framgår deskriptiv statistik från kategoriseringen av Svensk försäkrings data. Noterbart är fördelningen mellan de olika grupperna. Andelen skadeanmäl-

ningar som är av skadeart *vagn*, 30,4 procent, är relativt jämn med andelen av skadeart *referens*, 69,4 procent. Detta till skillnad från den andra kategoriseringen. I den andra kategoriseringen är andelen som är av skadeart *delkasko*, 88,3 procent men enbart 11,7 procent av skadeanmälningarna i skadearten *referens*.

Tabell 5: Kategorisering av Svensk försäkring.

Variabel	d_i	**Antal**	**Procent**
Kategori 1			
Referens	0	878 950	69,6
Vagn	1	383 521	30,4
Kategori 2			
Referens	0	147 358	11,7
Delkasko	1	1 115 113	88,3
Total	–	1 262 471	100

Källa: Författarnas beräkningar av data från Svensk försäkring (2018).

5 Metod

I forskningsöversikten redovisades olika tillvägagångsätt för att undersöka korrelationen mellan täckning och risk på bilförsäkringsmarknaden. Genomgående i vår metod och därmed i tolkningen av våra resultat finns det en distinktion som läsaren bör vara medveten om. I studien definierar vi risk som sannolikhet för en skadeanmälan. Tidigare forskning som vi undersökt har erhållit data som särskiljer på individers risknivå. Till exempel information om vållande av skada eller kunna exkludera datapunkter där två bilar varit inblandade. Anledningen till detta är självklar. Om vi estimerar skadeanmälningar som funktion av försäkring är det vedertaget att fler skador anmäls då en högre försäkring täcker fler typer av händelser. Genom vår implementering av en kontrollerande variabel försöker vi minimera skevheten som uppstår och som kan påverka våra resultat.

Vi särskiljer mellan empiri respektive metod använder. Skillnaden mellan de två avsnitten är att vi använder ålder som en kontinuerlig respektive diskret variabel. Anledningen till detta är att vi också vill undersöka Chiappori & Salaniés (2000) tillvägagångsätt där författarna jämför gruppen unga förare med icke-unga förare.

Detta i syfte att uppskatta den asymmetriska inlärningen en individ tillskansar sig vid erfarenhet av trafik. Denna erfarenhet gör individen mer varse om sin egen körförmåga vilket kan påverka resultaten.

5.1 Logit-modell

I denna uppsats använder vi en modell som är lik en binär probit-modell där sannolikheten för en händelse skattas med hjälp av olika oberoende variabler. Metoden vi använder är logistisk regression och mer specifikt en binär logit-modell. I litteraturen är logit- och probit-modeller likvärdiga gällande resultat men logit-modellen är lättare att förstå och därmed blir resultaten tydligare (Gujarati, 2003). Vidare antar vi att modellen är additiv med inga interaktionseffekter. Att modellen är binär medför att den beroende variabeln, y_i, $i = 1, \ldots, n$, kan anta 0 eller 1. I vårt fall representerar y_i två utfall. Detta åskådliggörs i (5.1),

$$y_i = \begin{cases} 1 & \text{Skadeanmälan under tidsperioden} \\ 0 & \text{Ej skadeanmälan under tidsperioden} \end{cases} \tag{5.1}$$

En ytterligare fördel med att använda en logit-modell jämfört med en linjär regression är vilka värden de oberoende variablerna kan anta. I jämförelse med en linjär regression kan de oberoende variablerna vara både diskreta och kontinuerliga (DiGangi & Moore, 2012). Detta är motiverat då våra förklarande variabler är av båda kategorierna. Ekvationerna vi vill estimera i vår uppsats representeras på följande vis.

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \delta_1 d_{i,1} + \delta_2 d_{i,2} + \delta_3 d_{i,3} + \delta_4 d_{i,4} + \delta_5 d_{i,5} \tag{5.2}$$

$$y_k = \beta_0 + \beta_1 x_{k,1} + \beta_2 x_{k,2} + \delta_1 d_{k,1} + \delta_2 d_{k,2} + \delta_3 d_{k,3} + \delta_4 d_{k,4} + \delta_5 d_{k,5} + \delta_6 d_{k,6} \tag{5.3}$$

Sannolikheten för en skada är en funktion av variablerna $x_{i,j}$, $d_{i,j}$ för y_i, $i = 1, \ldots, n$, i (5.2). Detta är modellen vilken behandlar ålder som en kontinuerlig variabel. $x_{k,l}$ och $d_{k,l}$ innefattar våra oberoende variabler för y_k, $k = 1, \ldots, n$, och behandlar ålder som en diskret variabel i (5.3). Detta motsvarar mer specifikt det som är känt hos försäkringsingsgivaren och i vårt dataset. Princeton (2007) menar att det uppstår ett problem eftersom vänsterledet kan anta värdena 0 eller 1 medan högerledet kan anta diskreta och kontinuerliga värdemängder. Vi löser detta genom att transformera (5.2), men samma villkor gäller också för (5.3). Sannolikheten y_i transformeras för att erhålla ett resultat där bägge leden har samma villkor. Vi justerar sannolikheten för en skada y_i till odds.

$$\text{Odds}_i = \frac{y_i}{1 - y_i} \tag{5.4}$$

Definitionen av odds är förhållandet mellan, till exempel, insats och vinst (Körner & Wahlgren, 2006). Fortsätter vi med Princetons (2007) metodik framgår följande. Om sannolikheten y_i för en händelse är 50 procent blir oddsen för förhållandet 1:1. I nästa steg tar vi den naturliga logaritmen av (5.4) för att få log-odds eller logit.

$$\pi_i = \text{logit}(y_i) = \log\frac{y_i}{1 - y_i} \tag{5.5}$$

Då transformationen i (5.5) är injektiv kan vi gå baklänges från logit värden, π_i till sannolikheten y_i. Lösning av (5.5) med avseende på y_i blir följande:

$$y_i = \text{logit}^{-1}(\pi_i) = \frac{e^{\pi_i}}{1 + e^{\pi_i}} \tag{5.6}$$

Med (5.5) respektive (5.6) i grunden skapar vi två logit-modeller. Dessa modeller avser de två typer av regressioner vi vill estimera i denna studie beroende på hur

vi behandlar ålder som variabel.

$$\text{logit}(y_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \delta_1 d_{i,1} + \delta_2 d_{i,2} + \delta_3 d_{i,3} + \delta_4 d_{i,4} + \delta_5 d_{i,5} \quad (5.7)$$

$$\text{logit}(y_k) = \beta_0 + \beta_1 x_{k,1} + \beta_2 x_{k,2} + \delta_1 d_{k,1} + \delta_2 d_{k,2} + \delta_3 d_{k,3} + \delta_4 d_{k,4} + \delta_5 d_{k,5} + \delta_6 d_{k,6} \quad (5.8)$$

I ekvationerna ovan representerar vänsterled *logit* av sannolikheten y_i (5.7) eller y_k (5.8). Högerled innehåller samtliga förklarande variabler som påverkar sannolikheten. I likhet med en linjär regression kan regressionskoefficienterna β_k och δ_k i (5.7) och (5.8) tolkas med avseende på den beroende variabeln. Det finns dock en viktig distinktion. I motsats till linjär regression avser β_k och δ_k förändringen i logit av sannolikheten y_i eller y_k när en av det förklarande variablerna förändras med en enhet och det andra hålls konstant (Rodríguez, 2007). Denna distinktion medför att det inte går att tolka koefficienterna på exakt samma sätt utan att göra ytterligare beräkningar. För att räkna ut det marginella effekterna av förklarande variabler kan man göra på olika sätt. Vi väljer att använda partiella derivator vilket också är den beräkning mjukvaran tillämpar i sin algoritm. Vi visade i (5.6) att sannolikheten kan omvandlas från logit till sannolikhet. Marginaleffekten av varje förklarande variabel $x_{i,j}$, $d_{i,j}$, $x_{k,l}$ och $d_{k,l}$ kan beräknas. Detta visas genom att implementera $x_{i,j}$ från (5.7) i följande ekvation som ett exempel.

$$\frac{\partial(y_i = 1)}{\partial x_{i,j}} = \frac{e^{\pi_{i,j}}}{(1 + e^{\pi_{i,j}})^2} \frac{\partial \pi_{i,j}}{\partial x_{i,j}} \quad (5.9)$$

Barron (2014) menar att förändringen av $x_{i,j}$ beror på alla andra oberoende variabler i modellen. Detta är i motsats till en linjär regression.

På liknande sätt som minsta-kvadrat metoden används för att analysera residualerna i en linjär regression används log sannolikhet, *log likelihood* för att bestämma parametrar i en logistisk regression (Menard, 2002). Log sannolikhet är värdet av att maximera (5.6) med metoden Maximum Likelihood, ML. En utförlig förklaring

av denna metod är bortom räckvidden för denna uppsats och därmed redovisas en enkel förklaring. Att erhålla en algebraisk lösning av (5.7) eller (5.8) genom maximering är svårt eftersom det finns många parametrar och dess sannolikhetsfördelning är icke-linjär. ML-metoden, icke-linjär optimering, bryter ner problemet numeriskt och löser det i delar där varje ny del utgörs av en iterativ process. Det viktigaste i algoritmen är att den lär sig göra bättre skattningar baserade på iterationen $(n-1)$ vilket leder till bättre resultat i modellen (Myung, 2003).

I en logistisk regression kan skillnaden mellan två log sannolikheter multiplicerat med -2 tolkas som ett χ^2-värde när ena modellen är ”nästad” i den andra (McCullagh, 2019). En modell är ”nästad” när ena modellen innehåller några men inte alla av de oberoende variablerna i den andra modellen samt innehåller inga av variablerna som inte inkluderas i den andra modellen (Menard, 2002). Tolkningen av χ^2 kan ses som ett test på signifikansnivå mot en nollhypotes.

H_0: $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = 0$

H_1: *En eller flera av koefficienterna är inte lika med noll.*

Om χ^2 är signifikant med 5 procent förkastar vi nollhypotesen och ger stöd åt mothypotesen att information om de oberoende variablerna låter oss göra bättre prediktioner av vår beroende variabel y_i (Menard 2000).

I en linjär regression mäts sambandets styrka med determinationskoefficienten, R^2. Där R^2 representerar hur stor del av den totala variationen för den beroende variabeln som förklaras av det linjära sambandet mellan variablerna (Körner & Wahlgren, 2006). Enligt litteraturen (Menard, 2002) är R_L^2[6] en av det mer robusta måtten för att undersöka hur väl en logistisk regressionsmodell passar data.

$$\mathrm{R}_L^2 = \frac{\chi^2}{D_0} \tag{5.10}$$

I (5.10) representerar D_0 Log-Likelihood-värdet för modellen med bara intercept

[6]Kallas även Pseudo-R^2 i statistiska litteraturen.

och utan förklarande variabler och χ^2 är värdet vi tidigare räknade ut (McFadden, 1974).

Slutligen är det motiverat att kommentera analys av residualtermerna i en logitmodell jämfört med linjär regression. I linjär regression tillämpas centrala gränsvärdessatsen för att inferera utifrån ett stickprov. Däremot måste man vara aktsam om standardavvikelserna är väldigt stora för ett litet stickprov då satsen bygger på en stor mängd datapunkter. Slutsatserna kan bli missvisande. Då utfallet i vår logistiska regression är binär följer residualerna en binomial fördelning. När $n \cdot \pi \cdot (1 - \pi) > 5$, kan binomialfördelningen approximeras med normalfördelningen.

$$\mathrm{Nf}(n \cdot \pi; \sqrt{n \cdot \pi(1 - \pi)}) \tag{5.11}$$

Där n är antalet händelser och π är sannolikheten för en händelse. Med detta i beaktande behöver vi inte vara lika försiktiga gällande ett litet urval. Resultatet behöver inte nödvändigtvis vara normalfördelat (Menard, 2002).

Multikollinearitet i logistisk regression innebär att en eller flera av det oberoende variablerna är korrelerade med varandra (Anderson, Sweeney, Williams, Camm & Cochran, 2016). Multikollinearitet utgör ett problem i en linjär regression då det kan vara svårt att utröna signifikansen av en oberoende variabels inverkan på den beroende variabeln. En tumregeln som tillämpas är att betrakta multikollinearitet som ett potentiellt problem om absolutbeloppet av korrelationen mellan två oberoende variabler är större än 0,7[7].

5.2 Procentuell förändring

Den metod analysen använder för att utvärdera den procentuella förändringen av att kategorisera om typer av skadeanmälningar från Svensk försäkring är något

[7]Korrelation är definierat för värden i intervallet $[-1,1]$. Absolutbeloppet täcker alla fallen.

Törnqvist, Vartia & Vartia kallar för procentuell förändring (1985).

$$\text{Förändring} = \frac{\Delta x}{x_i} \tag{5.12}$$

I (5.12) framgår metoden för att beskriva den procentuella förändringen av att kategorisera om skadeanmälningar. Δx beskriver skillnaden mellan det nya antalet skadeanmälningar och antalet skadeanmälningar för referensgruppen. I nämnaren på ekvation (5.12) återfinns beteckningen x_i vilket syftar till antalet skadeanmälningar för referensgruppen. Anledningen till att denna metod används i analysen är att vi vill kontrollera för att en ökad grad av täckning de facto inkluderar fler skadeanmälningspunkter. Med denna metod räknar vi ut den procentuella ökningen av skadeanmälningar vid att kategorisera om försäkringstyper. Sedan jämför vi med de marginella effekterna från föregående regressioners estimeringar.

6 Resultat

Denna studie utgår i att empiriskt undersöka informationsasymmetri på den svenska bilförsäkringsmarknaden. Denna fråga besvaras genom lämplig empiri och metod kopplad till korrelationen mellan täckning och risk.

6.1 Logit-modell

En logit-modell utvärderas utifrån, hur väl modellen passar datan samt statistiska tester av koefficienter (Peng, Lee & Ingersoll, 2002). Första steget vi gör är att utföra en logistisk regression med skadeanmälning som beroende variabel. Logistiska regressionen har följande oberoende variabler: årsmodell på bilen, premie, körsträcka, försäkringstyp samt ålder som kontinuerlig eller diskret variabel. Denna regression utfördes fyra gånger i olika syften. Första regressionen med ålder som kontinuerlig variabel inkluderade vagnförsäkring och delkaskoförsäkring exkluderades. Den andra regressionen med ålder som kontinuerlig variabel exkluderades vagnförsäkring och delkaskoförsäkring inkluderades. Sedan utfördes de två regressionerna med ålder som diskret variabel. Då, till och börja med, vagnförsäkring

inkluderad och delkaskoförsäkring exkluderad. Sedan med vagnförsäkring exkluderad och delkasko inkluderad. I samtliga regressioner är likelihood ratio av χ^2 signifikant ($p \leq 0{,}05$). Detta betyder att samtliga regressioner passar signifikant bättre än en modell med enbart intercept (se *tabell 6* & *7.*).

6.2 Ålder som kontinuerlig variabel

Resultatet av dessa regressioner presenteras i *tabell 6.* För att utvärdera hur väl vår logit-modell passar datan riktar vi fokus till modellens pseudo R^2-värde. Angivet värde på koefficienten speglar hur mycket variationen i modellen som minskar givet att samtliga oberoende variablerna inkluderas. (Nagelkerke m. fl., 1991). I *tabell 6* går det att se att båda regressionerna erhåller ett värde på 0,14 för pseudo R^2.

Genom χ^2 test med åtta frihetsgrader förkastar vi nollhypotesen (H_0) för båda regressioner som inkluderar ålder som kontinuerlig variabel. Detta innebär att vi inte kan förkasta mothypotesen (H_1). Alltså är någon eller några av de inkluderade variablerna signifikanta. För att utröna vilken eller vilka variabler som är signifikanta gör vi med grund i (5.9) beräkning av marginaleffekterna på våra oberoende variabler. Dessa resultat redovisas i *tabell 6.* Med signifikansnivån ($p \leq 0{,}05$) ser vi att årsmodell och försäkringstyp är signifikanta för bägge logit-modellerna. En intressant notering kring våra resultat är att ålder inte är en signifikant variabel för skadeanmälning i någon av regressionerna.

Marginella effekten av vår signifikanta variabel årsmodell är 1,5 procent i *test 1* och 1,4 procent i *test 2.* Alltså, när bilens ålder ökar med en enhet, givet att alla andra variabler hålls konstant, leder det till 1,4 procent förändring i sannolikheten för skadeanmälning, y_i. Vår andra signifikanta variabel, nivå av försäkring tolkas som följande: I första testet finns det ett samband mellan ökning av skadeanmälningar och att ha en försäkring som ersätter vid vagnskador. När en individ har vagnskadeförsäkring ökar y_i med 6,8 procent, givet allt annat konstant. I andra testet gällande delkasko var förändringen 9,1 procent för y_i. För samtliga ovanstående tester gäller resultaten den marginella effekten från baseline. Baseline i test 1 är de individer som har antingen trafikförsäkring, delkasko eller delkasko samt merförsäkring. I *test 2* representeras baseline enbart av individer med trafikförsäk-

ring. En ytterligare tolkning av resultaten görs om vi analyserar oddsförhållanden vilket representeras i *tabell 6* som odds ratio. Att ha en 1 år äldre bil medför 1,5 gånger större odds att bli ett skadeärende i *test 1* samt 1,4 i *test 2*.

6.3 Ålder som diskret variabel

När ålder kodas om från en kontinuerlig till diskret variabel förändras resultatet något. Samtliga resultat presenteras i en liknande tabell och finns att läsa i *tabell 7*. Precis som i testet med ålder som kontinuerlig variabel är χ^2 signifikant med åtta frihetsgrader. Alltså, förkastar vi nollhypotesen (H_0) för båda regressioner som inkluderar ålder som diskret variabel. Vi kan inte förkasta mothypotesen (H_1) vilket innebär att någon eller några av de variabler vi inkluderat i modellen är signifikanta. I både *test 1* och *test 2* är inte variabeln ung förare signifikant. Alltså kan vi inte påvisa med statistisk säkerhet att en person under 25 år är mer benägen att göra en skadeanmälan. Noterbart från dessa regressioner är att från *test 1* är en individ med en årlig körsträcka över 2500 mil signifikant ($p \leq 0{,}10$). En individ med en årlig körstäcka över 2500 mil utgör en 4,8 procent större sannolikhet för skadeanmälning. Detta innebär att de olika logit-modellerna påvisar olika resultat när man analyserar ålder som en kontinuerlig eller diskret variabel.

6.4 Multikollinearitet

I metoddelen beskrev vi en tumregel inom den statistiska litteraturen gällande multikollinearitet. Detta är ett potentiellt problem om värdet inte hamnar inom intervallet $i \in [-0{,}7, 0{,}7]$. Då multikollinearitet leder till missvisande slutsatser är en analys av multikollinearitet motiverat. Samtliga korrelationsmatriser till våra regressioner går att finna i bilagan till detta dokument. I regressionen med ålder som kontinuerlig respektive diskret variabel utgör ej variablernas korrelation med varandra ett problem då alla värden är inom intervallet $i \in [-0{,}7, 0{,}7]$.

6.5 Linearitet i logit-modellen

I vår metod har vi valt att utesluta test för linearitet i logit-modellen. Linearitet i logit-modellen innebär ett linjärt samband mellan logit-transformationen och

de oberoende variablerna (Menard, 2002) vilket är ett villkor i vår metod. Icke-parametrisk analys är komplicerat och relativt svårt att tyda. En Box-Tidwell-transformation[8] kräver avancerade kunskaper i numerisk analys och behandlas inte i denna uppsats.

[8]Potens-transformation som resulterar i en monoton funktion vilket möjliggör statistiska tester med avseende på linjäritet i logit-modellen. För en utförlig förklaring se Hosmer & Lemeshow (2013).

Tabell 6: Logistisk regression med ålder som kontinuerlig variabel.

	Test 1		Test 2	
	Vagnskada		Delkasko	
	Odds ratio	ME	Odds ratio	ME
Ålder	0,998	-0,001	0,001	0,001
	(0,003)	(0,001)	(0,003)	(0,001)
Årsmodell	1,149***	0,015***	1,138***	0,014***
	(0,009)	(0,001)	(0,009)	(0,001)
Premie	0,999	0,00	1,00	0,00
	(0,001)	(0,00)	(0,001)	(0,00)
Körsträcka				
1001–1500	1,158	0,016	1,117	0,012
	(0,113)	(0,010)	(0,109)	(0,105)
1501–2000	1,218	0,021	1,145	0,015
	(0,156)	(0,014)	(0,146)	(0,014)
2001–2500	1,110	0,011	1,024	0,002
	(0,237)	(0,023)	(0,217)	(0,022)
$>$ 2500	1,531	0,049	1,366	0,036
	(0,313)	(0,026)	(0,277)	(0,025)
Vagn	1,931***	0,068***		
	(0,202)	(0,010)		
Delkasko			3,068***	0,091***
			(0,739)	(0,025)
Konstant	0,000***		0,000***	
	(0,000)		(0,000)	
N	5 413	5 413	5 413	5 413
$LR\ \chi^2\ (8)$	627,75		614,62	
$Prob > \chi^2$	0,000		0,000	
Pseudo R^2	0,141		0,139	
Log Likelihood	-1 909,152		-1 915,721	

Not: Standardavvikelse inom parantes.

ME är den marginella effekten.

* $p < 0{,}1$; ** $p < 0{,}05$; *** $p < 0{,}01$.

Tabell 7: Logistisk regression med ålder som diskret variabel.

	Test 1		Test 2	
	Vagnskada		Delkasko	
	Odds ratio	ME	Odds ratio	ME
Ung förare	0,772	-0,026	0,606	-0,047
	(0,395)	(0,048)	(0,313)	(0,041)
Årsmodell	1,148***	0,015***	1,137***	0,014***
	(0,009)	(0,001)	(0,009)	(0,001)
Premie	0,999	0,00	1,000	0,000
	(0,000)	(0,000)	(0,000)	(0,00)
Körsträcka				
1001–1500	1,158	0,016	1,120	0,012
	(0,113)	(0,010)	(0,109)	(0,105)
1501–2000	1,217	0,021	1,143	0,014
	(0,156)	(0,014)	(0,146)	(0,014)
2001–2500	1,106	0,011	1,022	0,002
	(0,235)	(0,023)	(0,217)	(0,022)
$>$ 2500	1,516**	0,048*	1,362	0,035
	(0,310)	(0,026)	(0,277)	(0,025)
Vagn	1,896***	0,066***		
	(0,196)	(0,001)		
Delkasko			3,028***	0,091***
			(0,727)	(0,136)
Konstant	0,000***		0,000***	
	(0,000)		(0,000)	
N	5 413	5 413	5 413	5 413
LR χ^2 *(8)*	627,73		615,64	
Prob $> \chi^2$	0,000		0,000	
Pseudo R^2	0,141		0,139	
Log Likelihood	-1 909,163		-1 915,207	

Not: Standardavvikelse inom parantes.

ME är den marginella effekten.

* $p < 0{,}1$; ** $p < 0{,}05$; *** $p < 0{,}01$.

Tabell 8: Kategorisering av sammanfattande statistik.

Kategorisering	**Procentuell förändring**
Set 1	43,6
Set 2	756,7

Källa: Författarnas beräkningar av data från Svensk försäkring (2018).

6.6 Procentuell förändring

Resultatet av att tillämpa den procentuella förändringen på data från Svensk försäkring leder till att vi kan jämföra resultaten från logit-modellerna med en kontrollerande effekt. Applicerar vi (5.12) på data från Svensk försäkring resulterar det i följande värden (se *tabell 8*). Effekten av att kategorisera om skadeanmälningar till att innefatta fler punkter är vid *vagn* en 43,6 procentuell ökning av sannolikhet för skadeanmälning. För *delkasko* leder en liknande omkategorisering till 756,7 procentuell ökning. Skillnaden mellan andelen som inkluderar *vagn* och exkluderar *vagn* representerar vad vi förväntar att skadeanmälningar ska stiga för *set 1*. Resultatet från *set 2* är föga förvånande då andelen skadeanmälningar som hamnar under trafikens täckning är relativt låg.

7 Diskussion

Denna studie har sammanställt och använt tvärsnittsdata från Länsförsäkringar Uppsala samt hämtat data från Svensk försäkring. Genom fyra logistiska regressioner har korrelationen mellan täckning och risk uppskattats. Från dessa regressioner går det att analysera sannolikheten för en individ att göra en skadeanmälan beroende på försäkringens omfattning. Resultatet visade en positiv och signifikant effekt av korrelationen mellan täckning och risk. För att undkomma problemet att vi inte kan isolera typen av skadeanmälan på enskilda datapunkter kategoriserade vi försäkringstäckning i enlighet med två principer i fyra omgångar. Detta i avseende att också uppskatta effekten av att använda ålder som en kontinuerlig eller diskret variabel. Att olika försäkringstyper innefattar fler eller färre punkter för skadeanmälningar skapar problem när man uppskattar effekten av korrelationen mellan täckning och risk. Vid analys av den svenska bilförsäkringsmarknadens sta-

tistik, gällande skadeanmälningar och hur den kategoriseras, går det att uppskatta den procentuella förändringen av att utöka antalet möjliga skadeanmälningar för varje grad av försäkring. Tillsammans med estimeringar från de logistiska regressionerna, den marginella effekten av försäkringstyp som variabel samt procentuella förändringen finns en bedömningspunkt till vilken vi kan jämföra de två dataseten. Det är då möjligt att tillskriva effekten av hur graden av försäkringstyp förändrar sannolikheten av en skadeanmälning utöver att det råder en skillnad i täckning mellan olika försäkringstyper. Anledningen till detta är att vi vill uppskatta graden av informationsasymmetri som råder på den svenska bilförsäkringsmarknaden. Detta för att göra samma koppling som tidigare forskning gjort gällande informationsasymmetri och korrelationen mellan täckning och risk.

Denna studie har bidragit till forskningsfältet genom att kontrollera för effekten av att olika grad av täckning per definition innebär fler punkter för skadeanmälning. Kontrolleringen sker genom att kategorisera försäkringstyper i enlighet med vilka typer av skadeanmälningar som försäkringarna ger täckning för. Genom en förväntad ökning, hämtad från målpopulationen, förutspår metoden en marginell effekt försäkringstypen förväntas ha. Med logit-modellerna jämför vi sedan hur den faktiska effekten, en skillnad i försäkring, påverkar sannolikheten av en skadeanmälning. Resultatet av detta hämtar vi från föregående avsnitt och skillnaden är omfattande. I det första testet förutspår en förändring av försäkringstyp en skillnad på 43,6 procent. Estimeringen från logit-modellen med ålder som en kontinuerlig variabel redovisar 6,8 procent effekt för delkasko jämfört med baseline på delkasko. Estimeringen från logit-modellen med ålder som en diskret variabel redovisar 6,6 procent effekt för delkasko jämfört med baseline. I det andra testet förutspås en förändring på 756,7 procent. Estimeringarna från båda logit-modellerna med ålder som kontinuerlig respektive diskret variabel är 9,1 procent effekt för vagn jämfört med baseline. Vad detta säger är inte helt klart. Det går däremot att konstatera med statistisk säkerhet att en högre grad av försäkringstäckning leder till en högre sannolikhet för en skadeanmälning. Huruvida det förekommer informationsasymmetri på bilförsäkringsmarknaden är från tidigare forskning redan empiriskt undersökt. Till vilken grad denna informationsasymmetri existerar på materialet vi undersökt finner vi svårt att estimera. Detta eftersom den marginella effekten

och procentuella förändringen skiljer sig markant. Det är svårt att tillskriva den stora ökningen av sannolikhet av att en individ med högre täckning anmäler fler skador till att det skulle bero på informationsasymmetri mellan parterna. Trots att det är en av studiens prediktioner att sannolikheten av en skadeanmälning ökar vid ökad täckning, kunde vi inte empiriskt bevisa en tillräcklig ökning för att påvisa informationsasymmetri. Det vi kan konstatera med denna studien är dock att våra prediktioner stämmer. Huruvida detta bör anses som ett nymodigt eller ett självklart konstaterande låter vi vara osagt.

Begränsningen i denna uppsats är att vi inte har tillräckligt detaljerad data vilket möjliggör identifikation av individers riskprofil. Vår lösning på detta problem har beskrivits i metoden. Ytterligare en begränsning i vår uppsats är antalet variabler. Många tidigare studier har haft uppemot 30–40 variabler vilket bör leda till mer tillförlitliga resultat. Visserligen har flera av dessa variabler uteslutits på grund av att de inte är signifikanta. Inom denna uppsats gjorde vi ett antagande gällande individers nivå av riskaverson. Vi har antagit att individer har en konkav nyttofunktion och därmed är riskaversa. Det finns personer som är mer riskaversa och de som är riskälskande. Det kan motiveras att detta utgör en endogen variabel i vår analys och en begränsning av studien.

Utöver att en högre grad av försäkring de facto täcker fler anmälningspunkter, lyckades vi inte estimera en tillräckligt stor korrelation mellan täckning och risk för att påvisa informationasymmetri. Vidare forskning som saknar uppgifter om riskprofil bör vara medveten om vår avsikt att kontrollera för att högre täckning leder till fler anmälningspunkter. Vi föreslår också att framtida forskning bör fokusera på antagandet gällande riskaversion. En analys baserat på primärdata skulle kunna ge en insikt i enskilda individers uppfattning om sin egen risk. Förhoppningsvis leder detta till en bättre uppskattning av informationsasymmetri på bilförsäkringsmarknaden.

Källförteckning

Abbring, J. H., Chiappori, P.-A. & Pinquet, J. (2003). Moral hazard and dynamic insurance data. *Journal of the European Economic Association*, *1*(4), 767–820.

Abbring, J. H., Chiappori, P.-A. & Zavadil, T. (2008). Better safe than sorry? ex ante and ex post moral hazard in dynamic insurance data.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. I *Uncertainty in economics* (s. 235–251). Elsevier.

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D. & Cochran, J. J. (2016). *Statistics for business & economics*. Cengage Learning.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care (american economic review, 1963). I *Uncertain times* (s. 1–34). Duke University Press.

Autor, D. (2016, March). *Private information, adverse selection and market failure.* MIT. Microeconomic Thoery and Public Policy.

Barron, M. (2014, April). *Marginal effects in stata.* UCLA Berkeley.

Boyer, M. & Dionne, G. (1989). An empirical analysis of moral hazard and experience rating. *The Review of Economics and Statistics*, 128–134.

Bryman, A. (2016). *Social research methods*. Oxford university press.

Chen, Y.-H. & Jiang, B. (2019). Effects of monitoring technology on the insurance market. *Production and Operations Management*, *28*(8), 1957–1971.

Chiappori, P.-A. (2000). Econometric models of insurance under asymmetric information. I *Handbook of insurance* (s. 365–393). Springer.

Chiappori, P.-A. & Salanié, B. (2012). Asymmetric information in insurance markets: Empirical assessments. *Handbook of Insurance, Springer Verlag (Editor: Georges Dionne)*.

Cohen, A. & Siegelman, P. (2010). Testing for adverse selection in insurance markets. *Journal of Risk and insurance*, *77*(1), 39–84.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Dahlby, B. G. (1983). Adverse selection and statistical discrimination: An analysis of canadian automobile insurance. *Journal of Public Economics*, *20*(1), 121–

130.

Dahlby, B. G. (1992). Testing for asymmetric information in canadian automobile insurance. I *Contributions to insurance economics* (s. 423–443). Springer.

Desyllas, P. & Sako, M. (2013). Profiting from business model innovation: Evidence from pay-as-you-drive auto insurance. *Research Policy*, *42*(1), 101–116.

DiGangi, E. A. & Moore, M. K. (2012). *Research methods in human skeletal biology*. Academic Press.

Dionne, G., Gouriéroux, C. & Vanasse, C. (2001). Testing for evidence of adverse selection in the automobile insurance market: A comment. *Journal of Political Economy*, *109*(2), 444–453.

Dionne, G., Michaud, P.-C. & Dahchour, M. (2013). Separating moral hazard from adverse selection and learning in automobile insurance: longitudinal evidence from france. *Journal of the European Economic Association*, *11*(4), 897–917.

Einav, L. & Finkelstein, A. (2011). Selection in insurance markets: Theory and empirics in pictures. *Journal of Economic Perspectives*, *25*(1), 115–38.

Finkelstein, A. & McGarry, K. (2006). Multiple dimensions of private information: evidence from the long-term care insurance market. *American Economic Review*, *96*(4), 938–958.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424–438.

Gujarati, D. N. (2003). Basic econometrics. forth edition. *Singapura: McGraw-Hill*.

Heckman, J. J. & Borjas, G. J. (1980). Does unemployment cause future unemployment? definitions, questions and answers from a continuous time model of heterogeneity and state dependence. *Economica*, *47*(187), 247-283.

Holmström, B. (1979). Moral hazard and observability. *The Bell journal of economics*, 74–91.

Hosmer Jr, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied logistic regression* (vol. 398). John Wiley & Sons.

Israel, M. (2004). *Do we drive more safely when accidents are more expensive? identifying moral hazard from experience rating schemes* (forskningsrapport). CSIO Working Paper.

Körner, S. & Wahlgren, L. (2006). *Statistisk dataanalys* (vol. 4) (nr. 7). Studentlitteratur Lund.

Länsförsäkringar. (2021). *Om helförsäkring och halvförsäkring för bilen.* Hämtad från https://www.lansforsakringar.se/privat/forsakring/bilforsakring/om-helforsakring-och-halvforsakring/ (2021-05-11)

Länsförsäkringar Uppsala. (2021). *Utdrag från 2018 års kunddata.* (Erhållen via mejl: 2021-04-15)

Mas-Colell, A., Whinston, M. D., Green, J. R. m. fl. (1995). *Microeconomic theory* (vol. 1). Oxford university press New York.

McCullagh, P. (2019). Generalized linear models.

McFadden, D. (1974). The measurement of urban travel demand. *Journal of public economics*, *3*(4), 303–328.

Menard, S. (2002). *Applied logistic regression analysis* (vol. 106). Sage.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, *47*(1), 90–100.

Nagelkerke, N. J. m. fl. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.

Paefgen, J., Staake, T. & Fleisch, E. (2014). Multivariate exposure modeling of accident risk: Insights from pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice*, *61*, 27–40.

Pauly, M. V. (1968). The economics of moral hazard: comment. *The american economic review*, *58*(3), 531–537.

Pauly, M. V. (1978). Overinsurance and public provision of insurance: The roles of moral hazard and adverse selection. I *Uncertainty in economics* (s. 307–331). Elsevier.

Peng, C.-Y. J., Lee, K. L. & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, *96*(1), 3–14.

Puelz, R. & Snow, A. (1994). Evidence on adverse selection: Equilibrium signaling and cross-subsidization in the insurance market. *Journal of Political Economy*, *102*(2), 236–257.

Rodríguez, G. (2007, January). *Lecture notes on generalized linear models.* Princeton University.

Rothschild, M. & Stiglitz, J. (1978). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. I *Uncertainty in economics* (s. 257–280). Elsevier.

Rowell, D., Nghiem, S. & Connelly, L. B. (2017). Two tests for ex ante moral hazard in a market for automobile insurance. *Journal of Risk and Insurance*, *84*(4), 1103–1126.

SFS 2018:218. (2018). *Lag om kompletterande bestämmelser till eu:s dataskyddsförordning.* Hämtad från https://www.svenskforfattningssamling.se/doc/2018218.html (2021-05-11)

Shavell, S. (1979). On moral hazard and insurance. I *Foundations of insurance economics* (s. 280–301). Springer.

Svensk försäkring. (2018). *Försäkringsmarknaden.* Hämtad från https://statistik.svenskforsakring.se/SASVisualAnalyticsViewer/VisualAnalyticsViewer_guest.jsp?reportName=Marknadsstatistiken&reportPath=/SF/Extern&appSwitcherDisabled=true&reportViewOnly=true (2021-04-24)

Svensk försäkring. (2021). *Om oss.* Hämtad från https://www.svenskforsakring.se/om-oss (2021-04-24)

Törnqvist, L., Vartia, P. & Vartia, Y. O. (1985). How should relative changes be measured? *The American Statistician*, *39*(1), 43–46.

Trafikskadelag 1975:1410. (1975). *2 § trafikskadelag 1975:1410.* Hämtad från https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/trafikskadelag-19751410_sfs-1975-1410 (2021-05-13)

Transportstyrelsen. (2021). *Vägtrafikregistret.* Hämtad från https://www.transportstyrelsen.se/sv/vagtrafik/fordon/vagtrafikregistret/ (2021-05-25)

Varian, H. R. (2014). *Intermediate microeconomics with calculus: a modern approach.* WW Norton & Company.

Von Neumann, J. & Morgenstern, O. (2007). *Theory of games and economic behavior (commemorative edition).* Princeton university press.

Weisburd, S. (2015). Identifying moral hazard in car insurance contracts. *Review of Economics and Statistics*, *97*(2), 301–313.

Wilson, C. (1977). A model of insurance markets with incomplete information. *Journal of Economic theory*, *16*(2), 167–207.

Bilagor

Tabell 9: Korrelationsmatris för regression med vagn ålder som kontinuerlig variabel.

	Ålder	Årsmodell	Premie	1001–1500	1501–2000	2001–2500	> 2500	Vagn
Ålder	1,00							
Årsmodell	-0,07	1,00						
Premie	0,29	-0,27	1,00					
1001–1500	0,01	-0,16	-0,02	1,00				
1501–2000	0,00	-0,15	-0,08	0,45	1,00			
2001–2500	-0,03	-0,10	-0,09	-0,27	0,22	1,00		
> 2500	-0,06	-0,09	-0,13	0,28	0,23	0,15	1,00	
Vagn	-0,20	0,06	-0,45	0,00	0,02	0,03	0,07	1,00

Källa: Författarnas beräkningar av data från Länsförsäkringar Uppsala (2021).

Tabell 10: Korrelationsmatris för regression med delkasko och ålder som kontinuerlig variabel.

	Ålder	Årsmodell	Premie	1001–1500	1501–2000	2001–2500	> 2500	Delkasko
Ålder	1,00							
Årsmodell	-0,04	1,00						
Premie	0,25	-0,24	1,00					
1001–1500	-0,00	-0,15	-0,01	1,00				
1501–2000	0,02	-0,14	-0,06	0,45	1,00			
2001–2500	-0,01	-0,10	-0,07	0,27	0,22	1,00		
> 2500	-0,04	-0,10	-0,09	0,29	0,23	0,14	1,00	
Delkasko	-0,10	-0,16	-0,12	-0,05	-0,06	-0,03	-0,02	1,00

Källa: Författarnas beräkningar av data från Länsförsäkringar Uppsala (2021).

Tabell 11: Korrelationsmatris för regression med vagn och ålder som diskret variabel.

	Ung förare	Årsmodell	Premie	1001–1500	1501–2000	2001–2500	> 2500	Vagn
Ung förare	1,00							
Årsmodell	-0,04	1,00						
Premie	0,20	-0,27	1,00					
1001–1500	0,01	-0,17	-0,01	1,00				
1501–2000	0,02	-0,15	-0,09	0,45	1,00			
2001–2500	0,01	-0,10	-0,09	0,27	0,22	1,00		
> 2500	0,03	-0,10	-0,12	0,28	0,23	0,15	1,00	
Vagn	0,13	0,05	-0,44	0,00	0,02	0,03	0,06	1,00

Källa: Författarnas beräkningar av data från Länsförsäkringar Uppsala (2021).

Tabell 12: Korrelationsmatris för regression med delkasko och ålder som diskret variabel.

	Ung förare	Årsmodell	Premie	1001–1500	1501–2000	2001–2500	> 2500	Delkasko
Ung förare	1,00							
Årsmodell	0,03	1,00						
Premie	-0,19	-0,24	1,00					
1001–1500	-0,03	-0,15	0,00	1,00				
1501–2000	0,01	-0,14	-0,07	0,45	1,00			
2001–2500	0,01	-0,10	-0,07	0,27	0,22	1,00		
> 2500	0,02	-0,10	-0,09	0,28	0,23	0,14	1,00	
Delkasko	-0,05	-0,16	-0,11	-0,06	-0,06	-0,03	-0,02	1,00

Källa: Författarnas beräkningar av data från Länsförsäkringar Uppsala (2021).