NHH
INSTITUTT FOR SAMFUNNSØKONOMI
DEPARTMENT OF ECONOMICS

SAM 19 2010
ISSN: 0804-6824
AUGUST 2010

Discussion paper

Asymmetric information – evidence from the home insurance market

BY
KARL OVE AARBU

Norges Handelshøyskole
NORWEGIAN SCHOOL OF ECONOMICS AND BUSINESS ADMINISTRATION

Asymmetric information – evidence from the home insurance market

Karl Ove Aarbu*
Norwegian School of Economics and Business Administration (NHH)
July 2010

Abstract

In this paper I test whether asymmetric information is present in the home insurance market. To detect the existence of asymmetric information I apply the so-called positive correlation test to a dataset containing approximately 500 000 home insurance contracts gathered from a Norwegian insurer. In addition to the standard formulations of the positive correlation test I propose a method that encompasses joint modelling of frequency and severity. The results from these formulations show that frequency of claims increases in cover while claim costs are independent of cover. Asymmetric information may be driven by adverse selection or moral hazard and the empirical insurance contract literature has suggested different ways to disentangle these. I suggest two methods that can distinguish between these two possible explanations. The first method utilizes detailed claim information that allows me to separate out claims that most likely are driven by moral hazard. Second, I also conduct an instrumental variable regression that utilizes an exogenous reform that had an effect on the insurance price in this market. Both approaches indicate that adverse selection is the prime driver of the information problem. In a final step, I test whether risk aversion affects the results from the positive correlation test. Through detailed socio-economic information (SES) I construct proxies for risk aversion. These proxies turn out to be very important for understanding deductible choice but only marginally important for claim probability. The information problem increases when I control for risk aversion – indicating a small omitted variable bias in the positive correlation test.

* Thanks to Fred Schroyen, Frode Steen, Kjell Gunnar Salvanes, Ragnhild Balsvik, Øyvind Anti Nilsen, Gernot Doppehofer, Gregory Corcos, Eirik G. Kristiansen, Kjetil Gramstad and Miles Kimball for valuable suggestions. Moreover, I am deeply grateful to the company that provided the insurance data and to Statistics Norway who linked the insurance data to public administrative registers. A special thank to the different product specialists, actuaries and IT-specialists that explained, extracted and helped me with the understanding of the basic structure of the data. I will also thank the different persons in Statistics Norway who assembled the full data set. Moreover, I am very grateful to Mikis Stasinopoulos for providing me with the gamlss.rsm package in R and Roger Bivand for valuable comments and helping me to understand the basics of R.

1. Introduction

In the last decade there has been devoted considerable attention towards the empirical measurement of asymmetric information within different insurance markets. The vital first task in such an attempt is to measure the correlation between insurance coverage and ex-post claims. A positive correlation between cover and claims – conditional on the tariff variables that are used for setting the insurance premium - is an indication of asymmetric information. Several recent papers have investigated this property for different insurance markets as for example Chiappori and Salanié (2000), Chiappori, Jullien , Salanié, Salanié (2006), Finkelstein and Poterba (2004) and Cohen (2005).

The "positive correlation property" is the central outcome of the Rothschild and Stiglitz (1976) model where insurance buyers have private information about their risk type. The equilibrium result of this model is that ex-ante high risk types opt for full insurance cover, while low risk persons only seek partial cover. Thus, the outcome of this separating equilibrium is a positive correlation between demand for insurance cover and ex ante riskiness.

This "positive correlation property" can be tested empirically by measuring the correlation between insurance coverage and ex-post riskiness, conditional on all risk-characteristics used by the insurer. The reliance on ex-post claims to measure riskiness, however, opens up for that moral hazard rather than advserse selection is the prime cause for the information asymmetry (see for example Arnott and Stiglitz (1988) for a discussion of the moral hazard problem). Thus, the generic empirical test of positive correlation cannot identify the source of the information problem. Furthermore, despite convincing theoretical reasons for the existence of information asymmetries in insurance markets there have been surprisingly many studies that have retained the null hypothesis of zero correlation (see Cohen and Siegelman (2009) for a comprehensive list). Many authors suggest that one important reason for this result is that buyers differ considerably in their risk aversion. If heterogeneity in risk aversion is substantial and if risk aversion is negatively correlated with claims, the positive correlation test will be biased when risk aversion is unobserved.[1]

[1] See for example Cutler, Finkelstein and McGarry (2008), Einav, Finkelstein and Levin (2009) or Cohen and Siegelman (2009).

This paper provides a three-step empirical exploration of the asymmetric information problem, applying a dataset of approximately 500 000 home insurance contracts extracted from a Norwegian insurer.[2] In the first step I implement the positive correlation test by using three different methodological approaches, i) a reduced form regression, ii) a bivariate probit regression and iii) a model of total claim cost that simultaneously estimates the number and severity of claims. Both i) and ii) are, by now, standard approaches. However, to the best of my knowledge, the last method has not before been implemented within the empirical literature on asymmetric information. Irrespective of the method I find that the frequency of accidents is increasing in the coverage – a result that confirms selection. The model that provides a joint estimation of both the number of claims and the severity, however, shows that the expected severity is independent of the cover for "normal sized claims", which constitutes 98 percent of all claims filed. Thus, individuals who buy more insurance file more claims, but not higher claim sizes.

In a second step I ask whether the documented information asymmetry is caused by moral hazard or adverse selection. I suggest and implement two different methods in order to seek for the sources of the information asymmetry. First, I test whether the result of the positive correlation test is affected by removing claims that most likely can be attributed to moral hazard. Moreover, I also utilize an exogenous tariff reform in order to distinguish between adverse selection and moral hazard. Neither method indicates presence of moral hazard, which leads me to conclude that adverse selection is the primary explanation of the asymmetric information documented here.

One particular and unique feature with the data used here is the link between the insurance contracts and socio-economic characteristics (SES) that are obtained from public administrative registers. For each policyholder I have linked the major items of the annual tax return to the insurance contracts. That is, wage income, capital income, interest income, debt, gross wealth and wealth items as stock investment as well as information about education and main working sector. For the period 2004-2007 also household income items

[2] As to the best of my knowledge - the positive correlation test has not been applied to home insurance before. The only study on home insurance I am aware of is Justin Sydnor (2008) who collected insurance contracts for home insurance. He used this data set to investigate risk aversion but do not address the potential asymmetric information problem.

and other household characteristics are available.[3] The SES data allows me to test whether the presence of asymmetric information is affected by heterogeneity in risk aversion. Thus, in a third step I include risk aversion proxies and income splines in the bivariate version of the positive correlation test. Including these proxies increases the magnitude of the information problem marginally, which indicates that people with high risk aversion tend to choose low deductibles of other reasons than ex-ante riskiness, as suggested in recent literature. Interestingly and somewhat surprising, I find that income (irrespective of whether household income, personal income or gross wealth is used), conditional on the insurance tariff variables, does not have any effect on deductible choice or the claim probability.

The paper is organised as follows. The next section describes the positive correlation test and some of the caveats. Section 3 gives an overview of the relevant literature. Thereafter, in section 4, I give an overview of the market for home insurance in Norway. Next, the data set is described and descriptive characteristics are presented. Subsequently, in section 6, I give a more detailed explanation of the positive correlation test. Section 7 presents the benchmark results from the positive correlation test, where I employ three different formulations of the test:: a reduced form regression model, a biprobit model ,and a model of claim severity that provides a joint estimation of the number of claims and individual claim cost.. Section 8 addresses the question of whether it is moral hazard or adverse selection that drives the information asymmetry. In section 9 I test whether heterogeneity in risk aversion has an affect on selection into insurance contracts, an issue that is discussed at length in many recent articles. Section 10 concludes.

2. The positive correlation test and possible caveats

In the classic article of Rothschild and Stiglitz (1976) it is shown that the information problem leads to a separating market equilibrium where the contracts that are offered lead high-risk persons to choose full cover while low risk individuals select contracts with partial cover.[4] Rothschild and Stiglitz (1976) do not allow for moral hazard in their model. However, as discussed in numerous articles and analyzed in Arnott and Stiglitz (1988) - in addition to the selection problem the possibility of ex-post moral hazard always exist. After

[3] Several studies have linked aggregates of income, wealth and SES-information to insurance data by using either zip-codes, ward or municipality codes (Cohen and Einav 2007, Finkelstein and Poterba 2004, 2006).
[4] However, equilibrium may not exist at all.

the cover is bought the insured can pursue less attention, act with less caution or choose to involve in more risky activities compared to a case without insurance. Thus, it is a difficult task to empirically disentangle adverse selection from moral hazard because the observed outcome is the same.

The positive correlation test provides, however, an answer to whether a particular insurance market suffers from information asymmetries.[5] This test is in its generic form quite simple, although it requires very good and detailed data on insurance contracts. One needs data on claims, choice of deductibles and all the tariff variables that are used in the premium setting. Then, assuming that ex-post claims reflect ex-ante risk types - a researcher who documents that persons with a high claim probability select into contracts with low deductibles/low coinsurance, conditional on all the observable tariff variables, can infer that adverse selection is a problem in the market under scrutiny. Thus, if one finds this result there is at least one customer characteristic that is not taken into consideration in the insurance tariff and, moreover, positively correlated with the claim probability. Or, even in a case where all relevant risk factors are included – adverse selection can appear if one or several risk factors are priced incorrectly.[6]

The degree of market competition is, however, an issue that most of the empirical literature so far has touched little upon.[7] The classical Rothschild-Stiglitz model assumes full competition that drives premiums towards actuarial risk premiums. In such a setting, risk type is the only relevant screening characteristic as companies cannot offer a menu that, for example, screen customers with respect to differences in risk aversion. In reality various degrees of market power exist in almost all markets, which entails premium loading (may also be denoted as a mark-up) and opens up for selection along other dimensions than risk type, see Stiglitz (1977). For example, in a market where there is symmetric information

[5] Early implementations of this test can be found in Puelz and Snow (1994) who used car insurance data from US, Chiappori and Salanié (1996, 2000) who developed both parametric and non-parametric versions of this test.

[6] A special case of this possibility is discussed by Finkelstein and Poterba (2006). They document that insurance companies may abstain from using relevant observable information (they use the term "*unused* observable information"). An alternative interpretation is that insurance companies set the risk price to "zero" for some relevant observables, which effectively will lead to the same outcome as discussed by Finkelstein and Poterba. Therefore, it is not unlikely that companies can price some risk factors below their real risk price, because of imperfect information about the underlying risk or for political reasons. Thus, using the terminology of Finkelstein and Poterba, a risk price below the real risk price may be coined as "*underused* observable information".

[7] Chiappori, Jullien , Salanié, Salanié (2006)is, however, an exception.

about risk types but private information regarding risk aversion, companies will optimally offer menus in order to extract rent from differences in risk aversion. In such a case the positive correlation test will typically entail a zero result.[8] In the more extreme case of an insurance monopoly, Landsberger and Meilijson (1994) show that if agents that are equal along all dimensions expect for their risk aversion, the equilibrium result will entail full coverage for the most risk averse and partial coverage for the less risk averse. Again, the empirical counterpart of such a situation is a zero result from the positive correlation test. Jullien, Salanie and Salanie (2001) also consider a monopoly insurance model where policyholders differ in risk aversion and where moral hazard is present. Their model will under realistic circumstances also lead to an equilibrium result where there is no positive correlation between risk and coverage.

In perspective of the available theory, the main empirical problem that arises from this discussion is unobserved heterogeneity. That is, characteristics that companies can use to extract rent from their policyholders are often unobserved from the point of view of the econometrician. Therefore, these unobservable characteristics will weaken the power of the positive correlation test. One of the key components in the theoretical models discussed above is risk aversion and when this is unobserved in the data almost anything can happen to the empirical correlation between cover and riskiness. These considerations are important to have in mind when one evaluates the different results in this literature and the results shown in this paper. However, unlike most of the previous studies the data employed here contain a very rich set of administrative background information that allows me to address several aspects (but not all) of the unobserved heterogeneity problem. For example, I construct proxies for risk aversion and test whether these have a large impact on the benchmark results from the positive correlation test. I find that the results are robust towards this test.

[8] It may also very well be the case that highly risk averse people also are ex-ante low risk types or that they typically are less likely to engage in hazardous behaviour after the cover is obtained. Therefore, risk aversion can drive low risk customers to buy a high cover contract, which leads to advantageous instead of adverse selection.

3. Prior work on asymmetric information

The empirical literature that has applied the positive correlation test has produced diverse results. Puelz and Snow (1994) found evidence on asymmetric information in their dataset (US car insurance), but their result may stem from a misspecification of the empirical model, see Dionne, Gouriréoux and Vanasse (2001), Chiappori and Salanié (2000) and Chiappori (2000). Cawley and Philippson (1999) found no evidence of asymmetric information within the market for life insurance, employing data from a big US life insurer. Chiappori and Salanié (2000) cannot provide any evidence on asymmetric information in their French car insurance data, but Finkelstein and Poterba (2004) documents the existence of asymmetric information in the data they gathered from one large British annuity provider. Cohen (2005) does find clear evidence of asymmetric information in the dataset she collected from an Israeli car insurer, and so do Chiappori, Jullien , Salanié, Salanié (2006) using French car insurance data. Finkelstein and McGarry (2006) exploit data from the rather small US market of long term care insurance and find large heterogeneity regarding asymmetric information. In fact, they find evidence on both advantageous and adverse selection in their data – leading to the interesting result that these opposite selection forces cancels out in the aggregate, which in the next step leads to a negative result from the positive correlation test. This result, therefore, confirms some of the predictions in the theoretical literature discussed above. That also advantageous selection can exist in an insurance market is further documented by Fang, Keane and Silverman (2008).[9]

Interestingly, asymmetric information may also appear in a context where relevant observable characteristics are unused by the insurer. Finkelstein and Poterba (2006) find evidence on that insurers can choose not to utilize relevant information in their tariff setting. Using annuity data from a large UK life insurer they document that the insurance tariff they have access to can be improved upon by using zip-codes in the tariff setting. These zip-codes are observable for the insurer but still not used, thus the characteristic is an "unused observable". They attribute this situation to political concerns; an insurer may abstain from

[9] It is interesting to note that characteristics other than risk aversion may lead to advantageous selection.Fang, Keane and Silverman (2008) show that cognitive ability can explain the existence of advantageous selection in the US Medigap market. Their core story is that people with low cognitive ability fail to purchase Medigap insurance even though they are higher than average risk types.

using relevant information if it regards the variable as politically disputed and therefore may cause negative public relations if it is employed in the tariff setting. An analogous situation may occur when regulations prohibit insurers to use all relevant information, as the EU Gender Directive in 2004.[10]

The broad view of the literature so far is then that mixed results from the positive correlation test can appear of several reasons. One important explanation is heterogeneity in risk aversion – a characteristic that is typically not observed; another reason may stem from specific market characteristics; a third reason may be "unused observables"; and a fourth reason may be regulatory issues.

4. The home insurance market and comparisons of the sample with the population

The market I consider is both mature and quite advanced in terms of the risk classification. There are four major suppliers of home insurance in Norway and these four players covered altogether approximately 90 percent of the market over the period of investigation. The market statistics show, however, that the market fraction controlled by the biggest four companies declined steadily from 2002 to 2007 due to an influx of smaller companies. At the start of 2002 the market share of the big four was somewhat less than 95 percent. At the end of 2007 the market share was less than 90 percent.[11] This observation indicates increasing competition in the period under investigation. Another observation that reinforces the observation of a more competitive environment is that the number of suppliers in this particular market was 8 in 2002, 10 at the end of 2007 and 12 late in 2009.

The market for home insurance is important from an economic point of view. According to CEA (2009) page 9, property insurance (which includes home insurance) is the third largest non-life market after Motor- and Accident & Health insurance in Europe. However, the market for health insurance in Norway (and in the Nordic countries in general) is extremely thin - see for example Aarbu (2010). Thus, the market I consider here is the second largest non-life insurance line within the market for private insurance.

[10] Council directive 2004/113/EC

[11] The market share for the big four has also declined steadily in 2008 and 2009 and it is now at 84 percent. The market share numbers can be found at www.fnh.no/hoved/statistikk/kvartalsvise-statstikk-publikasjoner/kvartalsvise-statistikk-publikasjoner/premiestatistikk-skadeforsikring/, accessed Nov 24th 2009.

The supplier that has endowed me with the insurance data is among one of the biggest providers of non-life insurance and to the best of my knowledge it employs an advanced tariff that is comparable with the tariff of the other major competitors operating in this market. All suppliers of home insurance use tariffs that are dependent of a vast array of observable characteristics. These characteristics are typically collected over phone whenever a new customer signs for a contract. Moreover, some characteristics may be gathered through physical inspection of the home. Local surveyors typically conduct such inspections, but it is worth noting that such inspections are quite rare.

In general risk classification depends on object characteristics, subject characteristics and geography. First, a number of variables that describe the home are gathered. Typically the insurer will ask for the size of the house (in square metres), and other relevant home characteristics. The size of the house combined with an assessment of the standard of the house forms the base for the insurance value. Second, insurers also gather information about the insured, as the age, gender and other relevant socio-economic characteristics. Third, through address information an insurer may classify the object into risk zones.

One vital question is whether information from one specific insurer is sufficient in order to generalize up to a market level. I address this question along two dimensions: First, is the company that has provided me with data representative for the market? Second, are the customers of this company representative of the population? The first question is explored by conferring market statistics and market shares, while the second question is tackled by comparing various dimensions of the sample with similar population numbers (these comparisons were conducted by Statistics Norway).

The company offers housing insurance nationwide and all municipalities in Norway are represented in the data set. Moreover, the market share in this particular market during the investigation period was higher than 15 percent, but slightly declining from 2002 to 2007. However, the decline in the market share was within a normal range when we compare changes in market shares for this company with changes experienced by its competitors. To be more precise, the smaller players captured an increasingly higher share of the market at the expense of the four largest suppliers over the investigation period. Thus, both the large market share and the smooth changes in the market share from 2002 to 2007 indicate that the data analyzed here are likely to be representative for the market as a whole.

I have also compared important population measures with similar measures for the full sample (which is larger than the sample used in this paper). The details from this exercise are shown in appendix 1. There, it is shown that the customers of this company are comparable with the population along important income dimensions. The sample is slighty underweighted in the highest income deciles and the income distribution is more compressed in the sample. On the other hand, the distribution of specific income elements (as for example proprietary income) are almost a mirror image of the population distribution.

5. The data set – restrictions, definitions and descriptive statistics

I have collected all new insurance contracts that were sold between 2002 and 2007. Moreover, all renewals of these new contracts are included in the sample. This gives me a departing sample of 530 434 observations. Over this period I observe all details in the insurance tariff.

Several exclusions reduces the number of observations to around 492 000 observations. First, I remove observations with incomplete deductible information. Second, I remove observations that do not have complete age information and constrain the sample to include policyholders between 17 and 99 years of age. Third, I only include contracts that last either 366 or 365 days. Fourth, I observe less than 5 observations with more than 6 claims over the contract period.[12] These are also removed from the sample. Fifth, I remove observations where the insurance value of the house exceeds 10 million NOK. This exclusion removes approximately 100 contracts and secures that the data set contain standard home insurances. Finally, apartment houses with more than 20 rooms with water and with more than 5 floors are removed.[13]

It is important to underline that unit of analysis is the insurance contract. Typically, a customer buys one contract for the structure of the house and another contract for personal belongings (furniture and other movable property). People living in flats typically only need

[12] Experts in the company point to the possibility of error in registration of claims. This exclusion removes less than 10 observations.

[13] These restrictions are taken in order to secure that all contracts compared are derived from the same underlying tariff structure. For example, including homes with a higher insurance value than 10 million NOK increases the likelihood for manual underwriting, which I cannot observe in the data.

the latter insurance, while people that reside in their own home must buy both covers. This paper focuses solely on contracts for fixed property. [14]

The company that has provided the data registers all *events* that are reported by the customers. An event is not necessarily equivalent with a claim. An event will ultimately lead to three possible outcomes: a) The claim is legitimate and if the amount is larger than the deductible it will lead to a positive indemnity; b) the claim is legitimate but lower than the deductible which entails a zero indemnity; c) the claim is illegitimate and lead to zero indemnity. The two last claim types may be phrased as "zero claims", which I observe. Appendix 2 gives a detailed description of these claims and how these are recorded during the claim handling process. Based on the analysis given in Appendix 2 it is likely that the events reported to the company (irrespective of whether it leads to a positive indemnity or not) give an accurate picture of the underlying event risk. However, despite the evidence of a rather good environment for observing the "true" risk occurrence through registration of "zero claims", it is difficult to rule out that customers believe that the company is pursuing a penalising behaviour and therefore choose not to report an event - or that customers can assess the cost of the damage in an accurate way and therefore do not bother to report an event when the expected indemnity is lower than the deductible. I therefore apply the conservative decision of only including claims higher than the highest deductible in the empirical analysis, except for the reduced form benchmark regression where I present results both with and without "zero claims".

Another important empirical question is whether only first-year contracts should be included (as in Cohen (2005) and Cohen and Einav (2007))? The absence of a bonus-malus system in this market is an element that strengthens the case for including all contract years. Moreover, interviews with several company analysts and product specialists reveals that there is no automatic penalty at renewal if a customer reports an event, except if the customer is a

[14] Note that insurance cover can be changed in the contract period. The data I have access to, contain information about the contract that was valid at the end of the contract period. Thus, if a customer changes the contract in the middle of the contract year (for example the deductible or the insurance value), the change will occur in the data as it happened from the first day of that contract year. Moreover, the selection of contracts lasting 365 or 366 days is based on the start- and end-date for the insurance agreement (that normally includes several insurance cover types). This selection criterion may in some instances include contracts that have lasted less than 366 days because I do not observe whether the customer has cancelled the home insurance cover in – say – in the middle of the period. Interviews with company experts indicate, however, that the effect of these two "inaccuracies" on the results is probably quite low.

notorious claim reporter.[15] To check this further, however, I conducted a fixed effect regression (within regression) of the premium on all tariff variables and claims lagged one year. This regression (not reported) gave an insignificant value with a negative sign on the lagged claims parameter, which confirms that experience rating is not present in these data. Moreover, I checked whether the claim frequency is systematically declining in tenure. There is not much evidence of such a pattern.[16] The pragmatic solution I reach is therefore to show results for both the first year contracts and all contract years in most of the following tables. This is different from the majority of previous studies (see for example Cohen (2005) and Cohen and Einav (2007)) that predominantly focus the attention on the first year contracts.

Important descriptives

Table 1 shows the distribution of claims for the sample.[17] Table 1a includes all claims; i.e. zero claims are included, table 1b includes only claims that lead to a positive indemnity above the highest deductible (15 000 NOK) and table 1c show how claims are distributed across claim sizes.

[15] Note that a new customer is asked about the number of claims experienced the last two years. This variable is used in the tariff setting. The insurer is not allowed to verify this self-reported measure before the new customer eventually files a claim. Verification of the number of self-reported claims is possible because non-zero claims are registered in a national claim register. Thus, insurers are not allowed to check whether a new customer reveals past claim information truthfully at the time of the contract agreement. The possibility that the self-reported measure can be checked ex-post, though, will on the other hand give the customer an incentive to report the true number of claims.

[16] The probability of at least one claim above the highest deductible in the first contract year is 2.20 percent; it is 2.21 percent in the second; 2.19 in the third; 2.08 in the fourth; 2.06 in the fifth and it is 2.09 in the sixth year (the last I observe).

[17] Claims are counted for each contract within a contract period of 365 or 366 days. A customer can in principle possess more than one contract within the same contract period. A customer who for example owns two houses will possess two contracts. If an event causes damages on both houses simultaneously I will count one claim for each contract.

Table 1a. Number of claims on contracts signed 2002-2007. Including "zero claims"

Number of claims	Frequency	Percent
0	461201	93.74
1	28232	5.74
2	2334	0.47
3	222	0.05
4	23	0.00
5	4	0.00
6	1	0.00

Table 1b. Number of claims on contracts signed 2002-2007. Only claims>highest deductible.

Number of claims	Frequency	Percent
0	461,201	97.92
1	8,724	1.85
2	937	0.20
3	112	0.02
4	12	0.00
5	3	0.00

Table 1c. Claim size on contracts signed 2002-2007. Only claims>highest deductible.

Claim size in NOK	Frequency	Percent
<50 000	6,904	70.54
50 000-99 999	1,692	17.29
100 000- 149 999	441	4.51
150 000- 199 999	227	2.32
Over 200 000	524	5.35

The unconditional probability for at least one event is 6.2 percent for a one-year contract. The probability for two events during the period is approximately 0.5 percent. The probability for a claim that leads to a positive indemnity higher than 15 000 NOK, is much lower - around 2.2 percent. Compared to for example car insurance e.g. Cohen (2005) the event probability is only one eight of what is typical within the car insurance market (for claims above the highest deductible).[18]

[18] Almost 30 percent of the claims are connected to general building damages. Around 17 percent is caused by breakage of water pipes inside the building, and a little less than 13 percent is connected to similar breakages outside the building. Less than 10 percent is caused by fire, while the remainder of claims are spread over a broad menu of causes as theft, explosion etc.

Table 2 gives the menu of deductibles and the percent of contracts within each deductible group.

Table 2. The deductible menu for the period 2002-2007 and the distribution of deductible choice.

	Deductible in NOK	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	2000	3447	0.70	3447	0.70
	2500	10222	2.08	13669	2.78
	3000	44950	9.14	58619	11.91
	3500	1088	0.22	59707	12.14
Default deductible: 4000		83340	16.94	143047	29.07
	5000	192185	39.06	335232	68.13
	6000	121759	24.75	456991	92.88
	9000	30563	6.21	487554	99.09
	15000	4463	0.91	492017	100.00

Note first that the deductible menu is quite broad with 9 deductible categories ranging from NOK 2000 (approximately $300) to NOK 15 000 (approximately $2500).[19] The default deductible (NOK 4000) serves as an anchor for the premium calculation (see Appendix 3 for a general sketch of the relation between the premium and the deductible). At the default the deductible has a neutral impact on the premium. At higher deductibles the premium is lower and at lower deductibles the premium is higher.

The most common deductible is 5000 NOK, actually 1000 NOK higher than the default deductible. The second most common choice is 6000. Company executives explain one possible reason for this pattern is that their sellers was encouraged to sell on "high deductibles" and that selling "high deductibles" was more prevalent later in the period. I can confirm this in the data: If we restrict table 2 to only contain observations from 2002 and 2003 the modal value is 5000 (38.9 percent of the contracts) and the second most common value is 4000 (21.3 percent of the contracts), while 6000 was chosen by 10.8 percent of the sample. Another reason is inflation. Because the deductible menu was fixed in nominal terms in the investigation period the real value (or the purchasing power) of any deductible declined, which may lead policyholders to choose higher deductibles over time.

[19] The exchange rate in late April 2010 was 5.94 NOK per $, see www.valutakurser.no. Accessed April 20th 2010.

To get a first impression of the association between the event frequency and the choice of deductible table 3 shows the unconditional association between the deductible choice (first column) and reported events.

Table 3. Claim frequency within each deductible level. All claims (including "zero claims")

Deductible amount	Frequency all claims (percent)
2000	0.073
2500	0.077
3000	0.066
3500	0.068
4000	0.066
5000	0.069
6000	0.070
9000	0.062
15000	0.049

The numbers show that the frequency is highest at the low end and lowest at the high end of the deductible menu. However, the pattern is not monotonic, the frequency of claims between deductibles from 3000 to 6000 is almost constant with a floor of approximately 6.15 and a ceiling on about 6.35.[20] Thus, within this particular region it is not easy to see any clear association between frequency and deductible choice.

Tariff variables and important changes in the period from 2002-2007

The number of tariff variables that are used is between 40 and 50 and both to confidentiality reasons and the mere number of these covariates I will describe and report only on a subset of these. In table 4 this subset is shown along with their descriptive statistics.

[20] Note that the deductible level of 3500 is rarely used and it contain only 75 claims. The claim frequency of 6.89 in is therefore quite uncertain.

Table 4. Descriptive Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
Age	492017	48.21	13.35	17.00	99.00
Woman	492017	0.300	0.458	0	1.000
Non smoking household	492017	0.734	0.442	0	1.000
Fire alarm	492017	0.156	0.363	0	1.000
Water alarm	492017	0.0087	0.0927	0	1.000
Theft alarm	492017	0.170	0.376	0	1.000
Household size between 2 to 5	492017	0.776	0.417	0	1.000
Filed claims last two years (self reported)	492017	0.0236	0.173	0	9.000
Insurance value	492017	2.38E6	899629	16660	9.96E6
Number of floors	492017	1.590	0.730	0	5.000
Number of rooms with water tap	492017	3.908	1.235	0	19.00
Number of living units rented out	492017	0.112	0.362	0	5.000

The mean age of the policyholder is close to 50. Younger persons typically rent or own flats situated in larger apartment houses. Apartment houses are typically insured collectively and the occupier will normally only need to buy insurance for the movable belongings and these insurances are not considered here.

One interesting tariff parameter is whether the household is non-smoking or not. Table 4 shows that the fraction of non-smoking households is around 74 percent. According to Statistics Norway the population fraction of smokers was around 25 percent in 2004, quite close to what we observe in the data. A little less of 80 percent of the households are between 2 to 5 members, 16 percent possess fire alarm and 17 percent have theft alarm. Moreover, on average a house has 1.6 floors and close to 4 rooms with water connection and 1 out of 10 households offer units for let. The average insurance value in the period is around 2.4 million NOK.

The data I use in the benchmark analysis cover altogether 6 years. The main tariff structure was unchanged over this period, but some smaller changes were implemented. The table below give a short overview of when certain tariff changes took place.

Jan 1st 2002 to 31st Dec 2003	The tariff was unchanged
Jan 1st 2004	Gender was removed as tariff variable due to a decision by the Anti Discrimination Ombud
Apr 1st 2004	Inclusion of a new tariff variable
May 1st 2005	Inclusion of three new tariff variables

Note that the changes, apart from one, were initiated by the company itself. The one exogenous change was a consequence of a decision of the "Equality and Anti-discrimination

Ombud" (EAU). EAU prohibited the use of gender as a risk classification measure. The prohibition decision was taken May 2nd 2002 and insurance companies were given a maximum of two years to remove this tariff factor. In our case the removal of gender as a tariff factor was set into effect from January 1st 2004. [21] It is important to stress that the full effect of this regulatory change (if any) would appear throughout 2004 and into the last day of 2005, because the last renewal after the new tariff was implemented happened December 31st 2004.[22]

To assess the effect of the reform on premiums I have calculated mean premium changes for the same policyholders in 2003 and in 2004 (within changes), respectively. These are shown in the table below

	Men		Women		
	Mean change in NOK	N	Mean change in NOK	N	Relative premium increase for men compared to women
2003	536	15888	527	6988	1,02
2004	413	27783	353	12501	1,17
Average premium 2002					NOK 2442

Note first that the premium changes in 2003 and 2004 were large - reflecting a need for higher profitability within this particular insurance line. The absolute price change in 2003 was approximately the same for both genders. However, in 2004 the price increased relatively more for men compared to women, approximately 17 %. This is exactly what we should expect as the company after the reform needed to use a premium target level averaged over gender. To achieve this - insurance premiums for men had to increase more compared to women, which is confirmed in the table.

6. Implementing the positive correlation test

Estimating positive correlation can be analytically solved in several ways. Following Finkelstein and Poterba (2006) the basic idea behind this test can be illustrated with two simple equations. Let D denote the deductible and A the number of accidents/events.

[21] See www.ldo.no/no/klagesaker/arkiv/likestillingsombudets-klagesaker/2002/mars/kjonn/kjonn-som-faktor-ved-beregning-av-forsikringspremie-/, accessed Nov 24th 2009

[22] Based on interviews with company employees the mandatory change was implemented in a way that preserved expected profits. This implied higher premiums for men and lower for women.

Moreover, let X represent the variables used by the insurance company when setting premiums. Using the information contained in X one can then estimate:[23]

$$
(1) \qquad \begin{aligned} A &= X'\beta + \varepsilon \\ D &= X'\gamma + \upsilon \end{aligned} ,
$$

where ε and υ are residuals from the accident equation and the deductible equation respectively. Assume now that the vector X contains all relevant risk characteristics. Moreover, assume that that the insurer utilizes the full X-vector in the tariff setting and is able to price each risk factor correct. Lastly, assume that risk aversion and possible other types of preference based selection mechanisms play no role in the market under scrutiny. If these assumptions are fulfilled - the residuals - ε and υ will be uncorrelated.

If there, however, is one variable that has a positive impact on *A,* but not included in X the residual ε will contain the effect this variable has on A. A customer who is described by this variable will then normally have more accidents than the insurance company predicts. The company does not price this private information and the customer will typically self-select into a lower deductible (more cover). In such a case a high ε will typically work together with small υ and the correlation between the residuals will be negative. Thus, when the cover is measured through the deductible (higher deductible means less cover) a significant *negative* correlation between the residuals ε and υ indicates presence of asymmetric information in the market.[24]

If we take equation (1) directly to the data it lends itself to a biprobit specification, where A is 1 if the policyholder reports at least one claim and D is 1 if the policyholder has chosen the high deductible. However, under a conditional independence assumption (see footnote 22) it can also be estimated by a count model or with an OLS- model where A is the dependent variable and D is the independent variable (or the other way around). Most of the empirical literature has used the number of accidents (in a count model/OLS) or an

---

[23] For convenience I suppress subscripts for individual *j* and time *t*.
[24] An alternative explanation is provided by Dionne, Gourièoux and Vanasse (2001). They depart from the notion of "conditional independence". With the notation given in (1) conditional *independence* between claims and deductible choice can be written as $g(A \mid X, D) = g(A \mid X)$, where *g* is the function mapping from the independent variables to the claim probability. This means that the deductible choice does not contain any relevant information as soon as all the exogenous X – variables are taken into account. If it does, however, conditional independence should be rejected.

occurrence of a claim (in the biprobit) as the dependent accident variable, see for example Cohen (2005). The severity of a claim is to the best of my knowledge, seldom used within this context. One reason for this is that frequency is probably a better behavioural measure than the severity of the claim. Each claim – large or small – is triggered by an adverse event. Whether the claim grows large or stays small is often a function of luck and other exogenous events surrounding the event.[25] Moreover, if average claims are independent of the insurance cover (or even increases in the insurance cover) the frequency is a sufficient statistic in the sense that higher frequency means higher total losses for the company, which implies that the presence of asymmetric information (if any) has real economic impact. If however, a higher number of claims implies lower average claim sizes the link between occurrences and economic losses is not so clear anymore, as the total loss is the product of the number of claims times the average severity. In implementing the positive correlation test this paper will depart from the approaches taken in the prior literature, which means that the frequency of accidents is the dependent variable. However, I also implement the positive correlation test within a methodological framework developed by Heller, Stasinopoulos, Rigby and de Jong (2007) (referred to hereafter as HSRJ). Their method opens up for a joint modelling of the frequency and claim severity. To the best of my knowledge this approach reflects the state of the art in actuarial ratemaking, without having been used in the empirical literature on asymmetric information.

7. A positive correlation test for the home insurance market

This paragraph report results from three different variants of the positive correlation test. First, I estimate a reduced form model that involves regressing the number of claims on all tariff variables and the deductible menu (Cohen 2005, Finkelstein and McGarry 2006). Second, I specify a bivariate probit model that simultaneously estimates the probability for filing claims and the choice of deductible. This specification allows for an explicit estimate of the correlation between the equations given in equation (1), see (Chiappori and Salanié 2000, Cohen 2005, Finkelstein and McGarry 2006). It also provides us with the partial association

[25] One cannot exclude that the event grows large due to lack of action or even "too much action". For example, a policyholder can actively prevent the event to grow by doing the correct tasks. On the other hand the event can grow large due to passive behavior or by that the customer is acting in an erroneous way in an attempt to reduce the consequences of the event. Thus both "moral hazard" caused by passive behavior or lack of knowledge may cause an event to grow towards a large catastrophe.

between claim frequency, deductible choice and the risk characteristics. Third, in the last subsection I present a model for total claim size that jointly model the count process for the number of claims and the claim size process given a claim.

a) Reduced form regressions

The benchmark regression is specified as follows

(2) $$A_{it} = X_{it}'\beta + \eta D_{it} + u_{it},$$

where A is the number of claims, X is a vector that contains all tariff variables, D is the deductible choice and u_{it} is the residual.[26] The primary interest lies in the η parameter. If the tariff is specified in a way so that all relevant private information is revealed during the underwriting process and each risk factor is priced correctly there should be no incentive for the customer to strategically choose a deductible that fits her/his specific risk. If this is the case then η should be zero. If the tariff does not incorporate relevant information in a correct way, however, then one should expect that high-risk customers (measured by ex-post claims) strategically choose more cover i.e. opt for a lower deductible. An information asymmetry in favour of the customer will therefore result in a negative η. As should be clear from the discussion in section 2 one cannot rule out that η is greater than zero, which may indicate that the presence of some market power and heterogeneity in risk aversion allows the company to extract rents from customers.

Table 5 show regression results for different specifications of the benchmark model. The dependent variable is number of claims. The independent variables include all tariff variables and the deductible choice. I also control for regional risk classification by including

[26] The type of claim may tell something about the underlying cause of the information problem. The data available here contain all types of claims that belong under the umbrella "home insurance". Thus, broken water pipes, fire, theft and structural damages are included in the data. To test presence of asymmetric information I include all claims. However, in order to disentangle adverse selection from moral hazard it is possible to argue that some claim types must be more likely under a moral hazard story and other types may fit better into an adverse selection story. I will explore this opportunity later in the paper. Note also that claims caused by natural perils are covered by a separate insurance carrier "Norsk Naturskadepool" and these are not counted in the data employed here. This carrier is administered by insurers that are members of the Financial Services Organization.

a full set of county dummies. Norway has 20 counties, which leads to an inclusion of 19 county dummies.[27]. To take account of the gender reform that was discussed in section 5 I specify first a dummy variable that takes the value 1 after Jan 1st 2004 and zero before that date. This variable then, measures the average change (if any) between these two periods. Moreover, I include an interaction between the reform indicator and whether the customer is a woman. This variable takes out the separate effect for women (if any) the reform had on women's behaviour. The other tariff changes are also indexed with dummy variables in a similar way. Thus, an interaction between the variable and time is specified in order to control for the inclusion effect. Lastly, year dummies are included in all regressions.

[27] In unreported regressions I have used a much finer division by including municipality dummies. There are 430 municipalities in Norway and all municipalities are represented in the data. Moreover, I have also used zip-codes which imply an even finer division of the sample than obtained that municipality codes. However, the results are maintained and I therefore choose to report results from the regression that include a full set of county dummies.

Table 5. OLS and Negative Binominal regressions. Dependent variable is Number of Claims. All contract years.

	OLS – Including zero claims	NB – Including zero claims	OLS - Claims> highest deductible	NB - Claims> highest deductible
	b/se	b/se	b/se	b/se
Age	-0.000610***	-0.009392***	-0.000197***	-0.008730***
	(0.000033)	(0.000511)	(0.000020)	(0.000910)
Gender (Women=1)	0.017161***	0.229160***	0.005293***	0.213720***
	(0.002081)	(0.026520)	(0.001267)	(0.049123)
Non smoking household	-0.003914***	-0.057266***	-0.002056***	-0.088295***
	(0.000954)	(0.013809)	(0.000584)	(0.024511)
Fire alarm	0.006222***	0.084432***	0.000965	0.040609
	(0.001537)	(0.020319)	(0.000941)	(0.037218)
Water alarm	0.004947	0.058638	0.004541	0.156693
	(0.004922)	(0.059983)	(0.003375)	(0.108971)
Theft alarm	0.001952	0.021853	0.000426	0.023500
	(0.001664)	(0.022427)	(0.001017)	(0.040320)
Between 2 to five members of household	0.003158*	0.058703**	-0.000357	0.005670
	(0.001233)	(0.018946)	(0.000767)	(0.033978)
Number of floors	0.004377***	0.060007***	0.001252**	0.053250***
	(0.000623)	(0.008349)	(0.000384)	(0.015111)
Number of rooms with water tap	0.002213***	0.032574***	0.000694**	0.032437***
	(0.000416)	(0.005530)	(0.000257)	(0.009757)
Insurance value (1000 NOK)	0.000012***	0.000164***	0.000007***	0.000249***
	(0.000001)	(0.000008)	(0.000000)	(0.000014)
Claims last two years (self reported)	0.020007***	0.221534***	0.007318***	0.239162***
	(0.003051)	(0.028699)	(0.001844)	(0.048440)
Number of apartments rented out	0.019348***	0.200148***	0.009133***	0.255749***
	(0.001499)	(0.014693)	(0.001006)	(0.025697)
Reform Year (2004)	-0.007970**	-0.130979**	-0.001910	-0.077141
	(0.002695)	(0.048253)	(0.001664)	(0.071637)
Reform Year * Woman	-0.003143	-0.031353	-0.002027	-0.080131
	(0.002280)	(0.029538)	(0.001400)	(0.055061)
Deductible (1000 NOK)	-0.002446***	-0.037930***	-0.000584***	-0.025543***
	(0.000244)	(0.003947)	(0.000151)	(0.006637)
Constant	-0.025227	-9.699007***	0.152565	-4.957878
	(0.173297)	(2.673163)	(0.128176)	(4.608007)
N	4.92017e+05	4.92017e+05	4.70989e+05	4.70989e+05
r2/ll	0.006756	-1.23422e+05	0.003245	-5.07110e+04
Overdisperson parameter (alpha)		0.399825***		2.145963***
Year dummies	Yes	Yes	Yes	Yes
County dummies	Yes	Yes	Yes	Yes
All tariff variables included	Yes	Yes	Yes	Yes

* for p<.05, ** for p<.01, and *** for p<.001. Standard error clustered within customers.

The first column displays parameter estimates from an OLS regression while the second column shows similar results from a negative binominal regression. The negative binominal regression is especially well suited to these data. As seen from table 1 the data contains excess number of zeroes and there is therefore reason to believe that overdispersion is present. Moreover, as pointed out by Greene (2000) the negative binominal model can be shown to arise from an explicit formulation of unobserved (cross-section) heterogeneity. In the Poisson model (from which the negative binominal model departs) the mean and the variance is equal and equals the expected number of events per period. The negative binominal model is a generalized Poisson model where one introduces an individual unobserved effect in the conditional mean function. To see this, suppose that the arrival rate per period λ is random instead of being a deterministic function of the explanatory variables. Following Cameron and Trivedi (2005) one can enter the unobserved heterogeneity in a multiplicative way such that $\lambda = v\mu$, where μ is the deterministic part explained by the regressors and v is a random parameter. By assuming that v is gamma distributed one can obtain a closed form expression for the arrival process (the negative binominal distribution) that can be estimated using standard maximum likelihood estimation.

The results shown in table 5 reveal a consistent pattern across the different specifications and restrictions. Independent of specification the deductible choice has a small but highly statistically significant effect on the number of claims. The sign of the coefficient is negative and it tells us that the higher the deductible the lower the probability for a claim. The statistical significance is especially high in the specification that includes all reported events (i.e. both positive- and zero claims).

Age and gender are both significant predictors for the incidence of claims. Claims are more prevalent among young customers and women have a higher claim frequency than men.[28] Non-smoking households file fewer claims. Customers who are investing in alarms have typically higher event frequency, but this tendency is not reflected for claims higher than the highest deductible. However, it is still interesting to note that alarms apparently do not entail lower claim frequency. Why? One possible reason is that alarm systems are more

[28] Within home insurance (and also car insurance) there is not necessarily a unique link between the policyholder and the accident, apart from one-person households. Thus, it is somewhat surprising and interesting that gender turns out significant. I have also tried to include dummies for household type (this information is available from 2004 and onwards) but this do not change the result.

prevalent in neighbourhoods that have an elevated risk for theft or vandalism. Typically, however, it is common practice among all insurance companies to reduce the premium if such an alarm is installed (as most of the customers who do so expect that insurance premiums should be lowered). From the view of the insurance company this is fine if the company can set average risk premiums correctly at the neighbourhood level. Within each neighbourhood it is reasonable to think that homes with alarms are less exposed for the risk and therefore premiums should be lower compared to the neighbourhood average risk. On the other hand, if the insurance company is unable to estimate the average neighbourhood risk correctly, this discount practice may lead to mispricing of risk. Such a scenario is not unlikely. Even with a risk classification that involves zip-codes a company may not capture the full heterogeneity in risk within the zip-code area.Column 3 and 4 report results for regressions where zero claims are excluded. Only claims larger than the highest observed deductible is retained in the sample. This restriction removes the potential bias that may stem from underreporting of claims from customers with high deductibles. The overall results are maintained in regression 3 and 4. However, the magnitude of the informational asymmetry is lower, but the coefficient is still highly statistically significant at a 1 percent level. The estimate of alpha in the negative binominal – that measurers the degree of overdispersion is highly significant.

b) A bivariate probit specification

In this section I present results from a bivariate specification. One advantage of the bivariate specification is that it gives separate estimates on the association between the choice of deductibles and the insurance tariff characteristics. The equation that is estimated is

$$(3) \qquad \begin{aligned} a_{it} &= X_{it}'\beta + v_{it} \\ d_{it} &= X_{it}'\mu + u_{it} \end{aligned},$$

where a_{it} =1 if contract *i* experiences at least one accident during the policy period t and zero otherwise and d_{it} =1 if the deductible is higher than 6000 NOK and zero otherwise.

Table 6 shows the results from the model specification where only claims higher than the highest deductible are included – a restriction I will maintain throughout rest of the paper. The first column shows the parameter estimates for the claim equation, while the second column provides estimates for the deductible equation. The third and the fourth column show the same regression for the first contract year only. As I briefly mentioned above, Cohen (2005) and Cohen and Einav (2007) only use the first contract year in their analysis. They argue that including several years of contract history for each customer opens up for endogenous selection.[29] I argued, however, that endogenous selection may be a less of a problem here because there is no automatic screening of customers at the end of the contract year in order to sort out contracts with poor claim history. Therefore, I display results for all contract years as well as first-year contracts.

[29] This means that the insurance company assess each customer at the end of the year and may offer customers with poor claim history systematically inferior contract terms than customers with good claim history.

Table 6. Biprobit. Only claims higher than the highest deductible included.

	All contract years		First contract year	
	High deductible (>6000)	Have at least one claim	High deductible (>6000)	Have at least one claim
	b/se	b/se	b/se	b/se
Age	-0.007592***	-0.003118***	-0.007722***	-0.003324***
	(0.000480)	(0.000352)	(0.000471)	(0.000584)
Gender (Women=1)	-0.143243***	0.089750***	-0.128900***	0.100775***
	(0.025131)	(0.019838)	(0.024615)	(0.023851)
Non smoking household	-0.051558***	-0.039081***	-0.053328***	-0.038826*
	(0.012441)	(0.009801)	(0.012234)	(0.015899)
Fire alarm	0.058544***	0.014685	0.058315***	0.001342
	(0.017137)	(0.015024)	(0.017699)	(0.024953)
Water alarm	-0.001457	0.031529	0.052240	0.105521
	(0.048508)	(0.043956)	(0.049289)	(0.069767)
Theft alarm	-0.049513***	0.011103	-0.075049***	-0.003783
	(0.014926)	(0.016421)	(0.018668)	(0.028113)
Between 2 to five members of household	-0.058440***	0.007896	-0.063310***	-0.004018
	(0.016371)	(0.013378)	(0.016199)	(0.021887)
Number of floors	-0.041489***	0.018931**	-0.043659***	0.027713**
	(0.007909)	(0.006048)	(0.007832)	(0.010088)
Number of rooms with water tap	0.006825	0.011258**	-0.000325	0.015413*
	(0.005080)	(0.003935)	(0.004922)	(0.006532)
Insurance value	0.000219***	0.000101***	0.000210***	0.000088***
	(0.000007)	(0.000006)	(0.000007)	(0.000010)
Claims last two years (self reported)	-0.048277	0.093256***	-0.006447	0.084467**
	(0.031659)	(0.020308)	(0.036383)	(0.032233)
Number of apartments rented out	0.058554***	0.097226***	0.066703***	0.093910***
	(0.014053)	(0.010731)	(0.013795)	(0.017409)
Reform Year (2004)	0.392445***	-0.031761	0.511152***	-0.051690
	(0.023389)	(0.028489)	(0.034357)	(0.047520)
Reform Year * Woman	-0.068319**	-0.037435	-0.072372*	-0.032288
	(0.022656)	(0.022116)	(0.028383)	(0.030801)
Constant	-6.145759*	-0.605538	-2.375093	1.530689
	(2.952550)	(1.676489)	(2.389232)	(2.671933)
Loglik	-1.57039e+05		-5.92719e+04	
N	4.70989e+05		1.68279e+05	
Rho	-0.033469***		-0.034684*	
	(0.008523)		(0.013614)	
County dummies	Yes		Yes	
Year dummies	Yes		Yes	
All tariff variables included	Yes		Yes	

* for p<.05, ** for p<.01, and *** for p<.001. Standard error clustered within customers.

The main result that was found in the benchmark specification is confirmed by the results in table 6. The correlation coefficient is consistently negative and similar in size irrespective of whether one includes all contracts or just the first year contract. This result indicates presence of an information asymmetry.

However, the biprobit also adds some new and interesting information by separating out the effect of tariff variables on the claim frequency and the deductible choice, respectively.

Three clear and consistent patterns stand out. First, age is associated with lower claim frequency but at the same time older people choose lower deductibles. Second, women do have higher claim frequencies and choose lower deductibles. Third, non-smoking households file significantly less claims and choose lower deductibles. In addition the results indicate that various alarm systems do not reduce the frequency of claims.

The variables that aim to capture the effect of the gender reform yield interesting results. The time dummy for the reform turns out significantly positive for the deductible choice. This means that men tended to choose higher deductibles from 2004 and onwards, an expected effect because this group on average got higher premiums. The variable of particular interest is the interaction between gender and the reform. This is highly significant. The interpretation is that both men and women tended to choose higher deductibles after 2004, but the reform effect is markedly lower among women. We also note that the gender reform had no effects on the overall claim frequency.

c) Modeling expected severity – a joint model of the number and severity of claims

The aim of this paragraph is to implement the positive correlation test in a framework where the dependent variable of interest is the total claim severity. The total severity is the product of number of claims times the average individual claim cost. Thus, one can view total severity as a multiple of two processes, one process that describes the number of claims and one process that describes the individual claim cost (given that at least one claim has occurred). The number of claims can be modeled through a Poisson count type model. The claim cost takes on values from zero to infinity. As most loss distributions are right skewed a natural choice may be the gamma distribution or the inverse Gaussian (see for example Hogg and Klugman 1984).

To illustrate the idea, assume that a single insurance policy can generate $i = 0, 1, 2, 3 .. A$ claims over the policy period.[30] The claim cost (*S*) for a single policy is then

$$(4) \qquad S = \begin{cases} 0 & \text{if } A = 0 \\ Z_1 + Z_2 + ... + Z_A & \text{if } A > 0 \end{cases},$$

where Z_i is the claim cost connected to claim *i*. The expected total claim cost ES can be written as[31]

$$(5) \qquad ES = E(A)E(Z).$$

As seen from (5) the severity is the product of a count process and a claim process. The main idea is now to represent the two terms on the right hand side with a count model for the number of claims and a suitable model for claim size (conditional on the occurrence of a claim). These two parts will be modeled and estimated simultaneously.

Now, consider a single insurance policy *j*. The joint distribution of the number of claims (A_j) and the total severity of claims (S_j) can be written as

$$(6) \qquad f(S_j, A_j) = f\,(S_j \mid A_j) f\,(A_j),$$

where $f\,(S_j \mid A_j)$ is an appropriate distribution function for the severity and $f\,(A_j)$ is a count distribution function. I will in the following assume that the number of claims follows a Poisson-model with parameter λ_j and that individual claim sizes (Z_i) follow an inverse Gaussian distribution with mean μ_j and shape parameter ζ. The inverse Gaussian is suitable within this framework because it is right-skewed, which means that most of the probability mass is located in the low range of the random variable. Moreover, this

---

[30] See HSRJ (2007) for a detailed explanation of the different modeling steps.

[31] The expectation is the sum of the probability for a given count times the individual claim cost; that is:
$ES = 0 * \Pr(a = 0) + EZ_1 * \Pr(a = 1) + (EZ_1 + EZ_2) * \Pr(a = 2)) + ... + \sum_{i=1}^{A} E(Z_i) * \Pr(a = A)$.
Moreover, because the individual claim size is independent and identically distributed $EZ_1 = EZ_2 = ...EZ_A = EZ$.
By this assumption one can write $ES = \left[(\Pr(a = 1) + 2 * \Pr(a = 2)) + ... + A * \Pr(a = A)\right]EZ = E(A)E(Z)$.

distribution has a long tail. As seen from table 1c approximately 70 percent of the claims are lower than 50 000 NOK, while a quite small percentage of the claims exceed 200 000 NOK.

Since the sum of independent inverse Gaussian distributed random variables (S) is also distributed as inverse Gaussian, we can write $f\,(S_j \mid A_j)$ as[32]

$$(7) \qquad f(S_j \mid A_j) = f_{IG}(S_j; A_j \mu_j, A_j^{\,2}\zeta) .$$

I will assume that the shape parameter is constant, although the framework used here allows one to model ζ as a function of independent variables (see HSRJ (2007) for such an extension). Next, assume that the mean individual claim cost can be written as $\mu_j = e^{X_j'\beta}$. Moreover, I will follow the standard approach to model the mean of the Poisson process as $\lambda_j = e^{X_j'\gamma}$ (see for example Cameron and Trivedi 2005, page 668). This parameterization implies that $\log \lambda$ and $\log \mu$ can be written as linear functions of the explanatory variables, which is often denoted as a log-link formulation in the generalized linear model literature. For the purposes of the positive correlation test I will use the same X-vector for both severity and number of claims. However, the approach is flexible and one may use different X-vectors in the severity and the count regression.

Inserting equation (7) into (6) and taking the log of both sides yields

$$(8) \qquad \log\{f(S_j, A_j)\} = \log\{f_{IG}(S_j; A_j \mu_j, A_j^{\,2}\zeta)\} + \log\{f_P(A_j; \lambda_j)\} ,$$

where subscript P denotes the Poisson distribution. Equation (8) shows the two different elements in the likelihood contribution from policy j. One can note that the parameters in

---

[32] The pdf of the inverse Gaussian (IG) variable x is $p(x;\mu,\zeta) = \left[\dfrac{\zeta}{2\pi x^3}\right]^{\frac{1}{2}} e^{\frac{-\zeta(x-\mu)^2}{2\mu^2 x}}$, where μ is the mean and ζ is the shape parameter. The sum of IG random variables is distributed IG. In our specific case we then have $S = \sum_{i=1}^{A} Z_i \sim IG(S; A\mu, A^2\zeta)$.

the inverse Gaussian and the Poisson are unrelated and for insurance takers with no claims only the Poisson count model contributes to the likelihood.[33]

Summing (8) over all j=1 to N yields the full log likelihood, which is a function of the parameter vectors β, γ, ζ. As usual, we maximize this likelihood with respect to these parameters. Denote now the specific parameter connected to the deductible as β_D for the severity and γ_D for the count. These parameters measure the effect of the deductible on claim severity and the number of claims, conditional on all the X-variables. If these are found to be negative the positive correlation property is confirmed.

Table 7 below displays the parameter estimates under different data set restrictions that are explained in the table.[34]

Table 7.
Maximum likelihood parameter estimates of (7). Deductible measured in 1000 NOK. Only claims higher than the highest deductible included. Only the parameter estimate of β_D and γ_D is shown

	All contract years			First contract year		
	The effect of deductible on number of claims ($\hat{\gamma}_D$)	The effect of deductible on severity ($\hat{\beta}_D$)	N	The effect of deductible on number of claims ($\hat{\gamma}_D$)	The effect of deductible on severity ($\hat{\beta}_D$)	N
All insurance policies	-0.0254*** (0.0065)	0.0606*** (0.0065)	470989	-0.0224* (0.0110)	0.0530* (0.0216)	168279
Removal of the 15000 deductible group	-0.0283*** (0.0081)	0.0587**	466634	-0.0326* (0.0131)	0.0345 (0.0303)	166655
Removal of the 15000 deductible group and claims above NOK 600 000	-0.0336*** (0.0081)	-0.0019 (0.0087)	466429	-0.0332* (0.0135)	0.0000 (0.0095)	166568
Year dummies	Yes	Yes		Yes	Yes	
County dummies	Yes	Yes		Yes	Yes	
All tariff variables included	Yes	Yes		Yes	Yes	

* for p<.05, ** for p<.01, and *** for p<.001. Standard error clustered within customers.

[33] Most often there is one-to-one correspondence between the insurance taker and the insurance policy. However, some insurance takers have several insurance policies as one policy is written for each object the insured owns.

[34] I have cross-checked the estimates in table 7, which is based upon a STATA-ML routine, with the R-routine "rsm" in the "gamlss-package", see www.gamlss.com and HSRJ. To do this I extracted a sample of contracts containing the number of claims, the total severity, age, gender, insurance value and deductible choice. I ran this limitied model in STATA and R (using the four latter mentioned variables as explanatories) and obtained exactly the same results. The results are available upon request.

First, irrespective of the restrictions taken - there is always a positive dependency between the number of claims and coverage (that is: a negative dependence between number of claims and the deductible). The estimated parameter is statistically significant in all specifications. The estimate from the Poisson model indicates that an increase of NOK 1000 in the deductible leads to an approximate 3 percent reduction of the expected number of claims.

Second, from the third row in table 7 we see that for claim sizes below 600 000 NOK the results indicate a zero correlation between claim size and coverage. Furthermore, note that the restriction on the claim size only leads to that approximately 200/100 claims (all contracts/first year contracts) fall out of the estimation sample, which constitutes approximately 2 percent of all claims. Thus, conditional on "normal" claim sizes, one can conclude that there is zero correlation between claim size and deductible choice. Moreover, for first year contracts this conclusion also holds even when large claims are included (even though the point estimate is positive).

However, for all contracts and all claims included – this conclusion is no longer valid, see the two upper rows in column 3. Here, claim size increases in the deductible. This relation is highly significant and the point estimate is large. Actually, the point estimate in row 1-column 3 indicates that 1000 NOK in increased deductible increases expected claim cost with around 6 percent. This large effect is puzzling. By exploring the different tariff factors I find that the main reason for this surprising outcome is connected to homes and apartments that are rented. To be more concrete: policyholders who offer houses/units for let are overrepresented in the highest deductible group. Moreover, the incidence of large/catastrophic claims is clearly higher in this group. Thus, there is reason to believe that this constellation is the main reason for the large effects found in column 3. But, a puzzle still remains. Why is there a preference for high deductibles within this risk group? Two causes seem to explain this puzzle. First, premiums are much higher when a home/house/apartment is for let. This fact increases the incentive for choosing a high deductible. Second, there is also reason to believe that the owner of an apartment or a house for let may be able to *shift* the deductible over to the tenant through the rental agreement. Therefore, in these situations the risk type of the tenant is irrelevant for the insurance purchase decision – demand is most likely driven by a cost minimization motive of the owner, not the risk type of the tenant. This leads the owner to choose high deductibles, even though the insured home has a very high risk of damage due to the high risk of the tenants.

To sum up so far: The benchmark model, the bivariate model and the joint model all indicate that information asymmetries are present in the market under study. The claim frequency clearly declines in the deductible and claim costs are independent of the deductible choice, conditional on "normal" sized claims, which constitute around 98 percent of all claims. However, the joint model shows that when also the larger claims are included the claim size does in fact increase in deductible choice. It turns out that rented homes/apartments have a substantially higher "large claim" incidence rate than other risk groups. These insurance policies are also characterised by large deductibles. I argue that the owner of these objects typically can shift the deductible over to the tenant, which implies that demand for cover is driven by a cost minimization motive of the owner.

There are, however, several dimensions of this asymmetry that one should explore in more detail. At least five questions are of interest to explore further. First, is the information asymmetry present for all customer vintages or present only among the newest customers? Second, is the information asymmetry independent of the year the contracts are written? Third, is the asymmetry found above sensitive to alternative specifications of the models? Fourth, and maybe the most important, is it possible to disentangle moral hazard from adverse selection. Fifth, may it be that the test is biased due to preference based selection. The first three questions - that basically consist of different robustness tests - are addressed in Appendix 4, the moral hazard versus adverse selection issue is dealt with in the next section and the importance of preference based selection is discussed in section 9.

8. Moral hazard or adverse selection?

The next natural step in the exploration of the information problem that has been revealed is to examine the sources of the information problem. The most important task is to separate adverse selection from moral hazard. The literature has suggested different approaches and I will briefly recapture how these studies have assigned the information asymmetry to either adverse selection or moral hazard. Thereafter, I explain the approach I will use to discriminate between these two possibilities.

The study by Cohen (2005) finds that the information problem is non-existent among drivers with small or little experience and large and significant among drivers with 3 or more years of driving experience. She attributes this finding to learning – drivers are learning their risk type through driving experience. This finding lends support to adverse selection as the cause of the information problem. The only way this finding squares with moral hazard is that precautionary behavior among inexperienced drivers has an effect on claim probability while precautions taken by experienced drives have not. Moreover, Cohen investigates a car insurance market with no experience rating and no information sharing among the insurance providers, which is quite unusual. In a car insurance market with experience rating and information sharing one would expect presence of moral hazard because the effective price of the insurance is a function of past claims.[35]

Finkelstein and Poterba (2004) and Finkelstein and McGarry (2006) investigate the market for annuities and the market for long term health care insurance, respectively. Within these markets intuition tells us that the most likely explanation for an information asymmetry is adverse selection. In the annuity market the adverse outcome is a "long life". If one detects an informational asymmetry in this market, a moral hazard story must be that people who buy this insurance start to live healthy after they have obtained the cover. The adverse selection story is that people (for some reason) have private knowledge of the likelihood of a long life and those who believe they live longer than average buy this insurance more frequently than those who believe that their life span is relatively short. The same reasoning goes for the long term care health insurance market, where one typically has to survive beyond the average life expectancy in order to gain any utility from this type of insurance. Thus, both these examples describe markets where the utility gain arrives a long time after the insurance purchase, in many cases 10 to 20 years after the initial purchase decision. Thus, claiming that individuals strategically adjust their behavior to gain utility of a future uncertain insurance event is certainly a more heroic hypothesis than a hypothesis that involves private information of risk type.

Abbring et al (2003, 2008) develop a theoretical framework in order to develop a formal statistical test for moral hazard. They apply their framework to panel data on car

[35] Alternatively, one can loosely say that with experience rating the future price is a function of current behavior. The higher the future price becomes as a consequence of a claim today, the lower the likelihood for moral hazard.

insurance contracts from France (Abbring et al 2003) and The Netherlands (Abbring et al 2008). Both these markets are characterized by a bonus-malus system and information sharing between insurance companies. Within a bonus-malus system the future price of the insurance is dependent on current claim behavior. Filing a claim typically raises the future price. Moral hazard can be detected by looking for negative state dependence in claim probability. That is: a claim in year *t* that leads to a premium increase *t+1* should reduce the claim probability in *t+1* if moral hazard is present. For their French data they find no evidence on moral hazard, while they conclude with presence of moral hazard in their Dutch data.

The data set utilized here is comparable to the data by Cohen (2005) along two dimensions. First, there is no experience rating or bonus-malus system in this market. Second, the information sharing between companies is limited.[36] These two features make it difficult to implement the Abbring et al. methodology because there is limited hope of assigning any state dependence to changes in incentives. It is also difficult to argue – as in Finkelstein and Poterba/Finkelstein and McGarry - that moral hazard is unlikely in the home insurance market because one can imagine several ways of adjusting the behavior after an purchase of a home insurance that increases the likelihood for an adverse event.

Therefore, the first route I will proceed along in order to distinguish between adverse selection and moral hazard is to utilize the detailed information about claim types in the data. Whenever a claim is filed it will be assigned to a claim category. Altogether there are 69 such categories and I observe all of them. A quite simple idea is to use the description of the claim categories and sort out those which are more likely to be caused by moral hazard than by adverse selection. The precision of this method is of course debatable, but I believe that it is fruitful to use the detailed claim information as a first approach to the problem. The process for selecting claim types is based on a) the consequence of the claim and b) whether the

[36] Indeed – there exist a common register in Norway where all home insurance claims are registered (called Forsikringsselskapenes Sentrale Skaderegister, shorthand FOSS, see http://www.fnh.no/no/Hoved/Fakta/Skadeforsikring/Skadeforsikring-a---a/Forsikringsselskapenes-Sentrale-Skaderegister---FOSS/ , accessed April 22nd 2010. However, the use of this register strictly regulated and a company is in general not allowed to check the claim history of a new customer. However, if the customer has been asked to report past claims in the underwriting process, the company can check the claim history whenever he/she report a claim. Thus, there is a clear incentive to report truthfully when signing up with a new insurance company. In a case where untruthful reporting is detected, the insurance company can reduce the size of the indemnity and in extreme cases – deny any payout.

claim with a high likelihood can be caused by lack of concentration, laziness or other type of negligent behavior that typically can be induced by the degree of insurance coverage.

Recall at the outset that if moral hazard drives the information asymmetry a substantial number of insured must react with less caution as a consequence of having a low deductible compared to a case where the customer retains a high deductible. The maximal difference between the lowest and the highest deductible is here 13 000 NOK (approximately $2000). This amount may also be interpreted as the maximal gain of negligent behavior for a person who switches from the highest to the lowest deductible. The flip side of an insurance event, however, is administrative costs, the likelihood of being reviewed for negligent behavior and the possible disutility connected to moving from home for a shorter period (given that the event makes the home inhabitable). Thus, claims that incur costs of moving and other serious side-effects are probably not caused by moral hazard. This implies that broken water pipes, fire and other events that may cause serious consequences most likely can be attributed to adverse selection. What's left is claims that involve theft and other events that one intuitively believes can be avoided through cautious behavior. The list of claims I remove from is then: Theft of outside belongings, other claims caused by theft, snow on roof that causes damages or roofslides and claims connected to legal processes regarding the property. [37]

Table 8 show the result of a biprobit regression where claims just listed are removed

[37] The home insurance actually cover up to 100 000 NOK in legal expenses. The insurance event is typically a conflict between the homeowner and for example the seller of the home. Or it may be a conflict regarding furnishing of the home. Having this cover may give incentives to a higher conflict level. One could imagine that customers with low deductibles would use this opportunity more frequently than customers with high deductibles.

Table 8. Biprobit. Claims most likely connected to moral hazard removed from sample.

All contract years		***First contract year***	
Rho	***Chi-square***	***Rho***	***Chi -square***
(Standard error)		(Standard error)	
-0.0401***	19.05	-0.0390**	7.01
(0.0091)		(0.0147)	

* for p<.05, ** for p<.01, and *** for p<.001. Standard error clustered within customers.

The selection restriction described above removes approximately 18 percent of the claims in the biprobit that includes all contract years and 16 percent of the claims in the analysis of first contract year contracts.

Surprisingly, the correlation coefficient of the biprobit increases quite much (approximately 16 percent for both specifications) which indicates that adverse selection rather than moral hazard is the prime suspect for the information asymmetry.

The second approach is to utilize the exogenous gender reform that was implemented Jan 1st 2004. This reform caused an exogenous change in the price for all customers as the company pursued an adoption policy that aimed towards a neutral profit effect of the reform. In practice this meant a higher price for men after the reform and a lower price for women, which was confirmed in the table shown in section 5.

The basic idea can be outlined as follows: The price change affects the policyholders' coverage decision. Change of coverage should then change the prevention effort if moral hazard is present and if there is a strong effect from prevention to claim probability then this effect may also be of statistical importance.

The first requirement is that a change in the premium should have an effect on the coverage decision. Here, theory does not give us a clear prediction. To see this, assume that the insurance premium increases. If we assume decreasing absolute risk aversion (DARA) an increase in the premium will lead to a desire for more insurance through the reduction of income net of premium in both the accident and non-accident state. However, if we assume constant absolute risk aversion (CARA) the demand for cover would be unaffected. Moreover, the pure income effect of a higher price for insurance (or –equivalently – a lower price for self insurance) will probably lead the policyholder to demand less insurance. Thus, the net effect of this reform on the coverage decision is an empirical question.

To address this empirically I have first checked whether the deductible choice is sensitive to premium changes. To test this I have conducted a fixed effect regression of change in the deductible (the dependent variable) on changes in all tariff factors plus the lagged value of premium change.[38] Two different premiums are used on the right hand side of this regression. First, I use the actual premium change in *t-1* (that is the premium in *t*-1 minus premium in *t*-2) as the independent variable. Second, I use the change in the default premium in period *t* (that is: the default premium in *t* minus the default premium in *t-1*).[39] Both specifications yield a significant positive sign on the lagged premium, which means that a premium change conditional on all other possible changes in the other tariff factors has an effect on deductible choice. This result establishes an empirical link between premium changes and coverage choice.

The next step is to address of whether a deductible change lead to a change in the optimal risk preventing effort. Say now that claim probability can be written as $\pi(X, e^*(D))$, where X is the tariff characteristics, e^* is the optimal level of effort given the deductible D. Assume now that an exogenous price shock occurs. The price shock induces the policyholder to reconsider the deductible level. The basic idea is now that a change in the deductible will lead to a re-optimization of the effort. More precisely, a higher deductible leads to higher self-prevention effort, that is $e'(D) > 0$. This change in effort will in the next step lead to lower claim probability, that is $\pi'_e(X, e(D)) < 0$. If these mechanisms are at work, which is inherently an empirical question, the presence of moral hazard can be detected.

The third step is to assess the validity of the gender reform as a suitable instrument. First, the reform must have a notable effect on the deductible choice and should, moreover, be uncorrelated with the claim probability.[40] The different variants of the positive correlation test presented above indicate that this is actually the case - the reform had a clear effect on the deductible choice but not on the claim probability. Moreover, an important assumption is that the risk type of each policyholder is contained in the time constant individual effect.

---

[38] The regression results are not reported but are available upon request.

[39] Recall that the default premium is the premium that the policyholder would have paid under the default deductible. This premium is independent of the deductible choice and therefore exogenous. When using the actual premium I have to use the *lagged* change to avoid endogeneity.

[40] Think of a hypothetical situation where one knew that moral hazard was present. If a reform was implemented under these circumstances only a reform that was neutral regarding the retention of risk in the insurance population would produce a non-significant result on the claim probability. If, for example, the reform forced all policyholders to retain more risk (and moral hazard was present) the claim probability would fall. Thus, if the reform actually implied an aggregate incentive to retain more risk (which is a difficult empirical measurement problem) one can in fact interpret the non-significant effect of the reform on claim probability in table 6 as an indication of non-existence of moral hazard.

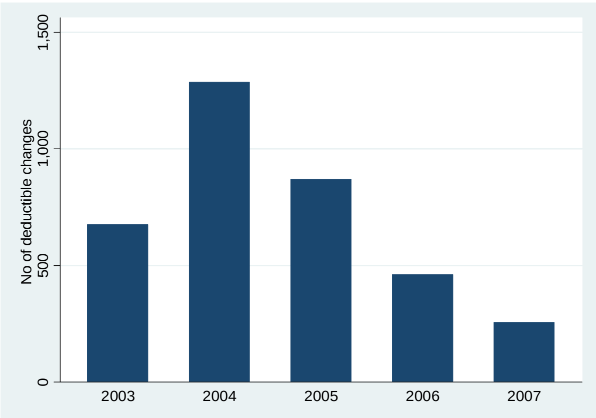
Because this time constant individual effect cancels out in a fixed effect or first difference regression the remaining spillover effect from the insurance cover to the claim probability must be caused by moral hazard. [41]

Details regarding changes in the deductible choice for the panel dimension of the data is shown diagram 1 to 6, where both the frequency of changes and the mean change (in NOK) are displayed.

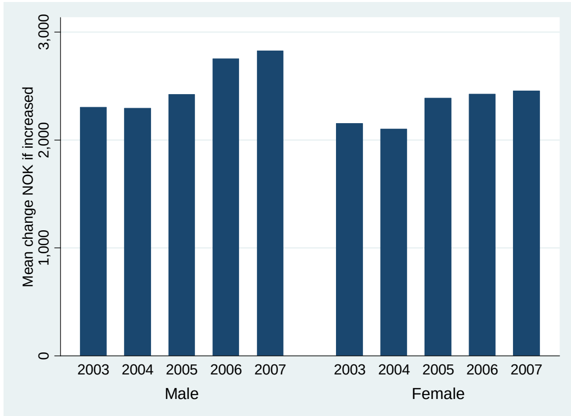
[41] Learning may be a variable that violates this requirement. Learning may be relevant within home insurance because one can imagine that the policyholder learns about the probability of adverse events over time. The house may, for example, contain some hidden problems that can only be discovered by residing in the house for a minimum of time. Thus, it is possible that a policyholder will react to learning by changing the deductible. This possible effect is not addressed here.

Diagram 1 - 6.
Changes in deductible over gender and year. Number of changes and mean change in NOK

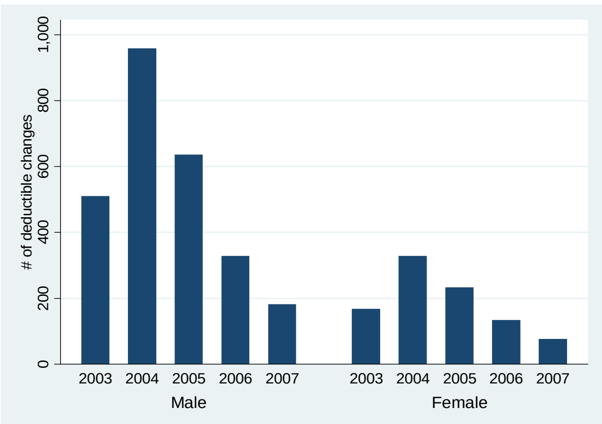
Number of changes in deductible, by year.



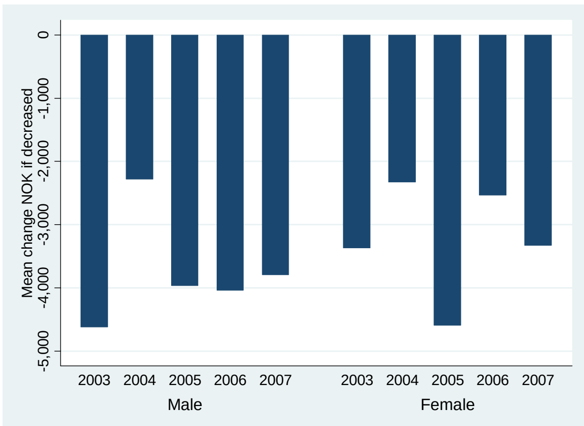
Mean change if up; by gender&year



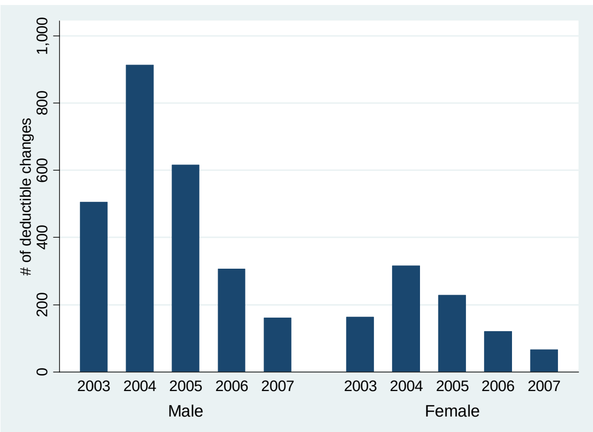
Number of changes; by gender&year



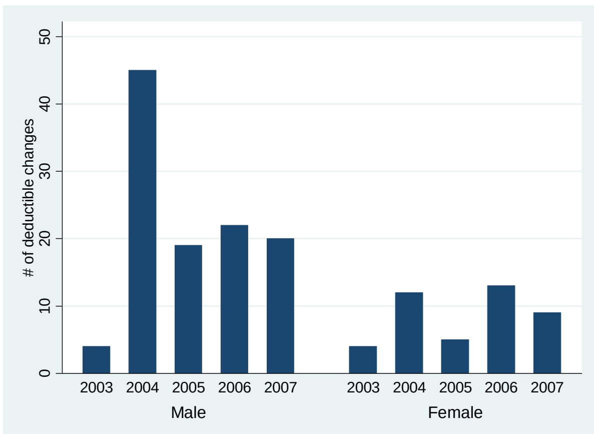
Mean change if down; by gender&year



Number of changes if up; by gender&year



Number of changes if down; by gender&year



The total number of deductible changes are 3 550 of a panel of 228 017 contracts with duration of at least two years. This constitutes approximately 1.5 percent of the panel. Thus, there is an extreme degree of state dependence in the deductible choice. Only 153 of these changes implicated a reduction in the deductible (see 6th diagram in panel 1). That is – 96 percent of the changes involved increases in the deductible. Nevertheless, the pattern of deductible changes is interesting. More than 35 percent of all observed changes took place in 2004 and if we lump together 2004 and 2005 – 60 percent of the observed changes came about in these two years. Even more interesting, the number of deductible changes almost doubled from 2003 to 2004 and –moreover - the 2004 number is approximately 1.5 times the 2005 number. Thus, the bar diagram indicates that the probability of a deductible change peaked in 2004. Furthermore, the high prevalence of deductible changes in 2005 fit into a model where policyholders need some time to re-adjust to the new regime. Thus, the descriptive evidence shown in diagram 1-6 indicates that using information about the gender reform may be a fruitful approach to identify moral hazard.

Having now established the basic arguments I proceed with a 2SLS fixed effect (FE) /first difference (FD) model. The first stage equation is specified as

$$(9) \qquad D_{it} = c_i + X_{it}'\tau + \theta R_t + u_{it},$$

where R is the instrument. The variable R is an indicator variable that takes the value zero before 2004 and one thereafter. The second stage is specified as

$$(10) \qquad A_{it} = a_i + X_{it}'\beta + \varsigma \hat{D}_{it} + v_{it},$$

where $\hat{D}$ is the predicted deductible from the first stage regression and A_{it} is the number of claims for individual i in year t. Recall now that the deductible and the tariff variables is chosen/set at the first day of the contract. Claims are dripping in from the that same day and onwards to the 365th day. This implies that one can regress A_{it} on $\hat{D}_{it}$ without incurring any simultaneity problem. Recall also that the main idea is that $\hat{D}_{it}$ mirror the change in the optimal effort level. That is – the exogenous price variation leads to reoptimization of the effort– which means that equation (10) gives the causal effect of a change in effort on claim probability. Above I assumed that the ex-ante riskiness (risk-type) is time-independent and is contained in the fixed individual effect a_i. As this unobserved effect cancels out in a FE/FD consistent estimates of parameters are obtained if there is no correlation between the residual and the X-vector, Wooldridge (2002 , page 266).[42]

The important parameter in equation (10) is ς. If ς is negative and statistically significant it indicates presence of moral hazard. If not – then adverse selection is probably the primary source of the information problem. The results are presented in table 9.

Table 9. **A test for moral hazard: 2SLS fixed effect and first differences regression. Dependent variable: Number of claims Instrument: Reform indicator**

	FD	***FE***
	Coefficient	***Coefficient***
Deductible	-0.0004281	-0.0001937
	(.0002745)	(0.0001285)
All tariff variables included	Yes	Yes
	First stage results	First stage results
Reform indicator	7.8699 **	16.4288***
	(2.8137)	(4.1593)

*for p<.05, ** for p<.01, and *** for p<.001. Asymptotic standard errors in parenthesis.

Note first that the first stage results indicate that the instrument is quite strong. The parameter ς turns out negative, which is in line with a moral hazard story – the higher the deductible the lower the likelihood for filing a claim. However, in both specifications the point estimate is statistically insignificant (a p-value of 0.119 in the FD and 0.132 in the FE regression). Thus, these both the FE and the FD-model give only weak and non-significant evidence on moral hazard.[43]

[42] Moreover, the choice to use FE or FD does not matter when the number of time periods is 2. With 2 or more periods and with serially uncorrelated residuals, the FE model is more efficient. With serially correlated residuals the FD model is the preferred one (Wooldrige (2002, page 284)). In order to address this I choose to present both FE and FD estimates which in essence provides at test of the assumption about the residual. If estimates from FD and FE are close to each other - one can infer that residuals are well behaved and that the strict exogeneity assumption probably is valid.

[43] It is also important to note that the benchmark regression involves three types of customers. First, there are customers that only are present before the reform (2002 and 2003). For these the reform variable will always be zero. Second, there are customers that enter the dataset after or in the reform year (2004 and later). For these the reform variable will be one. Finally, the sample contains customers who live through the reform. By

The results in this section – taken together - indicate therefore that the information problem documented over fit more into a story of adverse selection than of moral hazard.

9. The importance of preference based selection – the case of risk aversion

The discussion in section 2 provides several examples of cases where the positive correlation test may fail to produce a significant result. One important explanation for why this can happen is heterogeneity in risk aversion combined with an ability to extract rent from differences in risk aversion. Thus, if risk aversion matters for the deductible choice and claim probability then the positive correlation test may be biased because unobserved risk aversion is omitted from the test. As the analysis above has shown, however, the result in favor of an information asymmetry is present. Nevertheless, it is of interest to check whether this result is biased due to lack of controls for preference based selection. I will pursue this task in the following by including proxies for risk aversion in the positive correlation test.

The direction of the potential bias is an empirical matter and depends crucially on the possible link between risk aversion and claim probability. One possibility is that persons with high risk aversion typically are low-risk types than people with low risk aversion. If this is the case then highly risk averse people will choose low deductibles, but at the same time, file less claims than persons with low risk aversion. From the point of view of the insurance company this possibility may be coined as advantageous selection. One of the few studies that have investigated this issue in depth is Finkelstein and McGarry (2006) who find that self-reported measures as "preventive health care activities" and "seat belt use" have a significant impact on both the demand and the subsequent use of long term health care insurance. Another is Fang, Keane and Silverman (2008) who document a significant presence of advantageous selection in the highly regulated US Medigap market. Both studies utilize self-reported measures on risk preference and cautious behavior collected from nationally representative surveys.

including all these three different groups it is implicitly assumed that the reform affects these different groups in an equal manner. This may not be the case as one can imagine that customers that were present in the company in 2003 reacted differently to the reform than customers who were recruited in 2004 or later. For example, a customer that was recruited in 2003 may be less sensitive to price changes (due to habit, switching costs etc) than a new customer entering in 2004 (who typically had scanned offers from different insurance suppliers before a decision was made).

However, to the best of my knowledge, no one has directly used microeconomic public register information in order to construct proxies for risk aversion. One main advantage with administrative information is the accuracy of the information due to the fact that this information is used for tax purposes. Thus, the possibility of error in reporting due to low interest or random answering is reduced to a minimum. The first step is to define a set of risk aversion proxies. The second step is to include these proxies in the positive correlation test in order to address the possible omitted variable bias that may be caused by heterogeneity in risk preferences.

Finding suitable risk aversion proxies

Several studies have explored the connection between self-employment and risk preferences. Barsky, Juster, Kimball and Shapiro (1997) found that the probability of self-employment decreases in risk aversion. More recently, Cramer, Hartog, Jonker and van Praag (2002) found that risk aversion has a clear negative effect on entrepreneurship. Moreover, Ekelund *et al.* (2005) confirm a negative relationship between risk aversion and self-employment in Finland using psychometric survey questions. Also a recent study by Brown *et al.* (2008) using the US Panel Study on Income Dynamics (PSID), a negative relationship between self-employment and risk aversion is confirmed. Furthermore, a study by Fang and Nofsinger (2008) who use the 2004 Survey of Consumer Finance demonstrates that entrepreneurs do have a higher degree of risk tolerance than non-entrepreneurs. Finally, in recent article by Ahn (2010) the association between low-risk tolerance and self-employment is confirmed.

The main identifier for self-employment in the data is "income from self-employment". I define a customer as "self-employed" if more than 50 percent of the gross income stems from self-employment. Approximately 5 percent of the insured meets this requirement. Thus, I construct a dummy that takes the value one if more than 50 percent of the gross income consist of self-employment income, and zero otherwise.

Another well-investigated proxy for risk aversion is the sector of employment. For example Bellante and Link (1981) find that risk averse people are more likely to be employed in the public sector. Moreover, Rozkowski and Grable (2009) conclude that public sector employees show a higher degree of risk aversion than private sector employees. Also Buurman, Dur and Van den Bossche (2009) document that risk aversion is higher among public sector employees. They analyze a survey where respondents were offered a significant “response reward”. The respondents could choose either a risky option (a lottery ticket) or a risk free alternative (a redeemable gift certificate). The authors find that public sector employees were significantly less likely to choose the risky option than private sector employees. Finally, Aarbu and Schroyen (2009) analyze responses to hypothetical income gamble among a representative set of Norwegian residents. From these responses a cardinal measure of risk aversion for each respondent in the sample is constructed. Using this cardinal measure the authors are able to show that public sector employees are more risk averse than private sector employees.

The data contain administrative information (NACE-codes) on main employment sector. By using the two first digits of the NACE-code I am able to accurately identify customers employed in the public sector. Altogether a little more than 28 percent of the sample is employed in this sector. Of these - approximately 55 percent is women. I construct a dummy that takes the value 1 if the person is employed in the public sector, and zero otherwise.

Investment in stocks is associated with preference for risks. For example, in the standard portfolio problem the allocation of wealth to the risky asset is dependent on risk aversion. All else equal, a highly risk averse agent invests less in the risky asset than an agent with low risk aversion, see for example Gollier (2001) page 53. Moreover that risk aversion affects investment in stocks is documented by Barsky, Juster, Kimball and Shapiro (1997).

I define a variable that is the value of stock investment divided by the sum of bank deposits, mutual fund allocations and stock investment. Thus, the denominator can be interpreted as the sum of “liquid assets” the customer posses. Note that this is a continuous variable ranging from zero (no stock investments) to one (all liquid assets invested in stocks): The higher the fraction of stocks, the less risk averse one should expect the customer to be. I have also tried to use both stock investments and fund investments in the numerator; however, no significant changes in the results were found.

One possible source of concern when including these proxies in the regression is simultaneity bias. Therefore, I have tested whether the results change by using time means instead of using contemporaneous variables. For example, a customer observed in 5 consecutive years is defined as self-employed if the fraction of self-employment income over these 5 years exceeds 50 percent of gross income, while the same customer may be defined as self-employed only in 4 of the 5 years when using contemporaneous income definitions. Obviously for a customer observed only one year I have to use the contemporaneous proxy variables for risk aversion. These tests (not reported) do not lead to any changes in the results.

These three risk aversion proxies are included in the bivariate probit regression in order to control for risk preferences of the policyholders. In order to also control for possible income effects that may be correlated with these risk aversion proxies I also include a 10-piece income spline in order to open up for that outcome variables is non-linear in income.[44] Table 10 shows the results, where I only show the correlation coefficient along with the parameter values for the three risk proxies plus the income variable.

[44] The splines are constructed as follows: I partition the relevant income variable into 10 deciles so that income can be written as the sum of first decile income plus the second decile income and so forth. Thus, each policyholder's income is represented with 10 income categories, where the sum of these categories equals the total income. By representing income this way possible non-linear associations between income and outcome variables will be captured. The results shown in table 11 are based on personal gross income. I have also tried household income before tax without any change in the main results. Moreover, I have tested whether high income variation leads policyholders to choose lower deductibles, by including the variance of income as an explanatory variable. The results are more or less unaffected.

Table 10. Including risk aversion proxies in the positive correlation test. Biprobit. Claim equation :Having at least one claim higher than highest deductible Deductible equation: Deductible higher than 6000 NOK

	All contract years		First contract year	
Risk aversion proxy	**Have at least one claim**	**High deductible (>6000)**	**Have at least one claim**	**High deductible (>6000)**
Work in public sector	-0.033963**	-0.109737***	-0.029460	-0.111062***
	(0.011165)	(0.014216)	(0.018482)	(0.014775)
Stock investment i % of sum liquid assets	0.018233	0.098925***	-0.001244	0.129731***
	(0.020112)	(0.022174)	(0.034064)	(0.025074)
Self employed	-0.006064	0.158644***	-0.024391	0.192179***
	(0.018912)	(0.021274)	(0.031452)	(0.022562)
Income 1 decile	-0.006621	-0.019208	0.008079	-0.016447
	(0.010554)	(0.009965)	(0.017469)	(0.011872)
Income 2 decile	0.027651	-0.199292*	-0.073362	-0.218302*
	(0.083573)	(0.089522)	(0.133505)	(0.108668)
Income 3 decile	-0.486993**	0.083344	-0.441494	0.002816
	(0.176134)	(0.161947)	(0.284947)	(0.219962)
Income 4 decile	0.355866	-0.425327*	0.529403	-0.115160
	(0.246869)	(0.212642)	(0.401293)	(0.303012)
Income 5 decile	-0.178257	0.147252	-0.149592	0.007894
	(0.288920)	(0.234102)	(0.465452)	(0.351166)
Income 6 decile	0.115461	-0.393337	0.147143	-0.450694
	(0.304986)	(0.242400)	(0.505983)	(0.371762)
Income 7 decile	0.213066	0.119832	-0.028769	0.574488
	(0.272175)	(0.219183)	(0.469973)	(0.335408)
Income 8 decile	-0.149867	0.210452	0.083563	-0.216817
	(0.185333)	(0.154243)	(0.318221)	(0.227116)
Income 9 decile	0.033909	0.050743	-0.007168	0.122058
	(0.089551)	(0.087426)	(0.156493)	(0.114533)
Income 10 decile	-0.039798	0.044344	-0.060803	0.058847
	(0.023041)	(0.024315)	(0.043952)	-0.111062***
All tariff variables included	Yes	Yes	Yes	Yes
Rho	-0.03655***		-0.037209**	
N	4.64978e+05		1.66150e+05	
ll	-1.54574e+05		-5.83740e+04	

*for p<.05, ** for p<.01, and *** for p<.001. Standard error clustered within customers.

All risk aversion proxies yield results that are in line with the above mentioned literature. For example, self employed choose systematically higher deductibles and people with high share of their liquid assets invested in stocks are significantly more likely to choose a high deductible However, somewhat contrary to suggestions in the literature - neither self employed or policyholders with preference for stocks file significantly more claims. Public sector employees – choose significantly lower deductibles and file less claims – a pattern that is consistent with the suggestions in the literature. The estimate of the information asymmetry increases with around 5 to 6 percent. Thus, the result suggests that there is a very small bias in the positive correlation test that stems from selection based on risk aversion. Taken together, the results suggest that heterogeneity in risk aversion is an important factor for understanding deductible choices, but only marginally important for the measure of the information asymmetry.

It is also of isolated interest to note that income/wealth (irrespective of how I measure it) does not seem to have any effect on deductible choice – conditional on all the tariff variables. Unconditionally, however, I find that income/wealth and deductible choice is strongly positively correlated.

10. Concluding remarks

This paper has shown evidence on asymmetric information in the home insurance market. This finding is based upon a large data set containing approximately half a million contracts sold and renewed in the period from 2002 to 2007. By linking the insurance contract information with administrative register data I have shown that the insurance population analyzed here mirrors the full population very well. Moreover, the insurance data is gathered from a major insurer with several decades of experience in the market. Therefore, I believe that the results I provide in this paper can be generalized up to a market level. The test for asymmetric information – the positive correlation test – is implemented through three different approaches with consistent results. Especially, I have proposed a formulation of the test that provides joint modeling of the number and severity of the claim. This formulation shows that individuals who buy more insurance file more claims, but not higher claim sizes given normal sized claims. However, when large/catastrophic claims are included, the estimate from the joint model indicates that claim size increases in deductible choice. The most likely explanation for this phenomenon is that many catastrophic claims occur in homes that are rented. Owners of these houses are most likely able to shift the cost of the deductible over to the tenant, which implies that the owner focus to minimize the insurance premium.

Through two different methods I have tried to disentangle moral hazard from adverse selection, which is an empirically hard task. First, I have used detailed information about claim categories and removed claims that most likely can be attributed to moral hazard. This method does in fact lead to a larger asymmetry. Moreover, I have utilized an exogenous reform that changed the effective insurance price for home insurance. The reform was implemented in 2004. A subsequent change in the deductible level caused by this exogenous reform implied therefore a new optimal claim preventive effort level. Assuming that risk type is contained in the unobserved component of each individual - I implement a fixed effect instrumental variable regression to take out the adverse selection component, which leaves moral hazard (and possibly learning) to be the only causes for a claim. I do find that higher deductibles entail lower claim frequency – indicating moral hazard. However- neither the FE nor the FD estimate is significant and this result leads me to the conclusion of adverse selection as the main cause for the information asymmetry observed in this market.

Lastly, I ask whether risk preferences bias the result from the positive correlation test. Including three proxy measures of risk aversion; whether the policyholder is a public sector employee; whether he/she is self-employed; and the fraction of liquid assets invested in stocks – I find that the measure of positive correlation increases approximately 7 percent. This is a small and statistically insignificant increment, but the point estimate suggests that some preference based selection is at work in this market. The rather low increment may be due to the fact that home insurance for many is almost mandatory. One could for example imagine that selection based on risk aversion is much larger in for example the child insurance market or within other type of health related insurance markets.

Even though this paper gives quite clear evidence of adverse selection in the classical sense, there are several aspects that are left unresolved. First, it would be interesting to measure in more detail the economic consequences of the asymmetric information problem, an issue I have only vaguely touched upon. Moreover, whether insurance companies are able to extract different degrees of rents from the policyholders dependent on their coverage is also a question that should be more carefully addressed in future research.

References

Aarbu K.O. (2010), "Demand Patterns for Treatment Insurance in Norway", NHH Dept. of Economics Discussion Paper 11/2010 & CesIfo Working Paper no 3021.

Aarbu K.O. and F. Schroyen (2009), "Mapping Risk Aversion in Norway Using Hypothetical Income Gambles", NHH Dept. of Economics Discussion Paper No 13/2009.

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

Abbring, J., P.A.Chiappori and T. Zavadil (2008),"Better safe than sorry? Ex ante and ex post moral hazard in dynamic insurance data", CentER DP 2008-77.

Ahn, T. (2010), "Attitudes toward risk and self-employment of young workers", *Labour Economic*s **17**, 434-442.

Arnott, R. J, J.E. Stiglitz, (1988), " The Basic Analytics of Moral Hazard", *Scandinavian Journal of Economics*, Blackwell Publishing **90**, 383-413.

Barsky, R.B., F.T. Juster, M.S. Kimball and M.D. Shapiro (1997), "Preference parameters and behavioral heterogeneity: An experimental approach in the HRS", *Quarterly Journal of Economics* **112**, 537–579.

Brown, S., M. Dietrich, A. Ortiz and K. Taylor (2008), "Self-employment and risk preference", Working paper, Department of Economics, University of Sheffield.

Bellante, D and A. N. Link, 1981, "Are public sector workers more risk averse than private sector workers?," *Industrial and Labor Relations Review* **34**, 408-412.

Buurman, M. R. Dur and S. V. den Bossche (2009), "Public Sector Employees: Risk Averse and Altruistic?", IZA Discussion Paper no 4401 September 2009.

Cameron A. C. , P. K. Trivedi (2005), "MICROECONOMETRICS: Methods and Applications", Cambridge University Press, New York, May 2005.

Cawley, J., T. Philipson, (1999), "An Empirical Examination of Information Barriers to Trade in Insurance", *American Economic Review* **89**, 827-846.

CEA (Insurers of Europe) (2009), "CEA Statistics N°37: European Insurance in Figures", available at http://www.cea.eu/index.php?mact=DocumentsLibrary,cntnt01,details,0&cntnt01documentid=771&cntnt01returnid=104 , accessed May 6, 2010.

Chiappori, P.A (2000), "Econometric models of insurance under asymmetric information: in Handbook of Insurance: Edited by G.Dionne, Kluwer Academic Publishers.

Chiappori, P.A. & B. Salanie (1996), "Empirical contract theory: The case of insurance data", *European Economic Review* **41**, 943-950.

Chiappori, P.A. & B. Salanie (2001), "Testing for Asymmetric Information in Insurance Markets", *Journal of Political Economy* **108**, 56-78.

Chiappori P.A, B. Jullien B. Salanie` F. Salanie (2006), "Asymmetric information in insurance: general testable implications", *Rand Journal of Economics* **37**, 783-798.

Cohen, A. (2005), "Asymmetric Information and Learning: Evidence from the Automobile Insurance Market", *The Review of Economics and Statistics* **87**, 197-207.

Cohen, A and P. Siegelman (2009), "Testing for Adverse Selection in Insurance Markets ". Harvard Law and Economics Discussion Paper No. 651.

Cramer, J.S., J. Hartog, N. Jonker and C.M. van Praag (2002), "Low risk aversion encourages the choice for entrepreneurship: An empirical test of a truism", *Journal of Economic Behavior & Organization* **48**, 29–36.

Cutler D.M, A. Finkelstein, K. McGarry (2008), "Preference Heterogeneity and Insurance Markets: Explaining a Puzzle of Insurance", *American Economic Review* **98**, 157-162.

Dionne G., C. Gourieroux , C. Vanasse (2001), "Testing for Evidence of Adverse Selection in the Automobile Insurance Market: A Comment", *Journal of Political Economy* **109**, 444-473.

Einav, L, A. Finkelstein, J. D Levin (2009), "Beyond Testing: Empirical Models of Insurance Markets", NBER Working Paper No. w15241.

Ekelund, J., E. Johansson, M.R. Järvelin and D. Lichtermann (2005), "Self-employment and risk aversion–Evidence from psychological test data", Labour Economics 12, 649–659.

Fang, H., M. P. Keane, D. Silverman (2008), "Sources of Advantageous Selection: Evidence from the Medigap Insurance Market", *Journal of Political Economy* **116**, 303-350.

Fang, H. and J.R. Nofsinger (2008), "Risk Aversion, Entrepreneurial Risk, and Portfolio Selection" Department of Finance, Insurance and Real Estate, Washington State University, November 2008.

Finkelstein, A. and J. Poterba (2004), "Adverse Selection in Insurance Markets: Policyholder Evidence from the U.K. Annuity Market", *Journal of Political Economy* **112**, 183-208.

Finkelstein A., K. McGarry (2006), "Multiple Dimensions of Private Information: Evidence from the Long-Term Care Insurance Market" *American Economic Review* **96**, 938-958.

Finkelstein, A , J. Poterba, (2006), "Testing for Adverse Selection with "Unused Observables"," NBER Working Papers 12112, National Bureau of Economic Research, Inc.

Gollier, C. (2001), "The economics of risk and time" MIT Press, USA

Heller, G. Z., D. M. Stasinopoulos, R. A. Rigby and P. de Jong (2007), "Mean and dispersion modelling for policy claims costs", Scandinavian *Actuarial Journal* **4**, 281 – 292.

Hogg, R. V. and S. A. Klugman (1984) "Loss Distributions", John Wiley & Sons, New York, USA.

Landsberger, M., I. Meilijson (1994), "Monopoly Insurance Under Adverse Selection When Agents Differ In Risk Aversion", *Journal of Economic Theory* **63**, 392-407.

Rothschild, M. and J. E. Stiglitz (1976), "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information", *The Quarterly Journal of Economics* **90**, 630-649.

Roszkowski, M. J. and J. E. Grable (2009), "Evidence of Lower Risk Tolerance Among Public Sector Employees in Their Personal Financial Matters", *Journal of Occupational and Organizational Psychology* **82**, 453-463.

Stiglitz, J. E. (1977), "Monopoly, Non-linear Pricing and Imperfect Information: The Insurance Market", *Review of Economic Studies* **44**, 407-430.

Sydnor, J. (2010), "(Over)Insuring Modest Risks", *American Economic Journal: Applied Economics*, forthcoming

Appendix 1.
Comparisons of the home insurance sample with population statistics

In this appendix several population measures are compared with similar measures for the sample. The comparisons are conducted by Statistics Norway and are based on the full sample of home insurances; that is, it includes all customers in every year. Recall, that I restrict my focus on new customers arriving from 2002 and to 2007. This difference is most likely of little importance for the overall conclusions; that is, the conclusions drawn in the following will most likely also apply for the somewhat smaller analysis sample.

Three measures are shown in diagram 1 to 3; the first figure shows the level and development of mean wage in the sample, compared to the complete population (about 3.5 million persons). The second and third diagram show similar diagrams for proprietary income and household income after tax. As we see from all the diagrams, the level of wage, proprietary income and household income after tax are all significantly higher in the sample than in the population. The best explanation for this consistent pattern is that people with low incomes typically do not own their own house/home and are therefore not active in the market for home insurance. However, the development over time in these income definitions is almost perfectly correlated, as seen from the diagrams.

Figure A1 – A3. Comparisons of selected income definitions between the sample and the Norwegian Tax Return Register. For wage and proprietary income (2002, 2005, 2007). For household income after tax (2005, 2006, 2007)

Mean Proprietary Income

Mean Household Income after tax

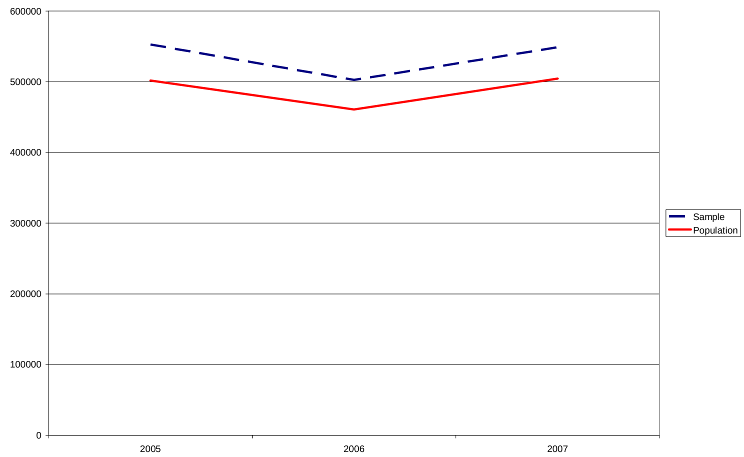


Table 1 gives additional information on the distribution of incomes. I select two income measures: household income after tax and proprietary income. Recall that household income after tax is only available from 2004 to 2007, while proprietary income is available for the whole period. Household income after tax gives the best available picture of the economic position of the household. Proprietary income is, on the other hand, a marginal income source for the majority of households. Thus, finding that the sample also is representative along this marginal income dimension is a reassuring feature for a market level interpretation.

Table A1. **Summary comparisons of the sample with the population**

	Household income after tax: Gini-coeffcient		
Year	2004	2007	
Sample	0.313	0.286	
Population	0.347	0.318	
	Household income after tax: Fraction in 1st decile		
Year	2004	2007	
Sample	2.5	2.5	
Population	2.3	2.3	
	Household income after tax: Fraction in 9th and 10th decile		
	2004	2007	
Sample	36.2	36.2	
Population	40.9	38.1	
	Positive proprietary income: Gini-coeffcient		
Year	2002	2005	2007
Sample	0.606	0.652	0.623
Population	0.599	0.653	0.613
	Positive proprietary income: Fraction in 1st decile		
	2002	2005	2007
Sample	10	9	8
Population	12	11	10
	Positive proprietary income: Fraction in 9th ant 10th decile		
	2002	2005	2007
Sample	41.2	48.7	43.3
Population	41.1	49.3	42.3

Note: The population and sample Gini coefficients plus the other summary measures are calculated by Statistics Norway. Household income is only available from 2004 and onwards.

As seen from the table - the Gini - coefficients are significantly lower in the sample than in the population. As stressed above, one important reason for this is that low income households typically do not own their house. But also the lower relative fraction of high income households in the sample explains the more compressed Gini-coeffcient in the sample. Finally, note that the company had a small relative increase in the fraction of high-income households from 2004 to 2007. For proprietary income, the correspondence between the sample and the population is quite accurate apart from a somewhat lower fraction of people with low proprietary in the sample.

Appendix 2. More about zero claims

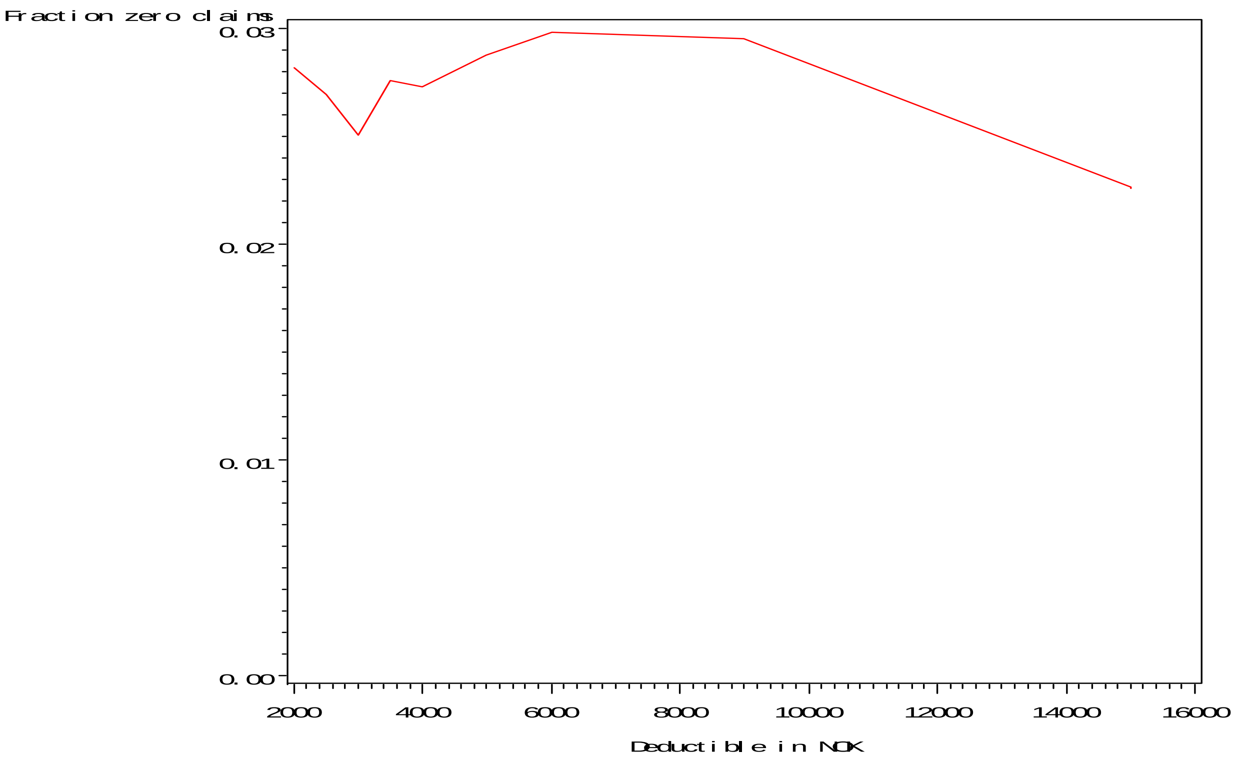
Recall that a zero claim is defined as a reported and registered claim that either is outside the cover of the specific contract or where the claim amount is lower than the deductible specified in the contract. Thus, a zero claim represents a potential claim and it should therefore be counted because it reveals important information about the underlying risk in the contract. Chiappori (2000) pp. 283 recognizes that "Although accidents involving no claims are generally not observed, adequate econometric techniques can be used". However, in this sample I observe "zero claims" and I will in the following give some important information about these.

The insurer that has provided the data registers all claims in it files irrespective of the size of the claim. The registration is an important part of the claims handling process in the company. Moreover and according to the product specialists - the insurance contract specifies that the insured is obliged to report an event without any unfounded delay. Of higher importance, however, is to what extent a customer has economic incentives to do so. The incentive to report a claim hinges on more than just the size of the deductible. First, if we assume that the insured has perfect knowledge about the deductible and, moreover, he/she can assess the likely indemnity caused by the event, then one should expect that reporting of claim events declines with the chosen deductible (holding the claim amount fixed). However, because claims within home insurance are rarer compared to for example car insurance it is likely that the insured often is ignorant regarding the deductible amount. Moreover, it is difficult to assess the seriousness of the event for most of the insured. For example, a water leakage may or may not have caused to damages to the structure, which is hidden for the insured.

Thus, one should expect that many customers prefer to contact the insurance company regardless the deductible choice or whether the event is considered small or large. Recall that there is no bonus-malus system within the home insurance market in Norway, a fact that increases the likelihood for "true" reporting of events.

Figure A4 pictures the fraction of zero claims in the dataset along the deductible menu that ranges from 2000 NOK to 15000 NOK.

Figure A4. **The fraction zero claims as a function of the chosen deductible**



There are two main points worth to emphasize considering the relation between zero claims and the deductible choice.[45] First, the percentage of zero claims is between 2.3 percent (15000 as deductible) and 2.9 percent (6000 as deductible). Thus, a significant fraction of zero claims is reported irrespective of the deductible choice. Second, the likelihood of zero claims increases in the range 2000 to 6000 and thereafter it declines.

Assume now that no insured perfectly can assess how serious an event is and therefore will contact the insurance company for advice. Then, we would expect that the probability for a zero claim will increase in the deductible menu. This is exactly what Figure A1 suggests for the range from 2000 to 9000 NOK. A t-test for whether the fraction of zero claims is equal for the 9000 deductible (N=33752) and the 3000 deductible (N=48850) is rejected (t-value 3.73). Thus, there is a significant higher fraction of zero claims in the 9000

[45] Note here that the percentage of contracts that specifies 2500 or 3500 NOK as a deductible is 2.17 and 0.22 respectively. This can explain the rather erratic picture at the low end of the deductible menu.

group than in the 3000 group. The table below show similar t-test for all the other interesting pairs:

Table A2. T-test of the difference in zero claim probability between pairs of deductibles.

T-value based on Satterthwaite test assuming unequal variances

Equality of zero claims	T-value	Implies
2000 vs 9000	-0.44	Lower fraction of zero claims in the 2000 group. Not significant.
3000 vs 9000	-3.64	Lower fraction of zero claims in the 3000 group. Significant.
4000 vs 9000	-1.99	Lower fraction of zero claims in the 4000 group. Significant.
5000 vs 9000	-0.71	Lower fraction of zero claims in the 5000 group. Not significant.
6000 vs 9000	0.28	Higher fraction of zero claims in the 6000 group. Not significant.

The zero-claim analysis show gives two important messages. First, the fraction of zero claims constitutes a significant amount in all deductible groups. Second, the probability of a zero claim increases in the deductible for the range from 2000 to 9000. Both these findings suggest that missing claims due to strategic reporting is low in these data.

Appendix 3: The relation between deductibles and premiums

Calculation of premiums departs from a advanced formula where characteristics of the object, the individual and the household matter. For purposes of this paper, however, we need only to consider the connection between deductibles and premiums. The premium is a function of a number of tariff factors and the deductible. Denote the tariff factors as *X* and the deductible as *D*. Then a general expression for the premium for individual *i* in year *t* for the particular cover *c* can be written as

(1) $$P_{ict} = f(X_{ict}, D_{ck}),$$

where X_{ict} is the characteristics of customer i who buy a cover c in year *t*. Here, D_{cK} denotes element *k* in the vector of possible deductibles and *f()* is a function that maps the different risk factors into a monetary premium. Altogether 9 possible deductibles are available for the customer, ranging from 2000 NOK to 15 000 NOK. Theoretically, the deductible menu can change from one year to another. However, this is not the case in our data and therefore we skip the subscript *t* from the formula in equation (1). The deductible choice affects the premium in a multiplicative way. Let $P^F{}_{ict}$ be the default premium if the customer choose the default deductible D_{cF}. Then, one can write the premium for another deductible, say D_{cK} as

(2) $$P^K{}_{ict} = P^F{}_{ict}\delta,$$

I

where $\delta > 1$ if $D_{cK} < D_{cF}$ and $\delta < 1$ if $D_{cK} > D_{cF}$. For deductibles less than D_{cF} this relation is always true. For deductibles higher than D_{cF} the formula (2) can be written as

(3) $$P^K{}_{ict} = \max\left[P^F{}_{ict}\delta, P^F{}_{ict} - M^K\right],$$

where *M* is the maximum allowed discount for the cover *c*. This ceiling kicks in at $\frac{M^K}{1-\delta}$.

Appendix 4. The robustness of the information problem

In order to test the robustness of the results above I have conducted several robustness tests. These are conducted by running the biprobit specification given by equation (4) under different data set restrictions. The results of these tests are presented in this section.

First, separate regressions for men and women are conducted. Second, I have run the biprobit for three distinct time periods, 2002-2003, 2004-2005 and 2006-2007. I have also conducted regressions where multi-contract customers are taken out and secondly - where both multi-contract and repeat customers are taken out. Thereafter, I conduct separate regressions within insurance value categories. Next, I split the sample into age categories and conduct separate analysis for each category. Furthermore, I ask whether the results are sensitive by replacing insurance value with the log-insurance value. The latter test is motivated by the possibility that the pricing of home insurance is non-linear in insurance value. Finally, I run the biprobit only including claims larger than twice the highest deductible. All these tests are conducted for a) - all customer vintages - and b) – only first contract year. The results of these robustness tests are presented in table A3.

Table A3. ***Various robustness tests. Biprobit.***

Claim equation : ***Having at least one claim higher than highest deductible***

Deductible equation: ***Deductible higher than 6000 NOK***

	All contract years			*First contract year*		
	Rho	Chi-square	P-value	Rho	Chi -square	P-value
Men	-0.0280	8.209	0.0042	-0.0236	2.295	0.1298
Women	-0.0555	10.183	0.0014	-0.0698	6.247	0.0124
Only 2002-2003	-0.0375	2.742	0.0977	-0.0284	1.759	0.1846
Only 2004-2005	-0.0350	5.816	0.0159	-0.0437	3.093	0.0786
Only 2006-2007	-0.0337	8.362	0.0038	-00306	2.223	0.1359
Only single contract customers	-0.0290	9.266	0.0023	-0.0425	7.655	0.0057
Only single contract and no-repeat customers	-0.0280	7.339	0.0067	-0.0404	5.509	0.0189
Insurance value < 1mill	0.0108	0.017	0.8948	-0.0035	0.009	0.9570
1-2 mill	-0.0101	0.308	0.5787	0.0124	0.217	0.6410
2-3	-0.0397	9.081	0.0026	-0.0709	10.63	0.0011
4-5	-0.0556	10.39	0.0013	-0.0516	3.087	0.0789
Above or equal 5 mill	-0.0185	0.572	0.4494	0.0231	0.303	0.5820
Age <30	-0.0204	0.543	0.4611	-0.0585	2.157	0.1419
30-40	-0.0345	4.856	0.0275	-0.0243	1.020	0.3125
40-50	-0.0419	6.314	0.0120	-0.0559	4.322	0.0376
Above 50	-0.0363	6.471	0.0110	-0.0260	1.146	0.2843
Log insurance value	-0.0348	16.65	0.0000	-0.0342	6.321	0.0119
Only claims>2 *highest deductible	-0.0199	3.544	0.0597	-0.0191	1.296	0.2549

Only the correlation coefficient, its associated chi-square statistic and the probability value is displayed. The estimated model contains exactly the same variables as in table 7.

The separate regressions for men and women reveal that the asymmetric information problem is more pronounced among women than among men. This is especially clear when one looks at new female customers. Apparently, the deductible choice among women is a strong signal of ex-post claim frequency.

The separate regressions for the different time periods reveal that the information problem has not been constant over time. It was particularly pronounced in 2004 and 2005. The same pattern is found for first year contracts.

The information problem is not particularly affected by removing multi-contract customers or repeat customers.[46]

The biprobit regressions for different categories of insurance value reveal a quite interesting pattern. For insurance values less than 1 million the information problem is practically non-existent. The problem increases in magnitude for higher insurance values up

[46] A repeat customer is identified by gaps in the contract history. For example, a customer may be insured in 2002, out of the sample in 2003, and observed in 2004 with the same object as in 2002.

to approximately 5 MNOK. This pattern is consistent with a hypothesis of a slight “under pricing“of middle to high insurance values.

The information problem increases with age up till around 50 and thereafter remains stable. However, the sharpness of the correlations coefficient reaches the peak for the age category that includes customers from 40 to 49. The second last row in table A3 reveals that the information problem is not affected by the functional form between insurance value and the outcome variables.

In the last row I include only claims higher than 2 times the deductible. This restriction leads to a clear reduction of the information problem, and the correlation coefficient is significant at a 6 percent level for all contracts, while the probability value is 25 percent for first year contracts.

Norges Handelshøyskole

Norwegian School of Economics
and Business Administration

NHH
Helleveien 30
NO-5045 Bergen
Norway

Tlf/Tel: +47 55 95 90 00
Faks/Fax: +47 55 95 91 00
nhh.postmottak@nhh.no
www.nhh.no