Data in Brief 39 (2021) 107639

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Case study data for joint modeling of insurance claims and lapsation

Montserrat Guillen[a,*], Catalina Bolancé[c], Edward W. Frees[b], Emiliano A. Valdez[c]

[a] Universitat de Barcelona, Spain
[b] University of Wisconsin-Madison and Australian National University, United States
[c] University of Connecticut, United States

ARTICLE INFO

Article history:
Received 3 September 2021
Revised 7 November 2021
Accepted 22 November 2021
Available online 26 November 2021

Keywords:
Motor insurance
Homeowners insurance
Customer retention
Loyalty
Ratemaking
Premium
Loss data
Dependence
Heavy tails

ABSTRACT

The dataset tracks 40,284 insurance clients over five years, between 2010 and 2015, who subscribed to both automobile and homeowners insurance. We have combined information on these customers. First, the characteristics including age, gender or driving experience, among others and dates of renewal for the two types of policies considered here. Note that we have only considered clients corresponding to persons and not commercial firms that can also underwrite home and motor insurance policies. Second, the policy data file for motor vehicle insurance consists of all vehicle insurance coverage including power, driving area or whether there is a second driver that drives the car occasionally. Third, the policy data file for homeowners insurance has information on the property such as value of the building (essentially the value of the home without any furniture, apparel and personal items), location and type of dwelling. Besides these three sources, we have access to data containing information on the number of claims and total cost of those claims per year and per policy type. So, for all policies that are in force, we finally have up to a five year record of the yearly

DOI of original article: 10.1016/j.eswa.2021.115552
* Corresponding author.
E-mail addresses: mguillen@ub.edu (M. Guillen), bolance@ub.edu (C. Bolancé), jfrees@bus.wisc.edu (E.W. Frees), emiliano.valdez@uconn.edu (E.A. Valdez).
Social media: (M. Guillen), (C. Bolancé), (E.A. Valdez)

https://doi.org/10.1016/j.dib.2021.107639

cost of claims in the motor insurance and in the home coverage. If the customer does not renew one of those two policies or both, we do not have more information after this lapse occurs. After summarizing the data, we provide the usual marginal analysis, where we fit regression models using Tweedie distributions for claims and a logistic model for lapse. Data can be used for joint analysis of insurance policyholders with more than one product.

Specifications Table

Subject	Business Intelligence and Data Warehousing
Specific subject area	Insurance analytics and risk analysis, dependent sources of risk, ratemaking
Type of data	Table
How data were acquired	Random sample of real insurance portfolio. A Spanish insurance company provided a sample of insured policyholders for the purpose of this research. A representative sample of all customers who had homeowners and motor insurance were included in the initial sample In 2010. There was a window of observation until 2015.
Data format	Raw
Parameters for data collection	A sample of customer in 2010 was selected and they were followed for five years. Only private customers were included of a major Spanish insurer. Some new policy holders enter the sample and some other lapse and leave the sample.
Description of data collection	Customers were selected randomly from a real insurance portfolio.
Data source location	Institution: University of Barcelona City/Town/Region: Barcelona Country: Spain Latitude and longitude (and GPS coordinates, if possible) for collected samples/data: Spain
Data accessibility	Observational data - Part of the data included in this article is a real dataset. Repository name: Mendeley Data identification number: DOI: 10.17632/vfchtm5y7j.1 Direct URL to data: https://data.mendeley.com/drafts/vfchtm5y7j Simulated data - Part of the data included in this article is simulated data. Data identification number: DOI: 10.5281/zenodo.5425210 Direct URL to code: https://github.com/ewfreesRes/Hybrid-Insurance-with-Dropout
Related research article	E. W. Frees, C. Bolancé, M. Guillen, E. A. Valdez (2021) Dependence Modeling of Multivariate Longitudinal Hybrid Insurance Data with Dropout, Expert Syst. Appl. 185, 115552 https://doi.org/10.1016/j.eswa.2021.115552

Value of the Data

- This unique real insurance dataset covers five years and customers who have more than one type of policy contract and may possibly lapse one or more policies.
- Researchers studying insurance ratemaking of more than one line of business or ratemaking dynamics can benefit from these data.
- These data may be reused to gain insight into modern artificial intelligence modelling in the area of insurance.
- Given that the information on claims is composed of frequency and cost, the data provide a way to analyse them jointly or separately.

1. Data Description

As in many countries, in Spain owners of automobiles are obliged to have some minimum form of insurance coverage for personal injury to third parties. Home insurance is optional. The reasons why citizens decide to have these two types of insurance may be quite different. So, we believe that our sample of customers having the two is not representative of the whole population, because not everyone selects to buy homeowner coverage or can afford to buy it. Even if ownership is vastly extended in Spain, and one may think that motor and home insurance should go together, home insurance coverage is often linked to a mortgage and, so it is not necessarily sold by the same insurance company that is covering motor insurance. In the recent years, many insurers are trying to cross-sell in their existing portfolios, so they have made an enormous effort to identify and offer a home insurance to those having only motor insurance and the other way around.

The dataset tracks 40,284 insurance clients over five years, between 2010 and 2014, who subscribed to both automobile and homeowners insurance. We have combined information on these customers. First, the characteristics including age, gender or driving experience, among others and dates of renewal for the two types of policies considered here. Note that we have only considered clients corresponding to persons and not commercial firms that can also underwrite home and motor insurance policies. Second, the policy data file for motor vehicle insurance consists of all vehicle insurance coverage including power, driving area or whether there is a second driver that drives the car occasionally. Third, the policy data file for homeowners insurance has information on the property such as value of the building (essentially the value of the home without any furniture, apparel and personal items), location and type of dwelling. Besides these three sources, we have access to data containing information on the number of claims and total cost of those claims per year and per policy type. So, for all policies that are in force, we finally have up to a five year record of the yearly cost of claims in the motor insurance and in the home coverage. If the customer does not renew one of those two policies or both, we do not have more information after this lapse occurs.

There is only one data file containing 122935 records. Each customer is uniquely identified by the policy identification number (PolID). Year of observation is recorded in column named "year" and goes from 1 to 5. Records are presented ordered by years and by policy number within each year. Given that some policy holders do not stay in the insurance company, the number of policies per year decreases over time.

The description of the data is the following: "gender" is 1 for male and 0 for female), "Age_client" is the age of the customer in years, "age_of_car_M" I s the number of years since the vehicle was bought by the customer, "Car_power_M" is the power of th vehicle, "Car_2ndDriver_M" equals 1 if the customer has informed the insurance company that a second occasional driver uses the vehicle, and 0 otherwise, "num_policiesC" is the total number of policies held by the same customer in the insurance company, "metro_code" equals 1 for urban or metropolitan and 0 for rural, "Policy_PaymentMethodA" equals 1 for annual payment and 0 for monthly payment in the motor policy, "Policy_PaymentMethodH" equals 1 for annual payment and 0 for monthly payment in the homeowners policy, "Insuredcapital_content_re" is the value of content in homeowners insurance, "Insuredcapital_continent_re" is the value of building in homeowners insurance, "appartment" equals 1 if the homeowners insurance correspond to an apartment and 0 otherwise "Client_Seniority" is the number of years that the customer has been in the company, "Retention" equals 1 if the policy is renewed and 0 otherwise,"NClaims1" is the number of claims in the motor insurance policy for the corresponding year, "NClaims2" is the number of claims in the motor insurance policy for the corresponding year, "Claims1" is the sum of claims cost in the motor insurance policy for the corresponding year,"Claims2" is the sum of claims cost in the homeowners insurance policy for the corresponding year, "Types" is equal to 1 when neither an auto nor a home claim, it is equal to 2 when the customer has an auto but not a home claim, it is equal to 3 when the customer does not have not an auto but

a home claim and it is equal to 4 when both an auto and a home claim. All monetary units are expressed in Euros. In motor insurance, only claims at fault are considered.

2. Experimental Design, Materials and Methods

Consider the case where we follow policyholders over time. During the year, there are three outcome variables of interest (more details on the data can be found in [1]). The claims outcomes are

Y_1 which represents claims from an auto coverage and
Y_2 which represents claims from a homeowners coverage.

As claims outcomes, these variables may take on value of zero (representing no claim) and are otherwise positive continuous outcomes (representing claim amount). We use subscripts i to distinguish among policyholders and t to distinguish observations over time. Thus, $Y_{1,it}$ and $Y_{2,it}$ represent auto and homeowner claims for the ith policyholder at time t. The third random variable, L, is a binary variable that represents a policyholder's decision to lapse one of the policies. Specifically, L_{it} equal to 1 indicates that the i-th policyholder in the t-th year decides to lapse one of the policies and, 0 indicates that the i-th policyholder in the t-th year decides to not lapse the two policies, i.e, to renew. Note that if $L_{it} = 1$ then we do not observe the policy at time t+1. In the same way, if $L_{it} = 0$, then we observe the policy at time t+1, subject to limitations on the number of time periods available. We use m to represent the maximal number of observations over time.

Associated with each policyholder is a set of (possibly time varying) rating variables x_{it} for the i-th policyholder at time t that is described below. We represent the marginal distribution of each outcome variables in terms of a generalized linear model. Specifically, following standard insurance industry practice, we represent the marginal distributions of the claims random variables using a Tweedie distribution so that the distributions have a mass at zero and are otherwise positive. The marginal distribution of the renewal variable is modeled using a logit function. Marginal distributions may use common rating variables and so are naturally related in this sense.

Lapsation dictates the availability of data which may be related to the outcomes, a violation of the statistical principle known as *missing at random*. This means that analyzing claims while ignoring lapse can lead to biased estimation. Thus, joint modeling of lapse and claims are critical because the claims model depends on the data observed through the lapsation/renewal process.

The R code to obtain the data summary is presented in Appendix A and the outcomes are presented in Table 1.

Marginal model estimation is typically done assuming that each year has the same set of parameters and that observations from different years are independent. This is not necessary

Table 1
Summary of the data outcome by year.

	2010	2011	2012	2013	2014
Number of observations	40284	29818	22505	17044	13284
Number of lapses	10466	7313	5461	3780	2296
Proportion of Lapses	0,26	0,25	0,24	0,22	0,17
Number of clients with positive motor claims	769	547	318	209	124
Average number of motor claims	0,04	0,03	0,03	0,02	0,02
Average cost of motor claims	1539.99	1689,84	2031.20	1629.18	1222.13
Number of clients with positive home claims	660	531	448	310	240
Average number of home claims	0.03	0.03	0.04	0.04	0.03
Average cost of home claims	447.85	501.59	410.73	348.10	508.86

realistic but provides a convenient starting point. The Tweedie model is commonly used in insurance applications for claims. In part, this is because it can be expressed as a generalized linear model. For type 1 (auto) claims model, we first need to find an initial parameter (p) for the Tweedie model, which is related to the variance function of the total cost of claims (see [2]). In order to make this procedure faster, we work with a random sample and then we proceed to estimate the optimal parameter with the whole sample.

For the Tweedie model we assume that the number of claims (frequency) follows a Poisson distribution:

$$\mu_{j,it} = exp\left(x_{it}{}'\beta_j\right),\ \ j = 1, 2,$$

while the cost of each claim (severity) is Gamma distributed and independent of the number of claims. The R code is shown in Appendix B for motor insurance and in Appendix C for homeowners' insurance. We then estimate the p parameter for the Tweedie model with the whole sample and find the optimal parameter which has minimum deviance. We also find the other parameters by maximum likelihood and the model fit. Note that the random samples used to search for the optimal p are simple random samples with replacement of size 10,000 from the original sample. Note that a seed number is selected in order to preserve reproducibility.

To model lapse, we employ a simple logistic regression (marginal) model. Thus,

$$Prob(L_{it} = 1) = exp\left(x_{it}{}'\beta_L\right)/\left(1 + exp\left(x_{it}{}'\beta_L\right)\right).$$

The R code is presented in Appendix D.

A comparison of GMM Lapse estimators is included in [2] and it uses simulated data. We generate samples of n policyholders observed over a maximum of 5 years. For the number of claims of each type (auto and homeowners) rating (explanatory) variables are considered: (a) x1, a binary variable that indicates whether or not an attribute holds, (b) x2, x3, x4, continuous explanatory variables, and (c) x5 a time trend. Simulation also requires generating claims severities for two lines of insurance (motor and homeowners). Total cost for each type of claims is simulated using the Tweedie distribution, a mean (μ), and two other parameters, ϕ_j (for dispersion) and p (the power parameter). So, the variance of total claim cost equals $\phi_j \mu^p$, $j = 1, 2$. We assume that mean cost is 1,000, $\phi_1 = \phi_2 = 500$ and that p=1.67. Correlation between each type of claim and lapse is also considered with a Gaussian copula, where correlation between lapse and, motor insurance and homeowners claims is set to -0.1 and 0.2, respectively. The correlation between motor and homeowners claims equals 0.2. . Samples of size 100; 250, 500 1,000 and 2,000 are generated. The number of replicates equals 500 for samples of size 100 and 250, and 100 for samples of size larger than 250. More details and extended simulation information can be found in [3].

Ethics Statement

Data are anonymized and adhere to all ethical requirements for publication in *Data in Brief*. Authors declare that they comply with reporting standards, data access and retention, originality and acknowledgement of sources and confidentiality. The authors declare no competing interests.

CRediT Author Statement

Montserrat Guillen: Conceptualization, Methodology, Software, Supervision, Validation, Writing- Original draft preparation, Writing- Reviewing and Editing. **Catalina Bolancé**: Conceptualization, Methodology, Software, Data curation, Writing- Reviewing and Editing. **Edward W. Frees:** Writing- Original draft preparation Visualization, Investigation, Supervision, Validation, Writing- Reviewing and Editing. **Emiliano A. Valdez:** Supervision, Validation, Writing- Reviewing and Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgments

We wish to express our thanks to the Spanish Ministry of Science, Fundación BBVA ayudas Big Data and ICREA Academia for their support.

Appendix A

R code to summarize the data by year.

```
SumClaims = matrix(0,9,5)
colnames(SumClaims) <- c("F2010","F2011","F2012","F2013","F2014")
rownames(SumClaims) <- c("Number of Obs","Number of Lapse","Proportion of Lapses",
"Number of Clients with positive Claims 1",
"Average Number of Claim 1","Average cost of Claim 1",
"Clients with positive Claims 2",
"Average Number of Claim 2","Average cost of Claim 2")
SumClaims[1,] <- aggregate(Lapse ~ year, data=SampleData, length)$Lapse
SumClaims[2,] <- aggregate(Lapse ~ year, data=SampleData, sum)$Lapse
SumClaims[3,] <- SumClaims[2,]/SumClaims[1,]
SumClaims[4,] <- aggregate(PosClaim1 ~ year, data=SampleData, sum)$PosClaim1
SumClaims[5,] <- aggregate(NClaims1 ~ year, data=SampleData, mean)$NClaims1
SumClaims[6,] <- aggregate(Claims1 ~ year, data=SampleData, sum)$Claims1/SumClaims[4,]
SumClaims[7,] <- aggregate(PosClaim2 ~ year, data=SampleData, sum)$PosClaim2
SumClaims[8,] <- aggregate(NClaims2 ~ year, data=SampleData, mean)$NClaims2
SumClaims[9,] <- aggregate(Claims2 ~ year, data=SampleData, sum)$Claims2/SumClaims[7,]
knitr::kable(SumClaims,digits=2, caption="Lapse and Claims Summary by Year")
```

Appendix B

R code for fiiting Tweedie model in motor insurance data.

```
# Randomly re-order data - "shuffle it"
n <- nrow(SampleData)
set.seed(12347)
shuffled_SampleData <- SampleData[sample(n), ]
subset_SampleData <- shuffled_SampleData[1:10000,]
out1 <- tweedie.profile(Claims1 ~year+Age_client+
Client_Seniority+metro_code+Car_power_M+Car_2ndDriver_M+
Policy_PaymentMethodA,data=subset_SampleData,
xi.vec=seq(1.1, 1.6, length=10), do.plot=TRUE)

funtweedie<-function(p){glm(Claims1 ~year+Age_client+
Client_Seniority+metro_code+Car_power_M+Car_2ndDriver_M+
Policy_PaymentMethodA,data=SampleData,

control = glm.control(maxit = 200),
family=tweedie(var.power=p, link.power=0))$deviance}
deviance_l_p_sample<-funtweedie(out1$xi.max-0.01)
deviance_p_sample<-funtweedie(out1$xi.max)
deviance_u_p_sample<-funtweedie(out1$xi.max+0.01)
p<-out1$xi.max-0.01
if(deviance_l_p_sample<deviance_p_sample)
{while(deviance_l_p_sample<deviance_p_sample){
deviance_p_sample<-deviance_l_p_sample
deviance_l_p_sample<-funtweedie(p-0.01)
p<-p-0.01}
} else{
deviance_p_sample<-funtweedie(out1$xi.max)
p<-out1$xi.max+0.01
while(deviance_u_p_sample<=deviance_p_sample){
deviance_p_sample<-deviance_u_p_sample
deviance_u_p_sample<-funtweedie(p+0.01)
p<-p+0.01}}

tweedie.fit <- glm(Claims1 ~year+Age_client+Client_Seniority+
metro_code+Car_power_M+Car_2ndDriver_M+
Policy_PaymentMethodA,data=SampleData,
control = glm.control(maxit = 200),
family=tweedie(var.power=p, link.power=0))
sum.tweedie.fit1 <- summary(tweedie.fit1)
knitr::kable(coefficients(sum.tweedie.fit1),digits=3,
caption="Tweedie Claims 1 (Auto) Model Summary")

dfTweedie1A <- ptweedie(SampleData$Claims1,xi=p,
mu=tweedie.fit1$fitted.values,phi=summary(tweedie.fit1)$dis)
SampleData$dfTweedie1 <- pmin(pmax( 1e-05,dfTweedie1A),.99999)
```

Appendix C

R code for fitting Tweedie model in homeowners insurance data.

```
# Randomly re-order data - "shuffle it"
n <- nrow(SampleData)
set.seed(12347)
shuffled_SampleData <- SampleData[sample(n), ]
subset_SampleData <- shuffled_SampleData[1:1000,]
out2 <- tweedie.profile(Claims2 ~year+Age_client+Client_Seniority+metro_code+
Insuredcapital_continent_re+appartment+
Policy_PaymentMethodH,data=subset_SampleData,
xi.vec=seq(1.1, 1.9, length=10), do.plot=TRUE)

funtweedie<-function(p){glm(Claims2 ~year+Age_client+
Client_Seniority+metro_code+
Insuredcapital_continent_re+appartment+
Policy_PaymentMethodH,data=SampleData,
control = glm.control(maxit = 200),
family=tweedie(var.power=p, link.power=0))$deviance}
deviance_l_p_sample<-funtweedie(out2$xi.max-0.01)
deviance_p_sample<-funtweedie(out2$xi.max)

deviance_u_p_sample<-funtweedie(out2$xi.max+0.01)
p<-out2$xi.max-0.01
if(deviance_l_p_sample<deviance_p_sample)
{while(deviance_l_p_sample<deviance_p_sample){
deviance_p_sample<-deviance_l_p_sample
deviance_l_p_sample<-funtweedie(p-0.01)
p<-p-0.01}
} else{
deviance_p_sample<-funtweedie(out2$xi.max)
p<-out2$xi.max+0.01
while(deviance_u_p_sample<=deviance_p_sample){
deviance_p_sample<-deviance_u_p_sample
deviance_u_p_sample<-funtweedie(p+0.01)
p<-p+0.01}}

tweedie.fit2 <- glm(Claims2 ~year+Age_client+Client_Seniority+metro_code+
Insuredcapital_continent_re+appartment+
Policy_PaymentMethodH,data=SampleData,
control = glm.control(maxit = 200),
family=tweedie(var.power=p, link.power=0))
sum.tweedie.fit2 <- summary(tweedie.fit2)
knitr::kable(coefficients(sum.tweedie.fit2),digits=3,
caption="Tweedie Claims 2 (Home) Model Summary")

dfTweedie2A <- ptweedie(SampleData$Claims2,xi=p,
mu=tweedie.fit2$fitted.values,phi=summary(tweedie.fit2)$dis)
SampleData$dfTweedie2 <- pmin(pmax( 1e-05,dfTweedie2A),.99999)
```

Appendix D

R code for fitting the lapsing logistic model to insurance data.

```
logistic.fit <- glm(Lapse ~ year+gender+Age_client+
Client_Seniority+metro_code,data=SampleData,
control = glm.control(maxit = 50),family=binomial(link=logit))
sum.logistic.fit <- summary(logistic.fit)
knitr::kable(coefficients(sum.logistic.fit),digits=3,
caption="Logistic Lapse Model Summary")
```

References

[1] M. Guillen, C. Bolancé, E.W. Frees, E.A. Valdez, Insurance data for homeowners and motor insurance customers monitored over five years, v1, 2021. https://doi.org/10.17632/vfchtm5y7j.1.
[2] E.W. Frees, C. Bolancé, M. Guillen, E.A. Valdez, Dependence modeling of multivariate longitudinal hybrid insurance data with dropout, Expert Syst. Appl. 185 (2021) 115552.
[3] [simulated data] E. W. Frees, C. Bolancé, M. Guillen, E. A. Valdez, Code to support the paper "dependence modeling of multivariate longitudinal hybrid insurance data with dropout", 2021. https://doi.org/10.5281/zenodo.5425210.