`"FireProp"` Fire and other property damage premium density. It includes damage or loss of property due to fire, explosion, storm, other natural forces, nuclear energy and land subsidence as well as other damage to property.

`"PecLoss"` Pecuniary loss premium density. It includes credit loss, surety loss and other miscellaneous financial losses.

`"GenLiab"` General liability premium density. It includes all liability other than motor vehicle, aircraft and ship liability.

`"AccSick"` Accident and sickness premium density.

`"OtherNL"` Other non-life premium density. It includes legal expenses, assistance and other miscellaneous insurance.

`"MRATE"` Motor vehicle ownership per capita.

`"NumAcc"` ?

`"Population"` Total population number.

Source

FreesBook-LPD

References

Browne, M. J., Chung, J. and Frees, E. W. (2000). *International property-liability insurance consumption*. Journal of Risk and Insurance, 73-90.

Frees, E. W. (2004). *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge University Press.

Examples

```
# (1) load of data
#
data(PnCdemand)
```

`pricingame` *French Motor Third-Part Liability datasets used for 100 percent Data Science game*

Description

`pg15training`, `pg15pricing` are the two datasets used for the 2015 pricing game of the French institute of Actuaries organized on November 5, 2015. `pg15training` contains 100,000 TPL policies for private motor insurance used to fit the models, whereas `pg15pricing` contains 36,311 policies of the same guarantee for which the premium is computed. Each record has been observed at most one year and contains risk features of the policyholder and the insured vehicle. For confidentiality reasons, most categorical levels have unknown meaning.

`pg16trainpol`, `pg16trainclaim`, `pg16test` are the three datasets used for the 2016 pricing game of the French institute of Actuaries organized on November 8, 2016. `pg16trainpol` contains 87,228 policies for private motor insurance and `pg16trainclaim` contains 4,568 claims of those 87,228

TPL policies. Policies are guaranteed for all kinds of material damages, but not bodily injuries. Both datasets are used to fit the models, whereas `pg16test` is used for training. For confidentiality reasons, most categorical levels have unknown meaning.

`pg17trainpol`, `pg17trainclaim` are the two training datasets used for the 2017 pricing game of the French institute of Actuaries organized on November 16, 2017. `pg17trainpol` contains 100,000 policies for private motor insurance and `pg17trainclaim` contains 14,243 claims of those 100,00 TPL policies. These training sets correspond to year $t = 0$. `pg17testyear1`, `pg17testyear2`, `pg17testyear3`, `pg17testyear4` are the four test datasets used for the pricing game: each has 100,000 rows of new policies (drivers willing to purchase insurance for Year t with $t = 1, 2, 3, 4$).

Usage

```
data(pg15training)
data(pg15pricing)

data(pg16trainpol)
data(pg16trainclaim)
data(pg16test)

data(pg17trainpol)
data(pg17trainclaim)
data(pg17testyear1)
data(pg17testyear2)
data(pg17testyear3)
data(pg17testyear4)
```

Format

`pg15training` and `pg15pricing` are two dataframes with the same columns:

`PolNum` The policy number.

`CalYear` The underwriting year.

`Gender` The gender of the car driver.

`Type` The car type (a single letter).

`Category` The car category (a string character).

`Occupation` The occupation of the driver (a string character).

`Age` The driver age, in years (in France, people can drive a car at 18).

`Group1` The group of the car.

`Bonus` The bonus-malus (French no-claim discount): `-30` means a 30 percent bonus while `+20` means a 20 percent malus; see details below.

`Poldur` The policy age (in year).

`Value` The car value (in euro).

`Adind` A dummy variable indicating a material cover.

`SubGroup2` The subregion of the driver home (unknown category).

`Group2` The region of the driver home (unknown category).

`Density` The density of inhabitants (number of inhabitants per km2) in the city the driver of the car lives in.

`Expdays` Exposure in days.

`Numtppd` The number of third-party material claims.

`Numtpbi` The number of third-party bodily injury claims.

`Indtppd` The total cost of third-party material claims (euro).

`Indtpbi` The total cost of third-party bodily injury claims (euro).

`pg16trainpol`,`pg16trainclaim`,`pg16test` are dataframes with the following columns:

`Year` The coverage year.

`BeginDate`,`EndDate` Beginning date and ending date of the coverage period (of class `"Date"`).

`Exposure` The exposure as a fraction of year, computed as the difference between `EndDate` and `BeginDate` divided by 365.

`PolicyID` The identification number of the policy.

`PolicyAgeCateg` The category of the policy age.

`PolicyCateg` The category of the policy.

`CompanyCreation` A dummy indicating if the company has been created.

`FleetMgt` The fleet management category.

`FleetSizeCateg` The fleet size category

`Area` The geographical area.

`PayFreq` The payment frequency.

`VehiclAge` The vehicle age category.

`VehiclNb` The number of vehicles

`VehiclCateg` The vehicle category.

`VehiclPower` The vehicle power

`LicNb` The license number of the vehicle.

`Deduc` The deductible category

`SumInsured` The category of the sum insured.

`BusinessType` The business type.

`ChannelDist` The distribution channel.

`ClaimNb` The claim number.

`ClaimCharge` The claim charge.

`DirectComp` As claims correspond only to material damage, the French claim convention (IDA) was applied. So the insurer may directly refund the insured (when `DirectComp=TRUE`) even if the insurer will sue the third-party insurer to recover the indemnity afterwards.

`CompRate` The rate of compensation (in percent).

`SettlYear` The settlement year.

`pg17***` are dataframes with the following columns:

`id_client` The client identification number: a string of the form `Annnnnnnn` (`A` followed by an 8-digit number). First client ID is `A00000001` and last is `A00091488`.

`id_vehicle` The vehicle identification number: a string of the form `Vnn` (a `V` followed by a 2-digit number). First vehicle is always numbered V01. If a client has multiple vehicles, then the numeration increases by 1. There is no particular ordering in the vehicles, so their rank should not represent anything valuable.

`id_policy` The policy identification number, a string of the form `Annnnnnnn-Vnn` resulting from appending `id_client` and `id_vehicle`.

`id_year` The year of coverage, Year ID begins at `"Year 0"` and ends at `"Year 4"`.

`pol_bonus` The policy bonus (French no-claim discount): `0.5` means a 30 percent bonus while `1.2` means a 20 percent malus; see details below.

`pol_coverage` The coverage category: The coverage are of 4 types : Mini, Median1, Median2 and Maxi, in this order. As you can guess, Mini policies covers only Third Party Liability claims, whereas Maxi policies covers all claims, including Damage, Theft, Windshield Breaking, Assistance, etc.

`pol_duration` The policy duration: Policy duration represents how old the policy is. It is expressed in year, accounted from the beginning of the current year i. Oldest policies in this portfolio can last since prehistoric ages of 45 years.

`pol_sit_duration` The policy current endorsement duration: Situation duration represent how old the current policy caracteristics are. It can be different from pol duration, because the same insurance policy could have evolved in the past (e.g. by changing coverage, or vehicle, or drivers, ...).

`pol_pay_freq` The payment frequency: The price of the insurance coverage can be paid annually, bi-annually, quarterly or monthly.

`pol_payd` A dummy indicating pay as you drive: a string with Yes or No, which indicates whether our client has subscribed a mileage-based policy or not. In those early ages of Year 0, Pay As You Drive was not that current, so they represent a minority in the portfolio.

`pol_usage` The policy usage: it describes what usage the driver makes from his vehicle, most of time. There are 4 possible values : `"WorkPrivate"` which is the most common, `"Retired"` which is presumed to be aimed at retired people (who also are presumed driving less kilometers), `"Professional"` which denotes a professional usage of the vehicle, and `"AllTrips"` which is quite similar to Professional (including pro tours). As for the coverage, it would be very surprising that this variable had no effect on frequency.

`pol_insee_code` The INSEE code of the French city/municipality where the policyholder lives: it is a 5-digits alphanumeric code used by the French National Institute for Statistics and Economic Studies (hence INSEE) to identify "communes" and departments in France. There are about 36,000 "communes" in France, but not every one of them is present in the dataset (there are only 18,000 of them). The first 2 digits of insee code identifies the department (they are 96, not including overseas departments). The insee code or department code can be used to possibly merge external data to the datasets: population density, OSM data, etc.

`drv_drv2` A character string indicating if there is a secondary driver: there is always a first driver, which characteristics (age, sex, licence) are provided, but a secondary driver is optional, and is present 1 time out of 3.

`drv_age1,drv_age2` The driver age of the ith driver: it is expressed in years counted from the beginning of the considered year. Then, `drv_age1` increases by 1 every year, like in real world... Legal age to drive is 18, so you shouldn't find any age below that limit. Due to the fact that the database is built on existing situations before Year 0, in fact the minimum age is 19 in Year 0 dataset. On the other side, you'll also find quite old drivers.

`drv_sex1,drv_sex2` The driver sex of the ith driver. European rules force insurers to charge the same price for women and men. But driver's gender can still be used in academic studies, and that's why drv sex1 is still available in the datasets, and can be used as discriminatory variable in this pricing game.

`drv_age_lic1,drv_age_lic2` The age of the driving license of the ith driver. As for the other ages, it is expressed in integer years from the beginning of the current year.

`vh_age` The vehicle age: This variable is the vehicle's age, the difference between the year of release and the current year.

`vh_cyl` The engine cylinder displacement is expressed in ml in a continuous scale. This variable should be highly correlated with din power of the vehicle.

`vh_din` The `vh_din` is a representation of the motor power. Highly correlated with din power, cylinder, speed and even value of the vehicle.

`vh_fuel` The vehicle fuel type: with mainly two values `"Diesel"` and `"Gasoline"`. Very few Hybrid vehicles can also be found, but, 6 years ago, the hybrid market was still at its beginning.

`vh_make` The vehicle carmaker. As the database is built from a French insurance, the three major brands are Renault, Peugeot and Citroen.

`vh_model` The vehicle model. As a subdivision of the carmake, vehicle is identified by its model name.

`vh_sale_begin,vh_sale_end` `vh_sale_begin` and `vh_sale_end` are the dates (in fact: ages) from the beginning of the current year of the beginning and the end of marketing years of the vehicle. This could for instance identify policies that covers very new vehicles or second-hand ones.

`vh_speed` The vehicle maximum speed (km/h), as stated by the manufacturer.

`vh_type` The vehicle type, either `"Tourism"` or `"Commercial"`. There are more `"Commercial"` types for `"Professional"` policy usage than for `"WorkPrivate"`.

`vh_value` The vehicle's value (replacement value) is expressed in euros, without inflation so it should be stable from a year to another.

`vh_weight` The vehicle weight (kg).

`id_claim` The claim identification number: a string of the form `CLnn` (`CL` followed by a 2-digit number). Numbering of the claims begins at 1 for every policy and each year. Then, the last value of id claim is the maximum number of claims for a vehicle in a year. Two-digits representation is sufficient : this maximum doesn't exceed 7 (but not on Year 0, where the maximum is 6).

`claim_nb` The claim number, as we are talking about individual claims, each claim nb has a value of 1.

`claim_amount` The claim amount: amounts range from (approx.) -2,000 to +300,000. Yes, there are negative values, they come from claims where our driver's liability is not engaged, so there's a legal recourse.

The bonus/malus system is compulsary in France, but we will only use it here as a possible feature. The coefficient is attached to the driver. It starts at 1 for young drivers (i.e. first year of insurance). Then, every year without claim, the bonus decreases by 5 percent until it reaches its minimum of 0.5. Without any claim, the bonus evolution would then be : $1 -> 0.95 -> 0.9 -> 0.85 -> 0.8 -> 0.76 -> 0.72 -> 0.68 -> 0.64 -> 0.6 -> 0.57 -> 0.54 -> 0.51 -> 0.5$. Every time the driver causes a claim (only certain types of claims are taken into account), the coefficient increases by 25 percent, with a maximum of 3.5. Thus, the range of bonus/malus coefficient extends from 0.5 to 3.5 in the datasets.

Source

Datasets from unknown private insurers.

See https://freakonometrics.hypotheses.org/20034 for the first pricing game.

See https://actinfo.hypotheses.org/69 for the second pricing game.

See https://actinfo.hypotheses.org/86 for the third pricing game.

Examples

```
# (1) load of data
#
data(pg15training)
data(pg15pricing)

data(pg16trainpol)
data(pg16trainclaim)
data(pg16test)


data(pg17trainpol)
data(pg17trainclaim)
data(pg17testyear1)


# (2) some check
# should be zero
sum(!pg16trainclaim$PolicyID %in% pg16trainpol$PolicyID)
# should be true
NROW(pg16trainclaim) == sum(pg16trainpol$ClaimNb)
```

---

sgautonb *Singapore Automobile claim count dataset*

Description

This dataset contains automobile injury claim number collected in 1993 in Singapore by the General Insurance Association of Singapore. Records contains individuals characteristics in addition to claim counts.

Usage

```
data(sgautonb)
```

Format

sgautonb is a data frame of 8 columns and 1,340 rows:

SexInsured Gender of insured, including male (M), female(F) and unspecified (U).

Female Numeric: 1 if female, 0 otherwise.

VehicleType The type of vehicle being insured, such as automobile (A), truck (T), and motorcycle (M).

PC Numeric: 1 if private vehicle, 0 otherwise.

Clm_Count Number of claims during the year.

Exp_weights Exposure weight or the fraction of the year that the policy is in effect.

LNWEIGHT Logarithm of exposure weight.

NCD No Claims Discount. This is based ont he previous accident record of the policyholder. The higher the discount, the better is the prior accident record.