

# Package ‘insuranceData’

July 22, 2025

**Type** Package

**Title** A Collection of Insurance Datasets Useful in Risk Classification  
in Non-life Insurance.

**Version** 1.0

**Date** 2014-09-04

**Author** Alicja Wolny--Dominiak and Michal Trzesiok

**Maintainer** Alicja Wolny--Dominiak <alicia.wolny-dominiak@ue.katowice.pl>

**Description** Insurance datasets, which are often used in claims severity and claims frequency modelling. It helps testing new regression models in those problems, such as GLM, GLMM, HGLM, non-linear mixed models etc. Most of the data sets are applied in the project ``Mixed models in ratemaking" supported by grant NN 111461540 from Polish National Science Center.

**License** GPL-2

**Depends** R (>= 2.10)

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2014-09-04 13:46:39

## Contents

AutoBi . . . . .	2
AutoClaims . . . . .	3
AutoCollision . . . . .	4
ClaimsLong . . . . .	5
dataCar . . . . .	6
dataOhlsson . . . . .	7
IndustryAuto . . . . .	8
SingaporeAuto . . . . .	9
Thirdparty . . . . .	10
WorkersComp . . . . .	11
<b>Index</b>	<b>12</b>

AutoBi

*Automobile Bodily Injury Claims***Description**

Data from the Insurance Research Council (IRC), a division of the American Institute for Chartered Property Casualty Underwriters and the Insurance Institute of America. The data, collected in 2002, contains information on demographic information about the claimant, attorney involvement and the economic loss (LOSS, in thousands), among other variables. We consider here a sample of  $n = 1;340$  losses from a single state. The full 2002 study contains over 70,000 closed claims based on data from thirty-two insurers. The IRC conducted similar studies in 1977, 1987, 1992 and 1997.

**Usage**

```
data(AutoBi)
```

**Format**

A data frame with 1340 observations on the following 8 variables.

CASENUM Case number to identify the claim, a numeric vector

ATTORNEY Whether the claimant is represented by an attorney (=1 if yes and =2 if no), a numeric vector

CLMSEX Claimant's gender (=1 if male and =2 if female), a numeric vector

MARITAL claimant's marital status (=1 if married, =2 if single, =3 if widowed, and =4 if divorced/separated), a numeric vector

CLMINSUR Whether or not the driver of the claimant's vehicle was uninsured (=1 if yes, =2 if no, and =3 if not applicable), a numeric vector

SEATBELT Whether or not the claimant was wearing a seatbelt/child restraint (=1 if yes, =2 if no, and =3 if not applicable), a numeric vector

CLMAGE Claimant's age, a numeric vector

LOSS The claimant's total economic loss (in thousands), a numeric vector

**Details**

<http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/DataDescriptions.pdf>

**Source**

<http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html>

**References**

Frees E.W. (2010), *Regression Modeling with Actuarial and Financial Applications*, Cambridge University Press.

**Examples**

```
data(AutoBi)
## maybe str(AutoBi) ; plot(AutoBi) ...
```

AutoClaims

*Automobile Insurance Claims***Description**

Claims experience from a large midwestern (US) property and casualty insurer for private passenger automobile insurance. The dependent variable is the amount paid on a closed claim, in (US) dollars (claims that were not closed by year end are handled separately). Insurers categorize policyholders according to a risk classification system. This insurer's risk classification system is based on automobile operator characteristics and vehicle characteristics, and these factors are summarized by the risk class categorical variable CLASS.

**Usage**

```
data(AutoClaims)
```

**Format**

A data frame with 6773 observations on the following 5 variables.

STATE Codes 01 to 17 used, with each code randomly assigned to an actual individual state, a factor with levels STATE 01 STATE 02 STATE 03 STATE 04 STATE 06 STATE 07 STATE 10 STATE 11 STATE 12 STATE 13 STATE 14 STATE 15 STATE 17

CLASS Rating class of operator, based on age, gender, marital status, use of vehicle, a factor with levels C1 C11 C1A C1B C1C C2 C6 C7 C71 C72 C7A C7B C7C F1 F11 F6 F7 F71

GENDER a factor with levels F M

AGE Age of operator, a numeric vector

PAID Amount paid to settle and close a claim, a numeric vector

**Details**

<http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/DataDescriptions.pdf>

**Source**

<http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html>

**References**

Frees E.W. (2010), Regression Modeling with Actuarial and Financial Applications, Cambridge University Press.

**Examples**

```
data(AutoClaims)
## maybe str(AutoClaims) ; plot(AutoClaims) ...
```

---

AutoCollision

*Automobile UK Collision Claims*


---

**Description**

This data is due to Mildenhall (1999). Mildenhall (1999) considered 8,942 collision losses from private passenger United Kingdom (UK) automobile insurance policies. The data were derived from Nelder and McCullagh (1989, Section 8.4.1) but originated from Baxter et al. (1980). We consider here a sample of  $n = 32$  of Mildenhall data for eight driver types (age groups) and four vehicle classes (vehicle use). The average severity is in pounds sterling adjusted for inflation.

**Usage**

```
data(AutoCollision)
```

**Format**

A data frame with 32 observations on the following 4 variables.

Age Age of driver, a factor with levels A B C D E F G H

Vehicle\_Use Purpose of the vehicle use: DriveShort means drive to work but less than 10 miles,  
DriveLong means drive to work but more than 10 miles, a factor with levels Business DriveLong  
DriveShort Pleasure

Severity Average amount of claims (in pounds sterling), a numeric vector

Claim\_Count Number of claims, a numeric vector

**Details**

<http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/DataDescriptions.pdf>

**Source**

<http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html>

**References**

Frees E.W. (2010), Regression Modeling with Actuarial and Financial Applications, Cambridge University Press.

Mildenhall S.J. (1999), A systematic relationship between minimum bias and generalized linear models, in: Proceedings of the Casualty Actuarial Society, 86, p. 393-487.

**Examples**

```
data(AutoCollision)
## maybe str(AutoCollision) ; plot(AutoCollision) ...
```

---

ClaimsLong*Claims Longitudinal*

---

**Description**

This is a simulated data set, based on the car insurance data set used throughout the text. There are 40000 policies over 3 years, giving 120000 records.

**Usage**

```
data(ClaimsLong)
```

**Format**

A data frame with 120000 observations on the following 6 variables.

policyID number of policy, a numeric vector  
agecat driver's age category: 1 (youngest), 2, 3, 4, 5, 6, a numeric vector  
valuecat vehicle value, in categories 1,...,6. (Category 1 has been recoded as 9.), a numeric vector  
period 1, 2, 3, a numeric vector  
numclaims number of claims, a numeric vector  
claim a numeric vector

**Details**

The dataset "Longitudinal Claims"

**Source**

[http://www.businessandconomics.mq.edu.au/our\\_departments/Applied\\_Finance\\_and\\_Actuarial\\_Studies/research/books/GLMsforInsuranceData/data\\_sets](http://www.businessandconomics.mq.edu.au/our_departments/Applied_Finance_and_Actuarial_Studies/research/books/GLMsforInsuranceData/data_sets)

**References**

De Jong P., Heller G.Z. (2008), Generalized linear models for insurance data, Cambridge University Press

**Examples**

```
data(ClaimsLong)
## maybe str(ClaimsLong) ; plot(ClaimsLong) ...
```

---

dataCar	<i>data Car</i>
---------	-----------------

---

### Description

This data set is based on one-year vehicle insurance policies taken out in 2004 or 2005. There are 67856 policies, of which 4624 (6.8

### Usage

```
data(dataCar)
```

### Format

A data frame with 67856 observations on the following 11 variables.

veh\_value vehicle value, in \$10,000s

exposure 0-1

clm occurrence of claim (0 = no, 1 = yes)

numclaims number of claims

claimcst0 claim amount (0 if no claim)

veh\_body vehicle body, coded as BUS CONV T COUPE HBACK HDTOP MCARA MIBUS PANVN RDSTR SEDAN  
STNWG TRUCK UTE

veh\_age 1 (youngest), 2, 3, 4

gender a factor with levels F M

area a factor with levels A B C D E F

agecat 1 (youngest), 2, 3, 4, 5, 6

X\_OBSTAT\_ a factor with levels 01101 0 0 0

### Details

dataset "Car"

### Source

<http://www.acst.mq.edu.au/GLMsforInsuranceData>

### References

De Jong P., Heller G.Z. (2008), Generalized linear models for insurance data, Cambridge University Press

### Examples

```
data(dataCar)
## maybe str(dataCar) ; plot(dataCar) ...
```

---

dataOhlsson*Motorcycle Insurance*

---

**Description**

The data for this case study comes from the former Swedish insurance company Wasa, and concerns partial casco insurance, for motorcycles this time. It contains aggregated data on all insurance policies and claims during 1994-1998; the reason for using this rather old data set is confidentiality; more recent data for ongoing business can not be disclosed.

**Usage**

```
data(dataOhlsson)
```

**Format**

A data frame with 64548 observations on the following 9 variables.

age The owners age, between 0 and 99, a numeric vector

sex The owners sex, between 0 and 99, a factor with levels K M

zone Geographic zone numbered from 1 to 7, in a standard classification of all Swedish parishes, a numeric vector

mcclass MC class, a classification by the so called EV ratio, defined as  $(\text{Engine power in kW} \times 100) / (\text{Vehicle weight in kg} + 75)$ , rounded to the nearest lower integer. The 75 kg represent the average driver weight. The EV ratios are divided into seven classes, a numeric vector

fordage Vehicle age, between 0 and 99, a numeric vector

bonuscl Bonus class, taking values from 1 to 7. A new driver starts with bonus class 1; for each claim-free year the bonus class is increased by 1. After the first claim the bonus is decreased by 2; the driver can not return to class 7 with less than 6 consecutive claim free years, a numeric vector

duration the number of policy years, a numeric vector

antskad the number of claims, a numeric vector

skadkost the claim cost, a numeric vector

**Details**

The dataset "mccase.txt"

**Source**

<http://people.su.se/~esbj/GLMbook/case.html>

**References**

Ohlsson E., Johansson B. (2010), Non-life insurance pricing with generalized linear models, Springer

**Examples**

```
data(data0hlsson)
## maybe str(data0hlsson) ; plot(data0hlsson) ...
```

---

IndustryAuto

*Auto Industry*


---

**Description**

The data represent industry aggregates for private passenger auto liability\medical coverages from year 2004, in millions of dollars. They are based on insurance company annual statements, specifically, Schedule P, Part 3B. The elements of the triangle represent cumulative net payments, including defense and cost containment expenses.

**Usage**

```
data(IndustryAuto)
```

**Format**

A data frame with 55 observations on the following 3 variables.

Incurral.Year The year in which a claim has been incurred, a numeric vector

Development.Year The number of years from incurral to the time when the payment is made, a numeric vector

Claim Cumulative net payments, including defense and cost containment expenses, a numeric vector

**Details**

<http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/DataDescriptions.pdf>

**Source**

<http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html>

**References**

Frees E.W. (2010), Regression Modeling with Actuarial and Financial Applications, Cambridge University Press.

Wacek M.G. (2007), A Test of Clinical Judgment vs. Statistical Prediction in Loss Reserving for Commercial Auto Liability, in: Casualty Actuarial Society Forum, p. 371-404.

**Examples**

```
data(IndustryAuto)
## maybe str(IndustryAuto) ; plot(IndustryAuto) ...
```



---

SingaporeAuto

---

*Singapore Automobile Claims***Description**

The data is from the General Insurance Association of Singapore, an organization consisting of general (property and casualty) insurers in Singapore (see the organization's website: [www.gia.org.sg](http://www.gia.org.sg)). From this database, several characteristics are available to explain automobile accident frequency. These characteristics include vehicle variables, such as type and age, as well as person level variables, such as age, gender and prior driving experience.

**Usage**

```
data(SingaporeAuto)
```

**Format**

A data frame with 7483 observations on the following 15 variables.

SexInsured a factor with levels F M U

Female a numeric vector

VehicleType a factor with levels A G M P Q S T W Z

PC a numeric vector

Clm\_Count a numeric vector

Exp\_weights a numeric vector

LNWEIGHT a numeric vector

NCD a numeric vector

AgeCat a numeric vector

AutoAge0 a numeric vector

AutoAge1 a numeric vector

AutoAge2 a numeric vector

AutoAge a numeric vector

VAgeCat a numeric vector

VAgecat1 a numeric vector

**Details**

<http://instruction.bus.wisc.edu/jffrees/jffreesbooks/Regression%20Modeling/BookWebDec2010/DataDescriptions.pdf>

**Source**

<http://instruction.bus.wisc.edu/jffrees/jffreesbooks/Regression%20Modeling/BookWebDec2010/data.html>

## References

Frees E.W., Valdez E.A. (2008), Hierarchical insurance claims modeling, „Journal of the American Statistical Association", 103(484), p. 1457-1469.

Frees E.W. (2010), Regression Modeling with Actuarial and Financial Applications, Cambridge University Press.

## Examples

```
data(SingaporeAuto)
## maybe str(SingaporeAuto) ; plot(SingaporeAuto) ...
```

---

Thirdparty

*Third party insurance*

---

## Description

Third party insurance is a compulsory insurance for vehicle owners in Australia. It insures vehicle owners against injury caused to other drivers, passengers or pedestrians, as a result of an accident.

This data set records the number of third party claims in a twelve-month period between 1984-1986 in each of 176 geographical areas (local government areas) in New South Wales, Australia.

## Usage

```
data(Thirdparty)
```

## Format

A data frame with 176 observations on the following variable.

`lga.sd.claims.accidents.ki.population.pop_density` a numeric vector

## Details

The dataset "Third Party Claims"

## Source

[http://www.businessandconomics.mq.edu.au/our\\_departments/Applied\\_Finance\\_and\\_Actuarial\\_Studies/research/books/GLMsforInsuranceData/data\\_sets](http://www.businessandconomics.mq.edu.au/our_departments/Applied_Finance_and_Actuarial_Studies/research/books/GLMsforInsuranceData/data_sets)

## References

De Jong P., Heller G.Z. (2008), Generalized linear models for insurance data, Cambridge University Press

## Examples

```
data(Thirdparty)
## maybe str(Thirdparty) ; plot(Thirdparty) ...
```

---

WorkersComp

*Workers Compensation*

---

### Description

Standard example in worker's compensation insurance, examining losses due to permanent, partial disability claims. The data are from Klugman (1992), who considers Bayesian model representations, and are originally from the National Council on Compensation Insurance. We consider  $n=121$  occupation, or risk, classes, over  $T=7$  years. To protect the data source, further information on the occupation classes and years is not available. Source: Frees, E. W., Young, V. and Y. Luo (2001). Case studies using panel data models. North American Actuarial Journal, 4, No. 4, 24-42.

### Usage

```
data(WorkersComp)
```

### Format

A data frame with 847 observations on the following 4 variables.

CL a numeric vector

YR a numeric vector

PR a numeric vector

LOSS a numeric vector

### Details

<http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression>

DataDescriptions.pdf

### Source

<http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression>

### References

Frees E.W. (2010), Regression Modeling with Actuarial and Financial Applications, Cambridge University Press.

### Examples

```
data(WorkersComp)
## maybe str(WorkersComp) ; plot(WorkersComp) ...
```

# Index

## \* datasets

- AutoBi, [2](#)
- AutoClaims, [3](#)
- AutoCollision, [4](#)
- ClaimsLong, [5](#)
- dataCar, [6](#)
- dataOhlsson, [7](#)
- IndustryAuto, [8](#)
- SingaporeAuto, [9](#)
- Thirdparty, [10](#)
- WorkersComp, [11](#)

- AutoBi, [2](#)
- AutoClaims, [3](#)
- AutoCollision, [4](#)

- ClaimsLong, [5](#)

- dataCar, [6](#)
- dataOhlsson, [7](#)

- IndustryAuto, [8](#)

- SingaporeAuto, [9](#)

- Thirdparty, [10](#)

- WorkersComp, [11](#)