Fallstudie: GLMs in der Motorradversicherung

Isaac David, Ramirez Limones

16. Oktober 2025

Inhaltsverzeichnis

1 Einführung

Mein persönliches Ziel für dieses Projekt ist es, mithilfe von dem Buch von Ohlsson und Johansson 2010, grundlegende, aber zugleich moderne Konzepte der Sachversicherung zu erlernen und diese auf einen echten Datensatz einer schwedischen Versicherung anzuwenden. Konkret handelt es sich um eine Motorradversicherung, für die eine reale Tarifierung aus dem Jahr 1995 vorliegt. Die Daten stammen aus den Jahren 1994 bis 1998. Unser Ziel ist es zu prüfen, ob die damalige Tarifierung noch angemessen ist oder ob sie angepasst werden sollte. Dieses Projekt ist im Buch [1] vorgeschlagen und basiere mich hauptsächlich auf dieses Buch.

Zunächst möchte ich einige Versicherungsbegriffe vorstellen, die ich im Laufe dieses Projekts gelernt habe und die für das Verständnis dieses Protokolls wichtig sind:

- Context in insurance (claims, pricing, risk modeling) - Objectives of this case study

2 Datenbeschreibung

Der Datensatz umfasst 64.548 Versicherungsnehmer. Wir haben vier Tarifmerkmale: geografische Zone, MC-Klasse, Fahrzeugalter und Bonusklasse. Außerdem stehen andere Informationen zur Verfügung, wie das Alter des Besitzers, Geschlecht, Vertragsdauer und Anzahl der Schadensfälle. Die Daten liegen in folgender Gestalt in einer .txt Datei vor:

0M141210.175342	0	0
4M36 91 0	0	0
5K331810.454795	0	0
5K412510.172603	0	0
6K212610.180822	0	0

Die erste Herausforderung bestand darin, das Format zu verstehen und die Inhalte so aufzubereiten, dass sie von RStudio verarbeitet werden können. Es stellte sich heraus, dass dieses Format typisch für Versicherungsdaten ist und sich mit der Funktion read_fwf leicht importieren lässt, sodass die Spalten besser erkennbar sind. Dasselbe lässt sich auch in Python mit Pandas umsetzen:

```
1    df = pd.read_fwf("data.txt", colspecs=colspecs, names
        =col_names)
```

2.1 Datenaggregation and Analyse

Zunächst gilt es, die Tarifmerkmale zu analysieren. Eine der Aufgaben aus dem Buch bestand darin, die empirische Schadenhäufigkeit und Schadenhöhe für jede Zelle zu bestimmen. Die Daten liegen jedoch nicht in Zellen vor, sondern jede Zeile entspricht den Angaben eines einzelnen Versicherungsnehmers. Deshalb muss zunächst eine Datenaggregation durchgeführt werden.

```
# Importiere Packet dplyr
# TotalClaims ist die Anzahl der Schadenanfaelle
# TotalClaimCost sind die Kosten der Schadenanfaelle in der Zelle
# TotalExposure ist die Summe der Vertragsdauern von jedem Versicherungsnehmer in der Zelle

Agg_data <- data %>%
  group_by(geographic.zone, mc_class, vehicle_age_class, bonus_class) %>%
  summarise(
    TotalClaims = sum(number_claims),
    TotalClaimCost = sum(claim_cost),
    TotalExposure = sum(duration)
  ) %>%
  mutate(
    Frequency = TotalClaims / TotalExposure,
    Severity = TotalClaimCost / TotalClaims
  )
```

3 Mathematisches Modell für die neue Tariffierung

In der Sachversicherung ist die Methode der Randsummen ein klassischer Ansatz. In diesem Projekt arbeiten wir jedoch mit Verallgemeinerten Linearen Modellen (GLM), insbesondere mit multiplikativen Modellen, die zu dieser Modellklasse gehören. Wir werden auch sehen, dass dieses moderne Modell im Fall von multiplikativen Modellen zu der klassischen Methode entsprechen.

3.1 Schadenhäufigkeit

Unter multiplikativem Modell ist die mittlere Schadenhäufigkeit für jede Zelle gegeben durch:

$$\mu_{ij} = \gamma_0 \cdot \gamma_{A,i} \cdot \gamma_{B,j} \cdot \gamma_{C,k} \cdot \gamma_{D,l}$$

Wir nehmen an, dass die Daten durch die relative Poisson Verteilung generiert wurden. Somit:

$$Y_{ijkl} \sim \Pi\big(\omega_{ijkl}\,\mu_{ijkl}\big),$$
$$P(Y_{ijkl} = y_{ijkl}) = \frac{e^{-\omega_{ijkl}\mu_{ijkl}}\,(\omega_{ijkl}\mu_{ijkl})^{y_{ijkl}}}{y_{ijkl}!}.$$

Die Log-Likelihood in der Zelle (i,j,k,l) ist gegeben durch

$$l_{ijkl} = y_{ijkl}\omega_{ijkl}\log\left(\omega_{ijkl}\mu_{ijkl}\right) - \omega_{ijkl}\mu_{ijkl} - \log(y_{ijkl}!)$$

Summe über alle Zellen liefert die totale Log-likelihood:

$$l = \sum_{i,j,k,l} \omega_{ijkl}\left(y_{ijkl}(\log\gamma_0 + \log\gamma_{A,i} + \log\gamma_{B,j} + \log\gamma_{C,k} + \log\gamma_{D,l}) - \gamma_0\cdot\gamma_{A,i}\cdot\gamma_{B,j}\cdot\gamma_{C,k}\cdot\gamma_{D,l}\right) + c$$

wobei die Konstante den Termen entspricht, die nicht von den γ-Koeffizienten abhängig sind. Ableiten nach diesen Koeffizienten liefert das folgende Gleichungssystem:

$$\gamma_0 = \frac{\sum_{i,j,k,l}\omega_{ijkl}\,y_{ijkl}}{\sum_{i,j,k,l}\omega_{ijkl}\,\gamma_{A,i}\,\gamma_{B,j}\,\gamma_{C,k}\,\gamma_{D,l}}$$

$$\gamma_{A,i} = \frac{\sum_{j,k,l}\omega_{ijkl}\,y_{ijkl}}{\sum_{j,k,l}\omega_{ijkl}\,\gamma_0\,\gamma_{B,j}\,\gamma_{C,k}\,\gamma_{D,l}} \quad \gamma_{C,k} = \frac{\sum_{i,j,l}\omega_{ijkl}\,y_{ijkl}}{\sum_{i,j,l}\omega_{ijkl}\,\gamma_0\,\gamma_{A,i}\,\gamma_{B,j}\,\gamma_{D,l}}$$

$$\gamma_{B,j} = \frac{\sum_{i,k,l}\omega_{ijkl}\,y_{ijkl}}{\sum_{i,k,l}\omega_{ijkl}\,\gamma_0\,\gamma_{A,i}\,\gamma_{C,k}\,\gamma_{D,l}} \quad \gamma_{D,l} = \frac{\sum_{i,j,k}\omega_{ijkl}\,y_{ijkl}}{\sum_{i,j,k}\omega_{ijkl}\,\gamma_0\,\gamma_{A,i}\,\gamma_{B,j}\,\gamma_{C,k}}$$

Dieses Gleichungssystem besitzt keine geschlossene Lösung. Allerdings lässt sich eine approximative Lösung iterativ bestimmen. Dieses multiplikative Modell lässt sich direkt auf R mit dem befehl glm() implementieren. Bei diesem Befehl muss man nicht nur die Daten angeben, aber auch die betrachteten Annahmen, die ich in der nächsten Tabelle besser beschreibe:

Theoretische Annahme	Argument auf R
Das verallgemeinerte Modell ist multiplikativ	`poisson(link = "log")`
Mittlere Schadenhäufigkeit wird im Modell durch die gegebenen Tarifmerkmale beschrieben	`Frequency geographic.zone + mc_class + vehicle_age_class + bonus_class`

Tabelle 1: Theoretische Annahmen und ihre Formulierung in R

```r
# Multiplikatives Modell fuer Schadenhaeufigkeit

agg$geographic.zone <- as.factor(agg$geographic.zone)
agg$mc_class <- as.factor(agg$mc_class)
agg$vehicle_age_class <- as.factor(agg$vehicle_age_class)
agg$bonus_class <- as.factor(agg$bonus_class)


fit <- glm(
  Frequency ~ geographic.zone + mc_class + vehicle_age_
     class + bonus_class,
  family = poisson(link="log"),
  offset = log(TotalExposure),
  data = agg
)
```

3.2 Schadenhöhe

Nun möchten wir ein Modell für die Schadenhöhe aufstellen. Dabei betrachten wir wieder ein multiplikatives Modell, berechnen die Log-Likelihood und bestimmen mit einem iterativen Verfahren eine Lösung der ML-Gleichungen. In R ist dieser Prozess wieder durch den Befehl glm() automatisiert, sodass wir ihn direkt wie folgt anwenden können. Davor muss man aber auf die Zellen aufpassen, wo keine Schadenanfälle stehen. Sie werden nicht relevant für die Modellierung sein.

```r
# Setze die Schadenhoehe NA, wo keine Schadenanfaelle
   stehen und trainiere fit2 Modell
agg$Severity[agg$TotalClaims == 0] <- NA
agg$Severity[agg$Severity == 0] <- NA

fit2 <- glm(
  Severity ~ geographic.zone + mc_class + vehicle_age_
     class + bonus_class,
  family = poisson(link="log"),
  offset = log(TotalExposure),
  data = agg
)
```

3.3 Berechnung der Pure Premium

Dieser Ansatz liefert zwei Modelle: eines für die Schadenhäufigkeit und eines für die Schadenhöhe. Um die Relativitäten zu bestimmen, müssen wir die Koeffizienten multiplizieren. Dieses Produkt entspricht dem erwarteten gesamten Schadenbetrag pro Exposure und pro Individuum. Die Ergebnisse sind auf Tabelle 2 dargestellt.

Rating Factor	Class	Relativity
Geographic zone	1	1
	2	0.23305440
	3	0.29860032
	4	0.02595793
	5	0.17687779
	6	1.57478008
	7	0.43286261
MC Class	1	1
	2	0.70049090
	3	0.03977275
	4	0.13728359
	5	0.30515049
	6	1.40609317
	7	51.18907225
Vehicle Age Class	1	1
	2	0.86349710
	3	0.01309964
Bonus Class	1	1
	2	1.68361823
	3	7.47178039
	4	9.21661330
	5	32.23672759
	6	2.80582578
	7	0.30732623

Tabelle 2: Relativities table for computing prices

Literatur

Ohlsson, Esbjörn und Björn Johansson (2010). *Non-Life Insurance Pricing with Generalized Linear Models*. EAA Series. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 174. ISBN: 978-3-642-10790-0. DOI: 10.1007/978-3-642-10791-7.