Modelowanie wysokości szkody z użyciem uogólnionych modeli liniowych (GLM)

sprobulski

7 lutego 2026

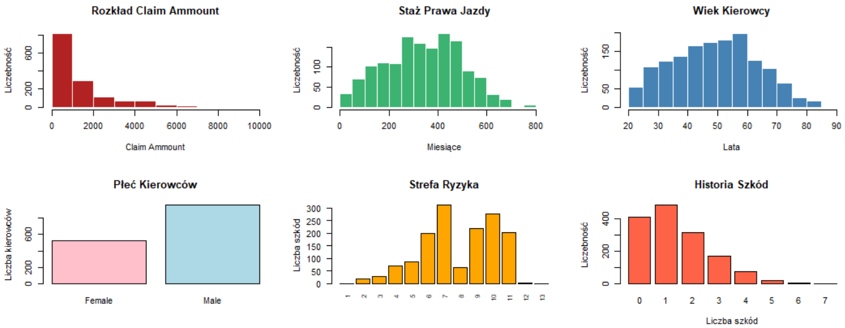
Cel badania i opis danych

Cel projektu: Budowa i porównanie modeli predykcyjnych dla **średniej wartości pojedynczej szkody** (*claim severity*). Analiza szkód niezerowych ($Y > 0$).

Zmienne wykorzystane w modelu:

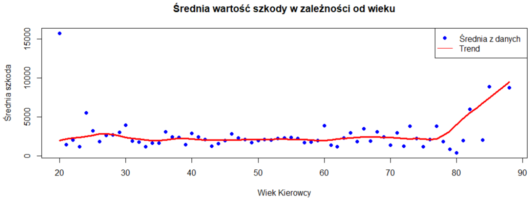
- **ClaimAmount** (Target): Wysokość odszkodowania.
- **DrivAge**: Wiek kierowcy (zmienna ciągła, lata).
- **LicAge**: Staż prawa jazdy (zmienna ciągła, miesiące).
- **Gender**: Płeć kierowcy (zmienna binarna).
- **RiskArea**: Strefa ryzyka (zmienna kategoryczna, poziomy 1-13).
- **PastClaims**: Liczba przeszłych szkód (zmienna ciągła).

Rozkład zmiennych



- **ClaimAmount:** Rozkład silnie prawoskośny z „ciężkim ogonem". Większość szkód < 2000, ale występują ekstremalne wartości.
- Ze zmiennej **RiskArea** usunięto poziom 1, poziomy 11, 12, 13 połączono w jeden.

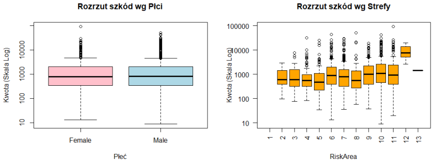
Zależność średniej wartości szkody od wieku



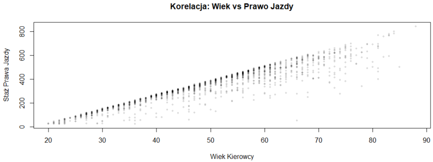
Wnioski:

- Zależność względnie stała.
- **Młodzi (< 30 lat):** Szkody nie są zauważalnie wyższe.
- **Seniorzy (> 80 lat):** Wzr ost wartości szkód.

Analiza czynników ryzyka i korelacje



Boxploty: Płeć i Strefa



Korelacja: Wiek vs Staż

- **RiskArea:** Silny predyktor (mediana szkody rośnie ze strefą).
- **Gender:** Brak widocznych różnic w rozkładach.
- **Współliniowość:** Silna korelacja *DrivAge* i *LicAge*.

Model 1: Gamma GLM (Specyfikacja)

Specyfikacja modelu:

- Rodzina: **Gamma**
- Funkcja łącząca: `log`
- Funkcja wariancji: $V(\mu) \propto \mu^2$

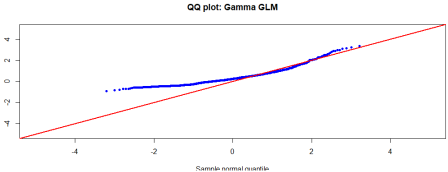
Selekcja zmiennych (Model zredukowany):

- Użyto testu F
- Usunięto zmienną **Gender** (nieistotna statystycznie).
- Pozostałe zmienne (`LicAge`, `DrivAge`, `RiskArea`, `PastClaims`) są istotne.

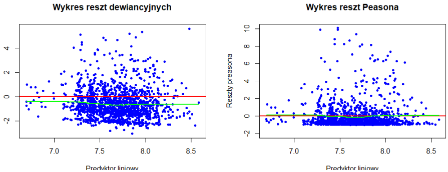
Ostateczna postać modelu

```
glm(ClaimAmount ~ LicAge + DrivAge + RiskArea + PastClaims,
family = Gamma(link = "log"), data_set)
```

Model 1: Gamma GLM (Diagnostyka Graficzna)



Q-Q Plot



Reszty

Ocena:

- **Reszty (Prawo):** Wyglądają poprawnie – są stabilne, oscylują wokół zera dla reszt Pearsona, brak wyraźnych trendów.
- **Q-Q Plot (Lewo):** Widoczne odchylenie punktów od linii teoretycznej sugeruje, że rozkład Gamma nie opisuje idealnie ogonów rozkładu szkód, mimo stabilności reszt.

Model 2: Inverse Gaussian GLM (Specyfikacja)

Specyfikacja modelu:

- Rodzina: **Inverse Gaussian** (Odwrotny Gaussowski)
- Funkcja łącząca: `log`
- Funkcja wariancji: $V(\mu) \propto \mu^3$

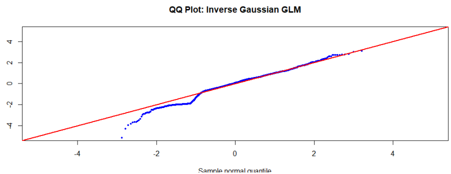
Selekcja zmiennych (na podstawie AIC):

- Użyto kryterium AIC.
- Usunięto zmiennej **Gender**, **LicAge**, **DrivAge**.
- Pozostałe zmienne: **RiskArea**, **PastClaims**

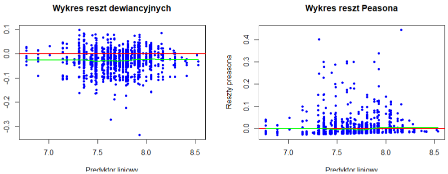
Ostateczna postać modelu

```
glm(ClaimAmount ~ RiskArea + PastClaims, family =
inverse.gaussian(link="log"), data_set)
```

Model 2: Inverse Gaussian GLM (Diagnostyka Graficzna)



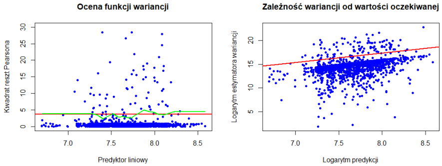
Q-Q Plot



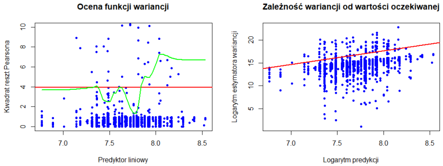
Reszty

- **Q-Q Plot (Lewo):** Punkty systematycznie odstają od linii, co oznacza, że model błędnie opisuje rozkład prawdopodobieństwa.
- **Wniosek:** Wyglądają poprawnie – reszty Pearsona są stabilne, niskie i oscylują wokół zera. Oznacza to, że model poprawnie estymuje wartość oczekiwaną (średnią).

Weryfikacja funkcji wariancji (Gamma vs IG)



Gamma (Nachylenie 2)

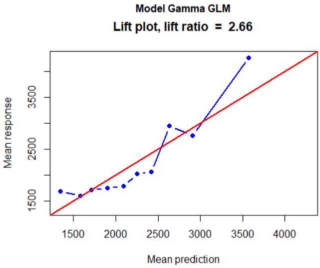


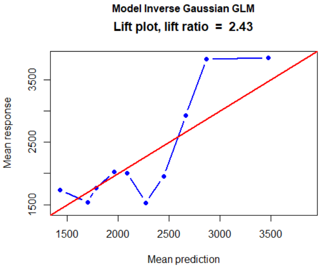
Inverse Gaussian (Nachylenie 3)

- **Gamma:** Dane podążają za linią o nachyleniu 2.
- **IG:** Dane są bardziej „płaskie” niż linia o nachyleniu 3, kwadrat reszt nie jest stały.
- **Wniosek:** Rozkład Gamma jest lepszym dopasowaniem niż rozkład odwrotny Gaussa.

Porównanie modeli podstawowych

Metryka	Gamma GLM	Inv. Gaussian GLM
AIC (niższe = lepsze)	25 536.95	**25 212.70**
Lift Ratio (wyższe = lepsze)	**2.66**	2.43





- **Inverse Gaussian:** Wygrywa w AIC (lepiej pasuje do ogona), ale przegrywa biznesowo. Na wykresie (prawo) widać niestabilność i **przeszacowanie** ryzyka dla najgorszych klientów.
- **Gamma:** Mimo wyższego AIC, oferuje lepszą segmentację (wyższy Lift Ratio). **Jest modelem lepszym do taryfikacji.**

Prognozy punktowe i przedziałowe

Metodologia:

- Wybrano dwóch ubezpieczonych z bazy danych.
- Wyznaczono predykcję wartości oczekiwanej $\hat{\mu} = \exp(\hat{\eta})$.
- 95% przedziały ufności obliczono **metodą Delta**, wykorzystując macierz kowariancji parametrów.

Tabela: Wyniki predykcji wartości szkody (w jednostkach pieniężnych)

Profil	Model	Dolna granica	Predykcja	Górna granica
Osoba 1 (id=10)	Gamma	1265.56	1739.23	2212.90
	Inv. Gauss	1337.90	1806.30	2274.71
Osoba 2 (id=200)	Gamma	1523.86	2184.76	2845.66
	Inv. Gauss	1449.63	2130.83	2812.03

Interpretacja:

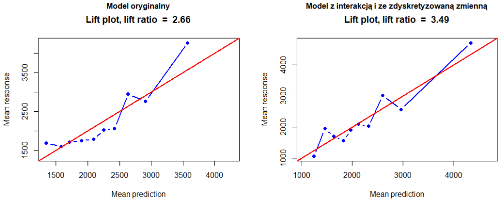
- Dla typowych profili klientów oba modele generują zbliżone prognozy punktowe (różnice rzędu $\sim$ 50 jednostek).

Wprowadzone modyfikacje:

1. **Dyskretyzacja:** Zmienna `PastClaims` potraktowana jako katogorie (factor). Rzadkie kategorie (5, 6, 7 szkód) połączono w grupę „5+".
2. **Interakcja:** Dodano składnik `LicAge:DrivAge`.

Tabela: Porównanie dopasowania modelu oryginalnego i zmodyfikowanego

Metryka	Model Oryginalny	Model Zmodyfikowany
AIC	25 536.95	**25 503.06**
Reszta Pearsona	5555.02	**4929.43**
Lift Ratio	2.66	**3.49**



Model Tweediego (Część 1: Specyfikacja Zmiennych)

Koncepcja: Model Tweediego jest rozszerzeniem rodziny wykładniczej. Pozwala nam zachować strukturę zmiennych z poprzedniego etapu, ale elastycznie dobrać rozkład błędu.

Struktura predyktora liniowego

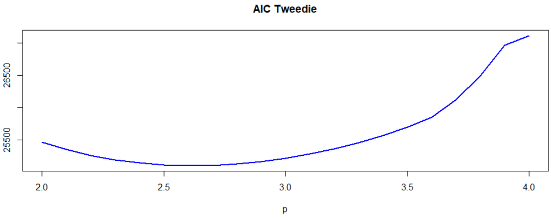
Model wykorzystuje **dokładnie ten sam zestaw zmiennych**, który wyłoniono w udoskonalonym modelu Gamma:

- **Zmienne bazowe:** `LicAge`, `DrivAge`, `RiskArea`, `PastClaims`.
- **Usunięto:** Zmienną `Gender` (brak istotności).
- **Przekształcenia:**
  - `PastClaims` jako zmienna kategoryczna (z grupą „5+").

Model Tweediego (Część 2: Optymalizacja parametru p)

Parametr wariancji p: Szukamy wykładnika w funkcji wariancji $V(\mu) = \mu^p$, który najlepiej pasuje do danych.

- $p = 2 \rightarrow$ Rozkład Gamma.
- $p = 3 \rightarrow$ Rozkład Inverse Gaussian.



Wynik optymalizacji:

- Minimum AIC osiągnięto dla $p \approx 2.6$.
- Oznacza to, że rozkład szkód ma „cięższy ogon" niż zakłada Gamma, ale nie tak ekstremalny jak w IG.

Model Double GLM (Podwójny uogólniony model liniowy)

Cel: Modelowanie nie tylko wartości oczekiwanej, ale i zmienności (dyspersji) w grupach klientów (Heteroskedastyczność).

Struktura modelu:

1. **Podmodel średniej (μ):** Standardowy GLM Gamma.

$$g(\mu_i) = x_i^T \beta$$

2. **Podmodel dyspersji (ϕ):** Parametr dyspersji zależy od cech kierowcy.

$$h(\phi_i) = z_i^T \gamma$$

Porównanie modeli: Kryterium AIC

Zestawienie wartości kryterium informacyjnego Akaike (im niższe, tym lepsze).

Tabela: Ranking modeli wg dopasowania statystycznego

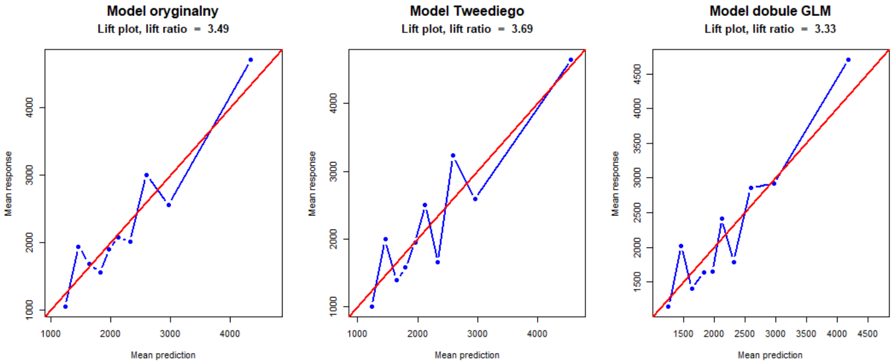
Model	**AIC**
Gamma (podstawowy)	25 537.55
Gamma (z interakcjami)	25 503.06
Double GLM	25 510.05
Tweedie ($p = 2.6$)	**25 099.36**

Wnioski:

- Model Tweediego jest najlepszy (najniższe AIC).

Ostateczna weryfikacja biznesowa (Lift Plot)

Porównanie zdolności segmentacyjnej trzech najlepszych modeli.



- **Tweedie (Środkowy):** Lift ratio 3.69 - najlepszy model

Rekomendacja końcowa

Najlepszy model - Tweedie

Jako ostateczny model do taryfikacji rekomenduje się **Model Tweediego GLM ($p = 2.6$)**.

Uzasadnienie: Model ten posiada najniższe AIC, najwyższy wskaźnik Lift Ratio (3.69).