Progetto Modelli Statistici Per Le Scienze Attuariali

Nicolò Amadori, Matteo Ferniani, Tommaso Zignani

15 Settebre 2024

1 Obiettivo dell'indagine

L'obiettivo principale di questa indagine è quello di sviluppare un modello per la determinazione dei premi assicurativi attraverso l'implementazione delle tecniche di pricing comunemente utilizzate dalle compagnie assicurative. Il premio assegnato a ciascun assicurato viene calcolato in funzione della sua appartenenza a una specifica classe tariffaria e sulla base di un'analisi della sua redditività.

Per raggiungere tale obiettivo, si parte da un'analisi descrittiva di un dataset relativo ai contratti assicurativi, con particolare attenzione alla classificazione e suddivisione delle varie classi tariffarie. Successivamente, si procede con la stima della frequenza dei sinistri, definita come il rapporto tra il numero di sinistri e l'esposizione, e della loro gravità (severity), utilizzando modelli di regressione basati sui GLM (Generalized Linear Models).

Una volta ottenuti i parametri di frequenza e gravità, il premio per ciascuna classe tariffaria viene calcolato moltiplicando i due valori stimati. Infine, si conclude con una simulazione tariffaria, che tiene conto di un bilancio complessivo, al fine di valutare la sostenibilità e l'efficacia del modello di pricing proposto.

2 Analisi Descrittiva

Il dataset di partenza è composto da 393.071 osservazioni, ognuna delle quali rappresenta una polizza assicurativa. Su queste osservazioni sono state raccolte informazioni relative a 12 variabili, tra cui:

- Gender: sesso dell'assicurato
  - DrivAge: età del guidatore, quindi dell'assicurato
  - VehYear: anno del veicolo
  - VehModel: modello di autovettura
  - VehGroup: gruppo del veicolo
  - Area: area geografica
  - State: stato
  - StateAb: abbreviazione degli stati
  - ExposTotal: esposizione
  - SumInsAvg: media della somma assicurata
  - ClaimNb: numero di risarcimenti durante il periodo di esposizione
  - ClaimAmount: ammontare dei risarcimenti

Il dataset in esame riguarda le polizze assicurative per sinistri auto in Brasile. In una fase iniziale del nostro processo analitico, abbiamo esaminato i dati relativi alle esposizioni per identificare eventuali istanze in cui il valore dell'esposizione risultava essere pari a zero. Utilizzando la funzione 'summary(dataset$ExposTotal)', abbiamo riscontrato un valore minimo di 0.000.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
ExposTotal	0.000	0.400	0.530	3.212	1.800	2346.500

Table 1: Summary of the variable ExposTotal

Abbiamo quindi deciso di escludere tali casi, in quanto le esposizioni con valore zero non contribuivano in modo significativo all'analisi, risultando prive di informazioni utili.

Per garantire un'analisi finanziaria e assicurativa coerente e comparabile, abbiamo proceduto alla conversione dei valori di esposizione presenti nel dataset in una scala temporale giornaliera. Questo è stato ottenuto dividendo ciascun valore di esposizione per 365, standardizzando così i dati su base quotidiana.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
ExposTotal	0.00008	0.001205	0.001589	0.009037	0.005151	6.428767

Table 2: Summary of the variable ExposTotal after conversion to a daily scale and no Zeros

Successivamente, abbiamo proceduto con la discretizzazione della variabile 'VehYear', suddividendo gli anni di immatricolazione dei veicoli in quattro categorie basate sui quartili della distribuzione. Questa suddivisione è stata realizzata utilizzando il seguente codice R, che ha trasformato la variabile 'VehYear' in formato numerico e successivamente l'ha segmentata in quattro intervalli: "<2003", "2003-2007", "2007-2009" e ">2009". Tale processo è stato reso possibile applicando la funzione 'cut()' e utilizzando i quartili calcolati tramite la funzione 'quantile()'.

```
ds$VehYear <- as.numeric(ds$DrivAge)

ds$VehYear <- cut(ds$VehYear,
                  breaks = quantile(ds$VehYear,
                  probs = seq(0, 1, 0.25),
                  na.rm = TRUE),
                  labels = c("< 2003", "2003-2007",
                  "2007-2009", ">2009"),
                  include.lowest = TRUE)
```

Inoltre, abbiamo rimosso le osservazioni contenenti valori mancanti (missing values) per migliorare la precisione e l'affidabilità dell'analisi. Questo passo è stato cruciale per garantire che i risultati non fossero influenzati da dati incompleti o errati.

3 Creazione delle celle tariffarie

Successivamente, abbiamo creato le celle tariffarie utilizzando la libreria 'MASS', un pacchetto R comunemente impiegato per la modellizzazione statistica. Per assicurare la corretta elaborazione dei dati, le colonne del dataset relative al numero di sinistri ('ClaimNb') e all'ammontare dei sinistri ('ClaimAmount') sono state convertite in formato intero. Questo passaggio ha permesso di ottenere una struttura del dataset più coerente e pronta per le successive fasi dell'analisi quantitativa.

Gender Levels	Corporate, Female, Male
DrivAge Levels	>55, 18–25, 26–35, 36–45, 46–55
VehModel Levels	Acura - Legend 3.2/3.5 Agrale - 13000 Turbo 2p (diesel) Agrale - 1600 D-rd 2p (diesel)
DrivAge Summary	>55: 71527, 18–25: 23585, 26–35: 61938, 36–45: 78533, 46–55: 72796

Table 3: levels of Gender Levels, DrivAge Levels, VehModel Levels, and DrivAge

Istogramma della variabile 'ClaimAmount' per valori maggiori di 0

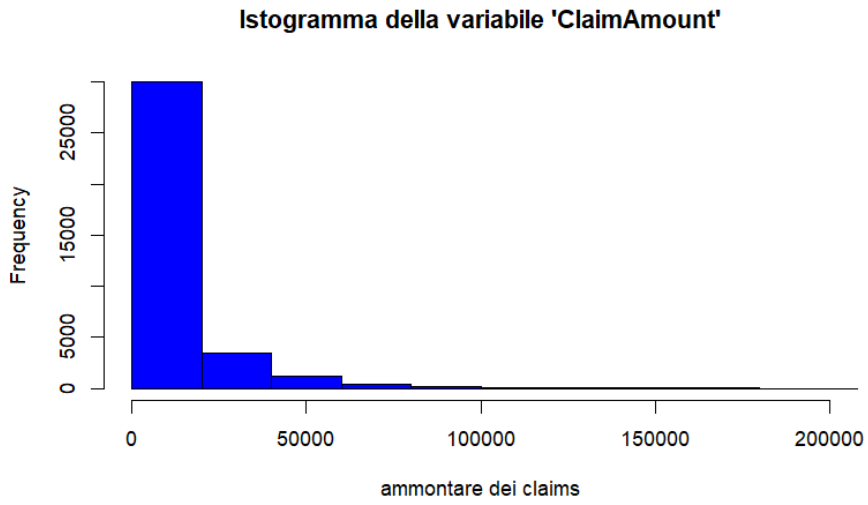

Figure 1: Istogramma della variabile 'ClaimAmount'

Da questo grafico osserviamo:

1. **Prevalenza di Sinistri di Basso Importo**: la maggior parte dei sinistri ha importi relativamente bassi. Questo è comune nei dataset di sinistri assicurativi, dove molti incidenti comportano danni minimi.

2. **Distribuzione positivamente asimmetrica**: la distribuzione degli importi dei sinistri è positivamente asimmetrica, con un gran numero di sinistri di basso importo e pochi sinistri di alto importo.

Calcolato il numero totale di sinistri e l'esposizione totale per determinare la frequenza dei sinistri, aggregando il numero di sinistri per area. Create tabelle per il numero di sinistri e per l'esposizione totale per area. Utilizzando tabFreq, è stata calcolata la frequenza dei sinistri per area.

Sono state create le celle tariffarie aggregando il numero di sinistri e l'esposizione totale per area delle celle (verificando la presenza di NA) e calcolata la frequenza dei sinistri per le celle tariffarie.

Area	**ClaimNb**	**ExposTotal**	**Freq**
Acre	108	2.5638713	42.12380
Alagoas	557	20.0319700	27.80555
Amapa	56	1.6122394	34.73429
Amazonas	354	10.8467739	32.63643
Bahia	2318	102.5178307	22.61070

Table 4: celle tariffarie coi claims, exposure total, and frequency by area

3.1 Regressione

Eseguiamo una regressione sul numero di risarcimenti, quindi sul numero di sinistri, utilizzando le variabili Gender, DrivAge e State

Coefficient	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-12.37183	39.22792	-0.315	0.752471
GenderFemale	15.86536	39.22780	0.404	0.68588
GenderMale	15.91923	39.22780	0.406	0.68478
DrivAge18-25	0.59464	0.01862	31.933	2e-16 ***
DrivAge26-35	0.23273	0.01208	19.257	2e-16 ***
DrivAge36-45	0.35193	0.01086	32.408	2e-16 ***
DrivAge46-55	0.15206	0.01205	12.623	2e-16 ***
StateAlagoas	-0.41925	0.10515	-3.987	6.69e-05 ***
StateAmapa	-0.18277	0.16467	-1.110	0.267172
StateAmazonas	-0.24168	0.10993	-2.199	0.027913 *
StateBahia	-0.59874	0.09845	-6.082	1.19e-09 ***
StateCeara	-0.31840	0.09959	-3.197	0.001388 **
StateDistrito Federal	-0.28074	0.09811	-2.862	0.004215 **
StateEspirito Santo	-0.33499	0.10021	-3.343	0.000829 ***
StateGoias	-0.26501	0.09812	-2.701	0.006918 **
StateMaranhao	-0.21599	0.10569	-2.046	0.040726 *
StateMato Grosso	-0.12126	0.10056	-1.206	0.227891
StateMato Grosso do Sul	-0.07802	0.10807	-0.722	0.47039
StateMinas Gerais	-0.37803	0.09648	-3.898	6.89e-05 ***
StatePara	-0.42412	0.10366	-4.091	4.29e-05 ***
StateParaiba	-0.32448	0.10336	-3.139	0.001701 **
StateParana	-0.39675	0.09714	-4.086	4.39e-05 ***
StatePernambuco	-0.40754	0.09744	-4.184	2.17e-05 ***
StatePiaui	-0.09067	0.10871	-0.834	0.404290
StateRio de Janeiro	-0.36905	0.09694	-3.808	0.000141 ***
StateRio Grande do Norte	-0.27085	0.10239	-2.645	0.008164 **
StateRio Grande do Sul	-0.55268	0.09723	-5.684	1.13e-08 ***
StateRondonia	-0.09244	0.12052	-0.767	0.443029
StateRoraima	-0.76349	0.13487	-5.662	1.47e-08 ***
StateSanta Catarina	-0.47304	0.09889	-4.784	1.74e-06 ***
StateSao Paulo	-0.68630	0.09584	-7.117	1.11e-12 ***
StateSergipe	-0.14296	0.10439	-1.370	0.17082
StateTocantins	-0.20918	0.11328	-1.847	0.064805 .

Si nota come la variabile Gender è poco significativa;

Si procede con un altro modello omettendo la variabile poco significativa Gender

Coefficients	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	3.506679	0.096570	36.312	¡ 2e-16 ***
DrivAge18-25	0.568716	0.018581	30.608	¡ 2e-16 ***
DrivAge26-35	0.219596	0.012007	18.290	¡ 2e-16 ***
DrivAge36-45	0.342111	0.010821	31.616	¡ 2e-16 ***
DrivAge46-55	0.136923	0.012020	11.391	¡ 2e-16 ***
StateAlagoas	-0.394917	0.105145	-3.756	0.000173 ***
StateAmapa	-0.190576	0.164672	-1.157	0.247147
StateAmazonas	-0.245248	0.109931	-2.231	0.025686 *
StateBahia	-0.592137	0.098446	-6.015	1.80e-09 ***
StateCeara	-0.307606	0.099587	-3.089	0.002010 **
StateDistrito Federal	-0.269377	0.098107	-2.746	0.006037 **
StateEsperito Santo	-0.333175	0.100203	-3.325	0.000884 ***
StateGoias	-0.254278	0.098123	-2.591	0.009558 **
StateMaranhao	-0.205611	0.106588	-1.929	0.053727 .
StateMato Grosso	-0.107698	0.100563	-1.071	0.284192
StateMato Grosso do Sul	-0.063164	0.100751	-0.627	0.530702
StateMinas Gerais	-0.387855	0.096975	-4.000	6.35e-05 ***
StatePara	-0.425763	0.103661	-4.107	4.00e-05 ***
StateParaiba	-0.314212	0.104021	-3.021	0.002522 **
StateParana	-0.404710	0.097108	-4.168	3.08e-05 ***
StatePernambuco	-0.684682	0.099419	-6.887	5.70e-12 ***
StatePiaui	-0.088315	0.108713	-0.812	0.416582
StateRio de Janeiro	-0.928535	0.097429	-9.530	¡ 2e-16 ***
StateRio Grande do Norte	-0.256620	0.102390	-2.506	0.012200 *
StateRio Grande do Sul	-0.550651	0.097226	-5.664	1.48e-08 ***
StateRondonia	0.006278	0.180323	0.035	0.972229
StateRoraima	-0.757129	0.203756	-3.716	0.000203 ***
StateSanta Catarina	-0.476157	0.097605	-4.878	1.07e-06 ***
StateSao Paulo	-0.687949	0.096434	-7.134	9.76e-13 ***
StateSergipe	-0.110208	0.104416	-1.055	0.291211
StateTocantins	-0.203138	0.113275	-1.793	0.072921 .

Questo modello presenta un criterio di informazione AIC maggiore

Si procede poi con una regressione stepwise utilizzando il criterio di selezione AIC.

Variable	**Deviance**	**AIC**
none	131355	215393
DrivAge	132890	216919
State	134329	218314

Ora si procede alla definizione del modello iniziale con tutte le variabili per ClaimAmount, tranne Gender, perchè non significativa.

Si noterà come la variabile State risulta poco significativa

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	8.6371471	0.1951631	44.256	¡ 2e-16 ***
DrivAge18-25	0.0484893	0.0427143	1.135	0.25630
DrivAge26-35	0.0615008	0.0300156	2.049	0.04047 *
DrivAge36-45	-0.0396177	0.0276392	-1.433	0.15176
DrivAge46-55	0.0867542	0.0292380	2.967	0.00301 **
StateAlagoas	-0.0237251	0.2158939	-0.110	0.91250
StateAmapa	-0.0006123	0.3302303	-0.002	0.99852
StateAmazonas	-0.2155279	0.2245954	-0.960	0.33725
StateBahia	0.1649021	0.2014968	0.818	0.41314
StateCeara	-0.0553773	0.2045401	-0.271	0.78659
StateDistrito Federal	-0.0445728	0.2002334	-0.223	0.82384
StateEsperito Santo	0.1439286	0.2045129	0.704	0.48159
StateGoias	0.1683011	0.2003655	0.840	0.40093
StateMaranhao	0.0591854	0.2179770	0.272	0.78599
StateMato Grosso	0.2041662	0.2069704	0.986	0.32392
StateMato Grosso do Sul	0.0308122	0.2072825	0.149	0.88183
StateMinas Gerais	0.0780021	0.1965909	0.397	0.69154
StatePara	0.1098927	0.2133828	0.515	0.60655
StateParaiba	0.0257795	0.2131742	0.121	0.90375
StateParana	0.0795892	0.1966845	0.405	0.68573
StatePernambuco	0.0243540	0.2042765	0.119	0.90510
StatePiaui	0.2173356	0.2214577	0.981	0.32641
StateRio de Janeiro	0.2787102	0.1974907	1.411	0.15818
StateRio Grande do Norte	0.0367002	0.2101874	0.175	0.86139
StateRio Grande do Sul	0.1377249	0.1970165	0.699	0.48452
StateRondonia	-0.2765508	0.3522459	-0.785	0.43240
StateRoraima	-0.2952017	0.3936149	-0.750	0.45327
StateSanta Catarina	0.0938223	0.1975988	0.475	0.63492
StateSao Paulo	0.0841068	0.1951013	0.431	0.66640
StateSergipe	-0.0188034	0.2149877	-0.087	0.93030
StateTocantins	0.4266206	0.2319535	1.839	0.06589 .

Si procede nuovamente con un modello con algoritmo di selezione stepwise basasto su AIC.

Variable	Df	Deviance	AIC
none	-	52017	723523
DrivAge	4	52107	723543
State	26	52264	723547

Table 6: risultato selezione stepwise

per la creazione delle celle tariffarie si selezionano come variabili DrivAge e State. Viene creata la colonna frequenza nelle celle tariffarie come rapporto tra

ClaimNb ed ExposTotal

DrivAge	State	ClaimNb	ExposTotal	ClaimAmount	Frequenza
>55	Acre	20	0.51577857	94155	38.776330
18-25	Acre	3	0.16292663	19276	18.413196
26-35	Acre	22	0.59302646	78273	37.097839
36-45	Acre	46	0.82585813	253242	55.699639
46-55	Acre	17	0.46628148	117683	36.458665
>55	Alagoas	122	4.87688722	709667	25.015957
18-25	Alagoas	27	0.80222931	132650	33.656212

Table 7: celle tariffarie (head)

Vengono inseriti poi i valori delle frequenze nel dataset e calcolate media e varianza del numero di sinistri

$$\text{avg} = 23.76307$$

$$\text{variance} = 3265.839$$

$$\phi = 137.4334$$

3.2 Modello per la Frequenza

Sviluppato un modello iniziale per stimare il numero di sinistri, abbiamo utilizzato le stesse variabili selezionate per la creazione delle celle tariffarie. Per

questo modello, si adotta un Generalized Linear Model (GLM) con una distribuzione di Poisson e un link logaritmico

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.506679	0.096570	36.312	¡ 2e-16 ***
StateAlagoas	-0.394917	0.105145	-3.756	0.000173 ***
StateAmapa	-0.190576	0.164672	-1.157	0.247147
StateAmazonas	-0.245248	0.109931	-2.231	0.025686 *
StateBahia	-0.592137	0.098446	-6.015	1.80e-09 ***
StateCeara	-0.307606	0.099587	-3.089	0.002010 **
StateDistrito Federal	-0.269377	0.098107	-2.746	0.006037 **
StateEsperito Santo	-0.333175	0.100203	-3.325	0.000884 ***
StateGoias	-0.254278	0.098123	-2.591	0.009558 **
StateMaranhao	-0.205611	0.106588	-1.929	0.053727 .
StateMato Grosso	-0.107698	0.100563	-1.071	0.284192
StateMato Grosso do Sul	-0.063164	0.100751	-0.627	0.530702
StateMinas Gerais	-0.387855	0.096975	-4.000	6.35e-05 ***
StatePara	-0.425763	0.103661	-4.107	4.00e-05 ***
StateParaiba	-0.314212	0.104021	-3.021	0.002522 **
StateParana	-0.404710	0.097108	-4.168	3.08e-05 ***
StatePernambuco	-0.684682	0.099419	-6.887	5.70e-12 ***
StatePiaui	-0.088315	0.108713	-0.812	0.416582
StateRio de Janeiro	-0.928535	0.097429	-9.530	¡ 2e-16 ***
StateRio Grande do Norte	-0.256620	0.102390	-2.506	0.012200 *
StateRio Grande do Sul	-0.550651	0.097226	-5.664	1.48e-08 ***
StateRondonia	0.006278	0.180323	0.035	0.972229
StateRoraima	-0.757129	0.203756	-3.716	0.000203 ***
StateSanta Catarina	-0.476157	0.097605	-4.878	1.07e-06 ***
StateSao Paulo	-0.687949	0.096434	-7.134	9.76e-13 ***
StateSergipe	-0.110208	0.104416	-1.055	0.291211
StateTocantins	-0.203138	0.113275	-1.793	0.072921 .
DrivAge18-25	0.568716	0.018581	30.608	¡ 2e-16 ***
DrivAge26-35	0.219596	0.012007	18.290	¡ 2e-16 ***
DrivAge36-45	0.342111	0.010821	31.616	¡ 2e-16 ***
DrivAge46-55	0.136923	0.012020	11.391	¡ 2e-16 ***

Molte delle variabili risultano essere statisticamente significative. Il segno dei coefficienti riflette l'effetto sull'entità del logaritmo del numero di sinistri rispetto al gruppo di riferimento.

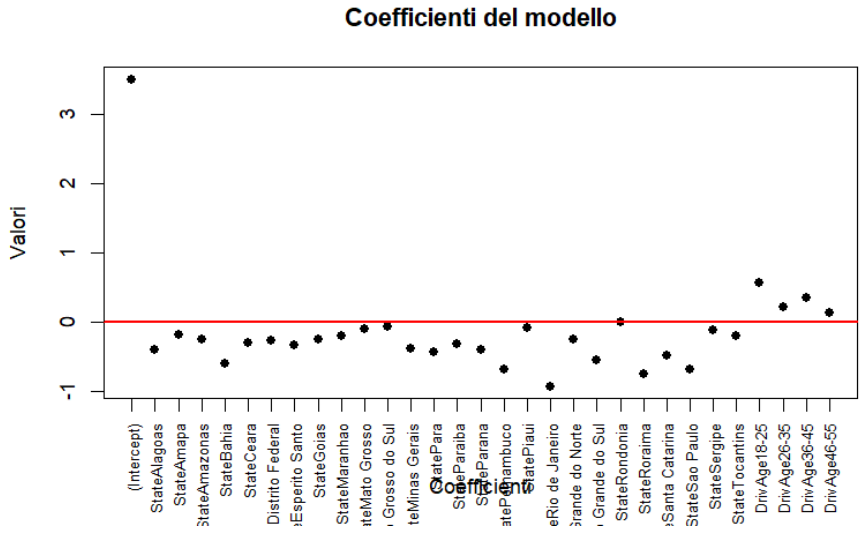

Figure 2: Coefficienti del Modello

Le categorie delle variabili tariffarie che presentano coefficienti superiori aumentano la frequenza di sinistri.

Esempio: il coefficiente associato alla variabile "DrivAge" ha una tendenza a essere maggiore, minore è il range, quindi, per questa categoria la frequenza dei sinistri aumenta di più rispetto alla fascia di età maggiore.

Facendo l'esponenziale dei coefficienti si ottiene un modello moltiplicativo

```
coef_mod_moltiplicat = exp(coefficients(mod_frequenza))
```

Utilizzando la distribuzione quasi-Poisson per creare un nuovo modello per la frequenza, che è utilizzata per gestire la sovradispersione (condizione in cui la varianza dei dati è maggiore della media)

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.506679	0.105152	33.349	¡ 2e-16 ***
StateAlagoas	-0.394917	0.114489	-3.449	0.000562 ***
StateAmapa	-0.190576	0.179306	-1.063	0.287848
StateAmazonas	-0.245248	0.119700	-2.049	0.040477 *
StateBahia	-0.592137	0.107195	-5.524	3.32e-08 ***
StateCeara	-0.307606	0.108437	-2.837	0.004558 **
StateDistrito Federal	-0.269377	0.106825	-2.522	0.011680 *
StateEsperito Santo	-0.333175	0.109108	-3.054	0.002261 **
StateGoias	-0.254278	0.106842	-2.380	0.017317 *
StateMaranhao	-0.205611	0.116060	-1.772	0.076462 .
StateMato Grosso	-0.107698	0.109499	-0.984	0.325339
StateMato Grosso do Sul	-0.063164	0.109705	-0.576	0.564772
StateMinas Gerais	-0.387855	0.105592	-3.673	0.000240 ***
StatePara	-0.425763	0.112873	-3.772	0.000162 ***
StateParaiba	-0.314212	0.113265	-2.774	0.005535 **
StateParana	-0.404710	0.105738	-3.827	0.000129 ***
StatePernambuco	-0.684682	0.108254	-6.325	2.54e-10 ***
StatePiaui	-0.088315	0.118374	-0.746	0.455629
StateRio de Janeiro	-0.928535	0.106087	-8.753	¡ 2e-16 ***
StateRio Grande do Norte	-0.256620	0.111489	-2.302	0.021350 *
StateRio Grande do Sul	-0.550651	0.105866	-5.201	1.98e-07 ***
StateRondonia	0.006278	0.196348	0.032	0.974495
StateRoraima	-0.757129	0.221863	-3.413	0.000644 ***
StateSanta Catarina	-0.476157	0.106279	-4.480	7.46e-06 ***
StateSao Paulo	-0.687949	0.105004	-6.552	5.70e-11 ***
StateSergipe	-0.110208	0.113695	-0.969	0.332382
StateTocantins	-0.203138	0.123341	-1.647	0.099566 .
DrivAge18-25	0.568716	0.020232	28.110	¡ 2e-16 ***
DrivAge26-35	0.219596	0.013074	16.797	¡ 2e-16 ***
DrivAge36-45	0.342111	0.011783	29.035	¡ 2e-16 ***
DrivAge46-55	0.136923	0.013088	10.461	¡ 2e-16 ***

La distribuzione quasi-Poisson consente la stima del parametro phi, che nel modello precedente, con la distribuzione Poisson era fissato a 1. Dalla valutazione dei risultati, il valore di phi si avvicina molto a uno, indicando una bassa dispersione nel modello.

3.2.1 Modello Nullo

Si definisce modello nullo per la frequenza dei sinistri con solo Offset

```
mod_nullo <- glm(ClaimNb ~ offset(log(ExposTotal)),
family = poisson(link = "log"),
                        data = ds, subset=ExposTotal>0, x= TRUE)
```

3.3 ANOVA test

Il modello nullo, ovvero che non include alcuna variabile tariffaria, ha lo scopo di effettuare un test ANOVA con likelihood ratio (LRT).

Questo test ci consente di confrontare il modello nullo con il modello sviluppato per la frequenza, al fine di valutare se quest'ultimo si adatta meglio ai dati rispetto al modello nullo.

I risultati del test indicano che il modello per il numero di sinistri fornisce un miglior adattamento ai dati rispetto al modello nullo.

Aspetti chiave di questo test:

Likelihood Ratio Test (LRT): Questo test viene utilizzato per confrontare la bontà di adattamento tra due modelli annidati (dove un modello è una versione semplificata dell'altro). L'ipotesi nulla è che il modello più semplice (nel nostro caso, modello nullo) sia sufficiente per spiegare i dati.

Modelli annidati: Un modello annidato significa che i parametri del modello più semplice sono un sottoinsieme dei parametri del modello più complesso.
In questo caso:
Il modello nullo (più semplice).
Il modello frequenza è il modello più complesso.

ANOVA con LRT: Confronta le log-likelihood dei due modelli e verifica se il miglioramento nella log-likelihood, passando dal modello nullo a quello più complesso, è statisticamente significativo.
Procedendo col bilanciamento tra i sinistri previsti e osservati, sia nel complesso che a livello singolo per ciascun livello di ogni fattore.

Calcoliamo il numero di sinistri e della frequenza previsti dal modello

```
celle$sinistri_previsti <- predict.glm(mod_frequenza,
                    newdata = celle, type = 'response')

celle$frequenza_prevista <- celle$sinistri_previsti
                            /celle$ExposTotal
```

E successivamente bilanciamo sulla frequenza

```
cat( sum(celle$Freq * celle$ExposTotal),"\n",
sum(celle$frequenza_prevista * celle$ExposTotal))
```

Confrontando i risultati

- Sinistri previsti totali: 70699
- Sinistri totali: 70699
- Differenza (arrotondata): 0

Con un arrontadamento alla sesta cifra decimale l'errore è zero.

La differenza tra sinistri totali reali e sinistri previsti ha un errore inferiore alla sesta cifra decimale.

Si procede col bilanciamento sul singolo livello di ogni fattore:

Trasponendo la matrice X1 ottenuta dal modello delle frequenza e facendo il prodotto matriciale con vettore (la differenza tra il numero di sinistri e il numero stimato di sinistri

Variabile	**Valore**
Intercept	-8.055973e-08
StateAlagoas	-1.419694e-10
StateAmapa	-3.869901e-11
StateAmazonas	-4.599528e-11
StateBahia	-5.392119e-10
StateCeara	-3.377784e-10
StateDistrito Federal	-5.006572e-10
StateEsperito Santo	-2.915179e-10
StateGoias	-4.877479e-10
StateMaranhao	-9.176131e-11
StateMato Grosso	-8.200624e-11
StateMato Grosso do Sul	1.483258e-11
StateMinas Gerais	-1.340735e-09
StatePara	-8.950560e-11
StateParaiba	-1.222851e-10
StateParana	-8.607412e-10
StatePernambuco	-5.710875e-10
StatePiaui	-7.524392e-11
StateRio de Janeiro	-1.310466e-09
StateRio Grande do Norte	-1.313899e-10
StateRio Grande do Sul	-8.304377e-10
StateRondonia	-2.485220e-10
StateRoraima	-6.618637e-08
StateSanta Catarina	-1.689355e-09
StateSao Paulo	-4.292653e-09
StateSergipe	-1.522870e-10
StateTocantins	-6.032080e-11
DrivAge18-25	-4.528872e-09
DrivAge26-35	-2.394255e-08
DrivAge36-45	-2.810578e-08
DrivAge46-55	-1.510297e-08

Il risultato è un vettore di 0s, indica che la somma pesata dei residui per il numero di sinistri è 0.

Successivamente si procede col prodotto matriciale tra il trasposto di X1 e la differenza tra le frequenze calcolate e quelle stimate moltiplicate per l'esposizione.

Variabile	**Valore**
Intercept	-8.055957e-08
StateAlagoas	-1.419814e-10
StateAmapa	-3.870167e-11
StateAmazonas	-4.598484e-11
StateBahia	-5.392155e-10
StateCeara	-3.377797e-10
StateDistrito Federal	-5.006697e-10
StateEsperito Santo	-2.915412e-10
StateGoias	-4.877659e-10
StateMaranhao	-9.175331e-11
StateMato Grosso	-8.199158e-11
StateMato Grosso do Sul	1.482703e-11
StateMinas Gerais	-1.340835e-09
StatePara	-8.950250e-11
StateParaiba	-1.222780e-10
StateParana	-8.606932e-10
StatePernambuco	-5.710589e-10
StatePiaui	-7.524387e-11
StateRio de Janeiro	-1.310438e-09
StateRio Grande do Norte	-1.313876e-10
StateRio Grande do Sul	-8.304581e-10
StateRondonia	-2.485235e-10
StateRoraima	-6.618637e-08
StateSanta Catarina	-1.689285e-09
StateSao Paulo	-4.292420e-09
StateSergipe	-1.522922e-10
StateTocantins	-6.031675e-11
DrivAge18-25	-4.528850e-09
DrivAge26-35	-2.394251e-08
DrivAge36-45	-2.810607e-08
DrivAge46-55	-1.510292e-08

Il risultato è un vettore di 0s, indica che la somma pesata dei residui per la frequenza moltiplicata per l'esposione è 0.

4 Analisi della Severity

Per la stima della Severity bisogna inizialmente rimuovere dal dataset tutte le osservazioni con un numero di sinistri pari a zero.

Il modello è costruito utilizzando un Generalized Linear Model (GLM) della famiglia gamma, con un link logaritmico e le stesse variabili utilizzate per creare le celle tariffarie.

Variabile	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	8.6371471	0.1951631	44.256	¡ 2e-16 ***
DrivAge18-25	0.0484893	0.0427143	1.135	0.25630
DrivAge26-35	0.0615008	0.0300156	2.049	0.04047 *
DrivAge36-45	-0.0396177	0.0276392	-1.433	0.15176
DrivAge46-55	0.0867542	0.0292380	2.967	0.00301 **
StateAlagoas	-0.0237251	0.2158939	-0.110	0.91250
StateAmapa	-0.0006123	0.3302303	-0.002	0.99852
StateAmazonas	-0.2155279	0.2245954	-0.960	0.33725
StateBahia	0.1649021	0.2014968	0.818	0.41314
StateCeara	-0.0553773	0.2045401	-0.271	0.78659
StateDistrito Federal	-0.0445728	0.2002334	-0.223	0.82384
StateEsperito Santo	0.1439286	0.2045129	0.704	0.48159
StateGoias	0.1683011	0.2003655	0.840	0.40093
StateMaranhao	0.0591854	0.2179770	0.272	0.78599
StateMato Grosso	0.2041662	0.2069704	0.986	0.32392
StateMato Grosso do Sul	0.0308122	0.2072825	0.149	0.88183
StateMinas Gerais	0.0780021	0.1965909	0.397	0.69154
StatePara	0.1098927	0.2133828	0.515	0.60655
StateParaiba	0.0257795	0.2131742	0.121	0.90375
StateParana	0.0795892	0.1966845	0.405	0.68573
StatePernambuco	0.0243540	0.2042765	0.119	0.90510
StatePiaui	0.2173356	0.2214577	0.981	0.32641
StateRio de Janeiro	0.2787102	0.1974907	1.411	0.15818
StateRio Grande do Norte	0.0367002	0.2101874	0.175	0.86139
StateRio Grande do Sul	0.1377249	0.1970165	0.699	0.48452
StateRondonia	-0.2765508	0.3522459	-0.785	0.43240
StateRoraima	-0.2952017	0.3936149	-0.750	0.45327
StateSanta Catarina	0.0938223	0.1975988	0.475	0.63492
StateSao Paulo	0.0841068	0.1951013	0.431	0.66640
StateSergipe	-0.0188034	0.2149877	-0.087	0.93030
StateTocantins	0.4266206	0.2319535	1.839	0.06589 .

Presenza di regressori non significativi

Si procede con algoritmo Stepwise basato su criterio di informazione AIC.

Modello	Df	Deviance	AIC
none	5	2017	723523
DrivAge	4	52107	723543
State	26	52264	723547

4.1 Grafici

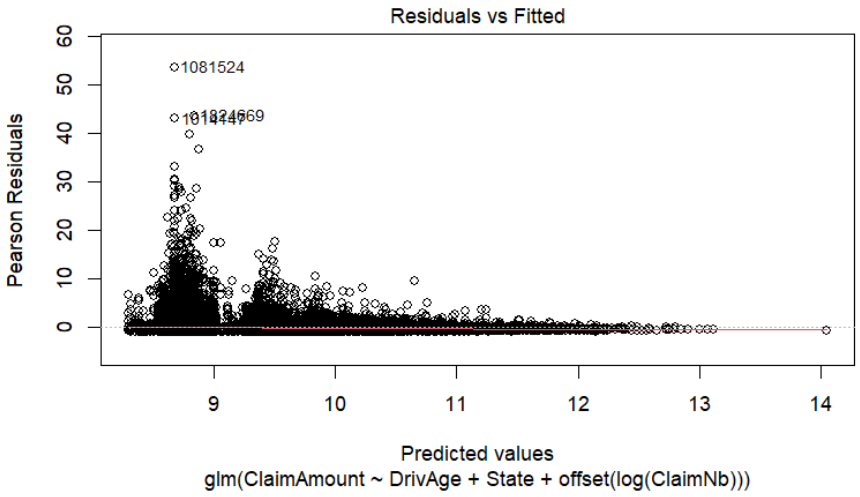

Figure 3: Residui Modello Severity

Dal grafico si nota una presenza di outlier, i punti sono prevalentemente concentrati intorno allo zero.

Gli outlier, esaminati singolarmente, non rappresentano le osservazioni con il maggior risarcimento totale, quindi non sono grandi sinistri.

La concentrazione dei valori intorno allo zero è un buon segno, suggerisce che il modello prevede accuratamente per molti dati.

La minore densità di punti man mano che i residui aumentano può suggerire pochi casi con grandi errori di predizione, (presenza di outliers o parte della variabilità non spiegata dalle nostre variabili è sempre da tenere in considerazione).

La dispersione crescente nel plot (varianza superiore alla media, eterosched) potrebbe spiegare gli elevati residui per i bassi valori predetti.

L'uso di una distribuzione quasi-Poisson o un modello di dispersione negativa potrebbero migliorare il modello.

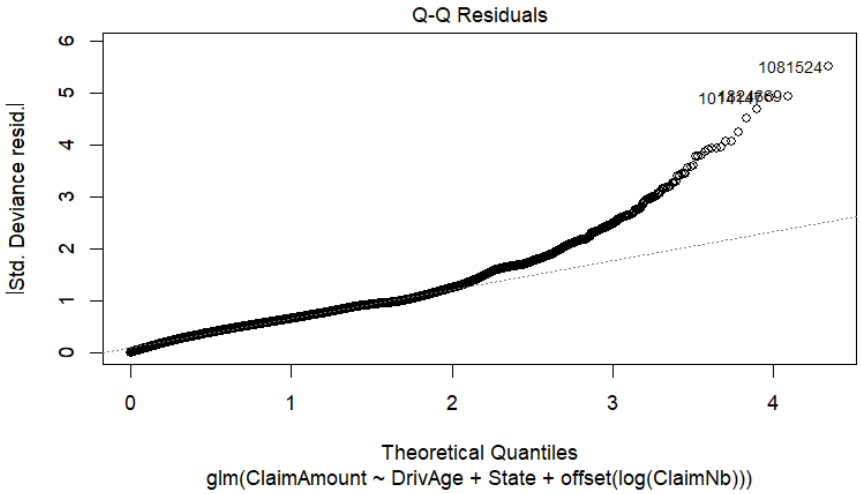

Figure 4: QQ plot Residui Modello Severity

Il Q-Q plot dei residui:

I punti iniziali che partono da zero indicano che i residui standardizzati sono centrati intorno a zero, il che è atteso per un modello ben adattato.

Una salita lenta all'inizio suggerisce che i residui nelle code inferiori (gli estremi negativi) sono vicini ai valori previsti dalla distribuzione normale teorica.

I punti iniziano a deviare dalla linea diagonale e salgono più rapidamente, indica che i residui nelle code superiori (gli estremi positivi) si discostano dalla distribuzione normale teorica.

Questo pattern può suggerire la presenza di outliers o di una distribuzione asimmetrica dei residui.

4.2 ANOVA test - Severity

Successivamente, abbiamo eseguito un test ANOVA che ci ha consentito di confrontare il modello nullo con il nostro modello:

$$p\text{-value} = 0$$

Il modello nullo risulta meno significativo rispetto al modello contenente le variabili.

Si aggiunge quindi al dataset la colonna severity, ovvero il rapporto tra l'ammontare dei risarcimenti e il numero di sinistri

5 Tariffazione e Bilancio

Definito il modello per la stima della frequenza e il modello per la severity;

Viene Calcolato il premio previsto moltiplicando la frequenza prevista per la severity prevista e il premio basato sulla frequenza osservata e la severity osservata

```
celle$premio_previsto <- celle$frequenza_prevista *
                         celle$Severity_prevista
celle$premio <- celle$Freq * celle$Severity
```

Valutare la performance dell'assicurazione::

1. Stimato l'ammontare dei risarcimenti che l'assicurazione deve pagare.
2. I costi totali in termini di risarcimenti ammontano a 394 870 208.
3. Le entrate (premi incassati dall'assicurazione) stimate: 449 049 784.
4. Calcolo del Bilancio totale: differenza tra le entrate (premi) e le uscite (costi):

$$\text{Bilancio totale} = 449\,049\,784 - 394\,870\,208 = 54\,179\,576$$

Saldo Positivo

Bilancio	**Importo**
Costi Totali	394 870 208
Entrate Totali	449 049 784
Bilancio	54 179 576

Infine si aggiugne il saldo (singolarmente) nelle celle tariffarie

```
celle$differenza <- celle$premio_previsto *
                    celle$ExposTotal - celle$ClaimAmount
```

5.1 Celle Tariffarie

	celle.Frequenza	celle.sinistri_previsti	celle.frequenza_prevista	celle.Severity
1	38.776330	17.194703	33.33737	4707.750
2	18.413196	9.592102	58.87375	6425.333
3	37.097839	24.624917	41.52415	3557.864
4	55.699639	38.762661	46.93622	5505.261
5	36.458665	17.825617	38.22931	6922.529
6	25.015957	109.537707	22.46058	5816.945
7	33.656212	31.820706	39.66535	4912.963
8	23.741568	84.842480	27.97630	5115.847
9	35.288747	211.482020	31.62261	4616.936
10	21.586554	119.317087	25.75645	7221.660

Table 8: celle tariffarie, prima metà

	celle.Severity_prevista	celle.premio_previsto	celle.premio	celle.differenza
1	5637.224	187930.25	182549.27	2775.3964
2	5917.305	348373.95	118310.92	37483.3932
3	5994.801	248928.98	131989.05	69348.4724
4	5418.257	254312.50	306641.04	-43215.9511
5	6148.118	235038.27	252386.18	-8089.0068
6	5505.055	123646.71	145516.39	-106655.9519
7	5778.568	229208.93	165351.72	51228.1233
8	5854.247	163780.17	121458.23	128347.8661
9	5291.221	167322.20	162925.90	29401.0787
10	6003.970	154640.93	155890.75	-5789.8254

Table 9: celle tariffarie, seconda metà