Application of Poisson regression in modeling insurance claims in Singapore

Stephen Kiilu
Supervised by
Prof. Viani Djeundje Biatat

December 17, 2020

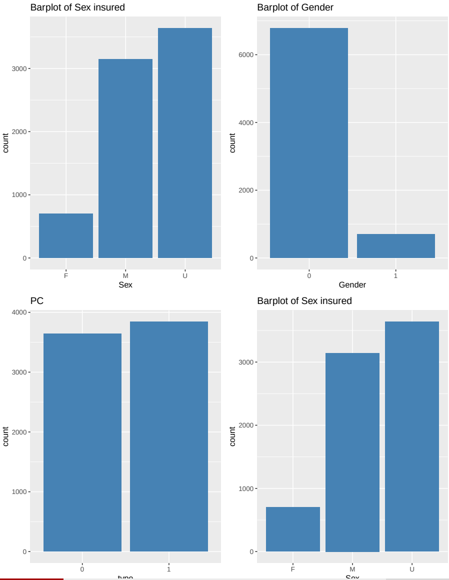
African Institute for Mathematical Sciences
RWANDA

Contents

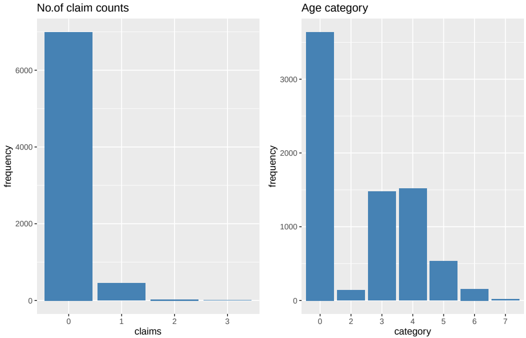
My main task is to work with SingaporeAuto (General Insurance Association of Singapore) data set and use regression model to determine main drivers of insurance claims. The data set provided

	Sex	Gender	VType	PC	Clm	Exp	LNW	NCD	AgeCat
1	U	0	T	0	0	0.67	-0.40	30	0
2	U	0	T	0	0	0.57	-0.57	30	0
3	U	0	T	0	0	0.50	-0.69	30	0
4	U	0	T	0	0	0.91	-0.09	20	0
5	U	0	T	0	0	0.54	-0.62	20	0
6	U	0	T	0	0	0.75	-0.28	20	0

Descriptive analysis



Descriptive analysis cont.



Model building

	Estimate	Std. Error	z value	$\Pr(>\lvert z\rvert)$	
(Intercept)	-1.9086	0.0913	-20.91	0.0000	***
NCD10	-0.2816	0.1247	-2.26	0.0239	*
NCD20	-0.4705	0.1274	-3.69	0.0002	***
NCD30	-0.3680	0.1947	-1.89	0.0588	.
NCD40	-0.7255	0.2436	-2.98	0.0029	**
NCD50	-0.7014	0.1428	-4.91	0.0000	***
Age2	0.3064	0.3110	0.98	0.3247	
Age3	0.4610	0.1169	3.94	0.0001	***
Age4	0.4089	0.1307	3.13	0.0018	**
Age5	0.2389	0.1972	1.21	0.2256	
Age6	0.7595	0.2595	2.93	0.0034	**
Age7	0.8584	0.7166	1.20	0.2310	

(Dispersion parameter for poisson family taken to be 1)
Null deviance: 2716.9 on 7482 degrees of freedom Residual deviance: 2664.6 on 7471 degrees of freedom
AIC: 3683.8
Number of Fisher Scoring iterations: 6

Interpretation

- Having a NCD10 versus NCD0, the log odds for number of claims changes by -0.2816
- Having a NCD50 versus NCD0, the log odds for number of claims changes by -0.7014
- Comparing age category one and age category two, the log odds for the number of claims increases by 0.3064
- Positive coefficient means that as the covariate increase, also the number of claims in a year increase

Confidence intervals

In logistic models, the confidence intervals can be obtained using log likelihood functions, but we can also get them based on the standard errors.

	Estimate	2.5 %	97.5 %
(Intercept)	-1.91	-2.09	-1.73
NCD10	-0.28	-0.53	-0.04
NCD20	-0.47	-0.72	-0.22
NCD30	-0.37	-0.77	-0.00
NCD40	-0.73	-1.24	-0.28
NCD50	-0.70	-0.99	-0.42
Age2	0.31	-0.36	0.87
Age3	0.46	0.23	0.69
Age4	0.41	0.15	0.66
Age5	0.24	-0.16	0.61
Age6	0.76	0.21	1.24
Age7	0.86	-0.95	2.01

Odds-ratio

Often we use odds ratios to interpret logistic models, by just using the same logic as with log odds.

	OR	2.5 %	97.5 %
(Intercept)	0.15	0.12	0.18
NCD10	0.75	0.59	0.96
NCD20	0.62	0.48	0.80
NCD30	0.69	0.46	1.00
NCD40	0.48	0.29	0.76
NCD50	0.50	0.37	0.65
Age2	1.36	0.69	2.38
Age3	1.59	1.26	1.99
Age4	1.51	1.16	1.94
Age5	1.27	0.85	1.84
Age6	2.14	1.24	3.45
Age7	2.36	0.39	7.49

- Holding other predictor variables constant, the odds for the number of claims for NCD10 over NCD0 is 0.75 times i.e it is lower as compared to our reference NCD0
- The odds for age category seven is 136 % higher compared to odd for age category 1.
- The log odd ranges from 0 to 1, while odd ratios range from 1 to infinity. As log odds increase, the odd ratios increase. We use 1 as bas to interpret odd-ratios.

Model fit - Likelihood ratio test

Often we may want to measure how our model fits. The test is too see if the predictors fits well compared to a null model. The test is the difference between deviance between two models. For this model the chi-square of 52.24953, with 11 degrees of freedom and a p-value <0.0001. This means our model fits significantly better than a null model.

Improving model

The goal is to have a good model that signficantly fits our data. Some of ways to improve the model include: adding relevant covariates, checking outliers, missing data and problem of overdispersion.

END.
THANK YOU!!

[1]

J. Bruin, “newtest: command to compute new test @ONLINE,” Feb. 2011.