ACTSC 632 – Assignment 1– Solutions

The purpose of these solutions is to show the results and analysis that you were asked to describe in your report. The solutions are not designed in the form of a report per se. The code used is in a separate file called `sol.R` on LEARN.

1. Download the data set `dataOhlsson` and briefly discuss the data found in there (e.g., how many rating factors, what are the levels for each, how is the exposure determined, etc.). If there are any problems with the data, explain how you dealt with them.

   Solution: as stated in the R documentation for this dataset

   > 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.

   In this data set we have 6 rating factors given by agarald (age of driver),which goes from 0 to 99, kon (sex of driver), either M (male) or K (female), zon (geographical zone, going from 1 to 7, generally going from more to less urban), mcklass (class of the vehicle, with 7 possible classes, as determined by the EV ratio of engine power to vehicle weight), fordald (age of vehicle, a numerical value between 0 and 99), bonuskl (bonus class based on experience, going 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). There are 64548 observations in this data set, with one for each driver observed. We see that for some observations the duration is 0. The exposure (or duration) is determined using policy-years, i.e., how long the contract was in effect for each driver. For each observation we also have the number of claims (antskad) and the claim cost (skadkost).

2. Use only the rating factors contained in the current tariff. Compute the exposure (in policy years), the claim frequency and the average claim severity for each rating factor (i.e., for each possible value of each rating factor).

   Solution: After grouping the classes for the vehicle age into 3 groups (0-2, 3-5, 6 and up) and the bonus class into three groups (1-2,3-4,5-7) we get

rating.factor class	duration	n.claims	totcost (divided by 1000)
Zone	1 6205.3096	183	5539.963
Zone	2 10103.0904	167	4811.166
Zone	3 11676.5726	123	2522.628
Zone	4 32628.4931	196	3774.629
Zone	5 1582.1123	9	104.739
Zone	6 2799.9452	18	288.045
Zone	7 241.2877	1	0.650
Vehicle class	1 5190.3507	46	993.062
Vehicle class	2 3990.1151	57	883.137
Vehicle class	3 21665.6794	166	5371.543
Vehicle class	4 11739.8821	98	2191.578
Vehicle class	5 13439.9260	149	3297.119
Vehicle class	6 8880.1342	175	4160.776
Vehicle class	7 330.7233	6	144.605
Vehicle age	1 4955.4027	126	4964.419
Vehicle age	2 9753.8109	145	5506.945
Vehicle age	3 50527.5972	426	6570.456
Bonus class	1 19893.3698	207	4558.072
Bonus class	2 9615.7644	121	3627.142
Bonus class	3 35727.6766	369	8856.606

   Note that the last column shows the total claim cost (in thousands).

3. Use a relative Poisson glm to determine relativities for the claim frequency, using the current rating factors. Provide a 95% confidence interval for each relativity. Comment on the overall fit of this model to the data.

Solution: we get

factor	factor class	multiplier	LB	UB
0	0	0.002345	0.001858	0.002959
Zone	1	5.156192	4.205633	6.321596
Zone	2	2.725123	2.215810	3.351503
Zone	3	1.708518	1.363530	2.140791
Zone	4	1.000000	1.000000	1.000000
Zone	5	0.906778	0.464785	1.769089
Zone	6	1.035100	0.638608	1.677762
Zone	7	0.727880	0.102011	5.193667
Vehicle class	1	1.478083	1.062390	2.056430
Vehicle class	2	2.103350	1.554868	2.845312
Vehicle class	3	1.000000	1.000000	1.000000
Vehicle class	4	1.321278	1.027817	1.698529
Vehicle class	5	2.045151	1.631045	2.564393
Vehicle class	6	3.979835	3.186922	4.970027
Vehicle class	7	3.311834	1.464417	7.489838
Vehicle age	1	3.239940	2.643934	3.970299
Vehicle age	2	1.894770	1.563719	2.295908
Vehicle age	3	1.000000	1.000000	1.000000
Bonus class	1	1.275967	1.067856	1.524635
Bonus class	2	1.443011	1.171850	1.776917
Bonus class	3	1.000000	1.000000	1.000000

With a deviance of 360.2168 on 389 degrees of freedom, the model seems a reasonable fit (p-value of 0.8495).

Note that I also accepted answers based on the quasi-Poisson family where a dispersion parameter is estimated, causing the CIs to be different from the above.

4. Use a Gamma glm (with log link function) to determine relativities for the severity, still using the current rating factors. Provide a 95% confidence interval for each relativity. Comment on the overall fit of this model to the data.

   Solution: we get the multipliers

factor	factor class	multiplier	LB	UB
0	0	1.570e+04	1.132e+04	2.177e+04
Zone	1	1.300e+00	9.679e-01	1.747e+00
Zone	2	1.370e+00	1.019e+00	1.842e+00
Zone	3	9.364e-01	6.771e-01	1.295e+00
Zone	4	1.000e+00	1.000e+00	1.000e+00
Zone	5	9.634e-01	3.658e-01	2.537e+00
Zone	6	7.845e-01	3.878e-01	1.587e+00
Zone	7	1.765e-02	1.048e-03	2.975e-01
Vehicle class	1	7.459e-01	4.661e-01	1.194e+00
Vehicle class	2	6.673e-01	4.302e-01	1.035e+00
Vehicle class	3	1.000e+00	1.000e+00	1.000e+00
Vehicle class	4	7.976e-01	5.560e-01	1.144e+00
Vehicle class	5	8.330e-01	6.010e-01	1.155e+00
Vehicle class	6	1.035e+00	7.504e-01	1.427e+00
Vehicle class	7	1.433e+00	4.354e-01	4.716e+00
Vehicle age	1	2.556e+00	1.910e+00	3.420e+00
Vehicle age	2	2.345e+00	1.774e+00	3.101e+00
Vehicle age	3	1.000e+00	1.000e+00	1.000e+00
Bonus class	1	8.356e-01	6.455e-01	1.082e+00
Bonus class	2	1.031e+00	7.664e-01	1.386e+00
Bonus class	3	1.000e+00	1.000e+00	1.000e+00

   The severity model doesn't seem to be a very good fit. We reject the hypothesis that the data comes from this model based on the residual deviance, given by 351 on 164 degrees of freedom, as its corresponding p-value is 1.13842e-15.

5. Assess whether rating factors for the policyholder's age and sex would have a significant impact. Include an interaction term. (I leave it up to you to decide how to handle the age variable, e.g., group it into intervals etc.).

   Solution: we have combined the ages into 2 groups given by the breaks 0, 30, 100. When including age and sex with an interaction term, for the frequency the deviance goes to 742.76 on 1277 degrees of freedom. The LRT statistic we get

   742.76- 360.22 =382.55 on 1277-389=888 degrees of freedom. The corresponding p-value is 1, suggesting that the simpler model is a better fit.

   For the severity however, the LRT statistic is given by 253.12 on 124 degrees of freedom, with corresponding p-value given by 7.09×10^{-11}. Hence in this case, the age and sex do appear to provide a model that has a better fit. We also see that for both the frequency and the severity models, while sex is not a significant factor, the interaction term between age and sex and the coefficient for the age group 0-30 are both significant.

6. Now combine your multiplier estimates for the frequency and severity data to get multipliers (and associated 95% confidence intervals) for the premium overall and then propose a new tariff based on your analysis. Compare the results to the old tariff.

factor	factor class	new multiplier	LB	UB	old multiplier
		36.81174	2.465e+01	54.9782	0.000
Zone	1	6.70508	4.684e+00	9.5988	7.768
Zone	2	3.73268	2.601e+00	5.3568	4.227
Zone	3	1.59982	1.078e+00	2.3747	1.336
Zone	4	1.00000	1.000e+00	1.0000	1.000
Zone	5	0.87358	2.693e-01	2.8336	1.734
Zone	6	0.81205	3.456e-01	1.9079	1.402
Zone	7	0.01285	4.117e-04	0.4011	1.402
Vehicle class	1	1.10257	6.206e-01	1.9588	0.625
Vehicle class	2	1.40354	8.237e-01	2.3915	0.769
Vehicle class	3	1.00000	1.000e+00	1.0000	1.000
Vehicle class	4	1.05390	6.790e-01	1.6359	1.406
Vehicle class	5	1.70368	1.145e+00	2.5345	1.875
Vehicle class	6	4.11778	2.786e+00	6.0857	4.062
Vehicle class	7	4.74583	1.120e+00	20.1095	6.873
Vehicle age	1	8.28051	5.805e+00	11.8111	2.000
Vehicle age	2	4.44415	3.167e+00	6.2372	1.200
Vehicle age	3	1.00000	1.000e+00	1.0000	1.000
Bonus class	1	1.06615	7.792e-01	1.4587	1.250
Bonus class	2	1.48750	1.036e+00	2.1367	1.125
Bonus class	3	1.00000	1.000e+00	1.0000	1.000

We see that the new tariff is much higher than the old one for Vehicle Age 1 and 2. We also see that for Zone 5 to 7, the new tariff suggests to lower the premium compared to the baseline of Zone 4, while for the old tariff they were all higher than the baseline. The number of claims for these 3 zones is very small though, so it seems like it might be best to not make such an important change based on such a small sample.

7. Comment on any further analysis that should be considered before deciding on a final tariff.

Solution: we saw that, especially for the severity, the age and sex seem to be significant factors that should be considered. On the other hand, looking at the results for coefficients in each of the frequency and severity models (reproduced below), we see that Zones 5,6,7 (note that although the results below are based on the ordering of levels by decreasing order of duration, and therefore do not necessarily correspond to the original numbering, for these 3 particular classes are actually the same as in the original ordering) do not seem to be significant, and should therefore probably be combined with the baseline of Zone 4. Finally, we should explore the use of models other than the gamma distribution for the severity, to see if a better fit can be obtained.