# Report II

*Isac Karlsson*
*Ludvig Wärnberg Gerdin*

March 11, 2020

# Contents

Table 1: Data example

| RiskYear | VehicleAge | Weight | Climate | ActivityCode | Duration | NoOfClaims | ClaimCost |
|---|---|---|---|---|---|---|---|
| 2010 | 009 | 3830 | North | Construction | 0.63 | 1 | 627099 |
| 2008 | 001 | 400 | South | Missing | 0.59 | 1 | 253850 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

# 1 Introduction

## 1.1 Background

Most of the tractors in Sweden have a third party liability insurance, because they are required by law. In southern Europe a few large players have dominated the sales of tractor insurances. Our main task this project is to create our own tractor tariff. The price $p_i$ for tariff cell $i$ is defined as

$$p_i = \gamma_0 \prod_{k=1}^{M} \gamma_{k,j} \tag{1}$$

Here $\gamma_0$ corresponds to the base level and $\gamma_{k,j}$ are the risk factors for variable $k$ and variable group $j$. For example, letting the variable `VehicleAge` correspond to variable number $k = 1$ and letting a particular variable group `VehicleAge` $< 2$ correspond to $j = 2$, then the risk factor for the `VehicleAge` of a tractor that is 1 year old would be $\gamma_{12}$l

## 1.2 Data

*If P&C* have provided us with data to train this price model, example given in the Table 1.

## 1.3 Problem description

### 1.3.1 Risk Differentiation and Grouping

Using GLM analysis we aim to make each group Risk homogeneous and that they contain enough data to get a stable GLM analysis, meanwhile handling imperfections in the dataset.

### 1.3.2 Levelling

Here we aim to calculate $\gamma_0$ such that the forecasted claim costs for each insurance are covered by the the price for each insurance on a full year basis. We use a ratio between the estimated claim cost and the total premium of 90%. Lastly we calculate the base level $\gamma_0$ from the formula given in (1)

# 2 Methods and Methodological Considerations

## 2.1 Grouping and Risk Differentiation

The criteria on which we based our grouping was

1. Each group should be risk homogeneous, and

2. Each group should have enough data to make the GLM estimates stable.

Greater emphasis were placed on fulfilling criteria 2) due to it being more concrete. In order to do that we choose cut-offs that placed a fairly equal shares of data in each risk group.

The resulting cut-offs and risk groups are found in section 3.

## 2.2 Levelling

From the results of section 2.1, we get risk factor estimates for each level of each predictor. These are henceforth referred to as the "group factors".

For each corresponding sub-assignment presented under Levelling in the project description, we conducted the following:

1. From the original data, we selected those rows (or tractors) that had a `RiskYear` 2016. That way the GLM analysis were only conducted on the active customers, leaving out those that weren't customers to If in 2016.

    Following the GLM-script each row were aggregated to tariff cells. From the aggregated data we calculated the expected yearly claim-cost per tariff cell by dividing the claim cost by the duration for each row. The rationale for this was to enable us to analyse the yearly cost, even if the insurances had not been active for all of 2016.

    The estimated claim cost for the coming year is then simply the sum of all of the estimated yearly claim costs for each tariff cell.

2. Recall first that the *If P&C* has a ratio target of 0.9 between the estimated claim cost and the total premium. Using that the total premium is $P = \sum_i p_i$, where $p_i$ is the premium or price for tariff cell $i$ defined in (1), and the total expected yearly cost is $C = \sum_i c_i$, where $c_i$ is the expected yearly cost for tariff cell $i$, we get

$$\frac{C}{P} = 0.9 \tag{2}$$

    Using that we have an estimated value for the total expected claim cost, we reorder the equation to get

$$\frac{C}{P} = 0.9 \iff \frac{C}{0.9} = P \tag{3}$$

3. The formula for the price of tariff cell $i$ is defined in (1). As a reminder, this was

$$p_i = \gamma_0 \prod_{k=1}^{M} \gamma_{k,j}$$

    Inserting (1) into the formula for the total premium we get

$$P = \sum_i p_i = \sum_i \left( \gamma_0 \prod_{k=1}^{M} \gamma_{j,k} \right)_i = \gamma_0 \sum_i \left( \prod_{k=1}^{M} \gamma_{j,k} \right) \qquad (4)$$

and subsequently inserting (4) into (3) and solving for $\gamma_0$

$$\frac{C}{0.9} = \gamma_0 \sum_i \left( \prod_{k=1}^{M} \gamma_{j,k} \right) \implies \frac{C}{\sum_i \left( \prod_{k=1}^{M} \gamma_{j,k} \right)_i} = \gamma_0 \qquad (5)$$

In other words: Row by row in the aggregated data, we map the row characteristics to the corresponding risk factor in the group factor table, and subsequently take the product of all the risk factors to obtain the total risk factor for that tariff cell. By (5), we obtain the base level $\gamma_0$ by dividing the total expected cost by the sum of all total risk factors.

## 3  Results

We adjusted the data slightly to adjust for some odd rows. Moreover, the rows with where `Duration` $= 0$ were removed. Since we are interested in the tractors that has been exposed to risk, and the tractors where `Duration` $= 0$ make up a small fraction of the total data, we decided to remove these rows.

   We noted that some rows had suspiciously low values for the `Weight` predictor, e.g. a weight of 0. Since the fraction of rows with `Weight` $= 0$ were small (0.02%), removing the rows would not have a large impact on the results. However, since many of these rows correspond to a particular ActivityCode (namely Middle H - Hotels and restaurants), we believe that we are missing the contextual dimensions needed to decide whether these should be counted as wrong inputs in the dataset. In the end we decided to leave them in the data as a part of the group $< 1000$. An alternative would have been to include those rows as a separate level, however in that case the risk factor corresponding to this level would have been inflated (since a claim on a small duration on would have resulted in a predicted claim frequency), which is undesirable.

   As an other note we identified the use of Other as a factor level to `ActivityCode`. We assume that this level can be mapped to more specific types of businesses internally by If. In a future version of this model, the model could input more granular groups of `ActivityCode` to potentially improve the performance of the model.

   The variable groups and corresponding estimated risk factors are presented in Table 2. For each sub-question in the Levelling assignment we got the following results

1. The total expected cost for 2017 would be 170033.9 kr.

2. The total premium would be 188926.6 kr, and

3. Mapping the risk factors to each respective tariff cell and calculating $\gamma_0$ we got

$$\gamma_0 = 238.8046$$

In order to evaluate our model we used the Akaike Information Criterion (AIC). Considering e.g. the qualitative importance of the age of the insured tractors when estimating the claim severity, we decided to test whether leaving out this predictor would reduce the predictive performance of the model. The results of this evaluation are presented in Table 3. The AIC was lower for the both models when keeping the `VehicleAge_group` predictor, hence we conclude that this predictor should be included in the model.

A more exhaustive and thorough way of evaluating the model would have been to run all-possible regression with AIC in order to evaluate the importance of the other predictors. This, however, is left for a future analysis.

Table 2: Variable groups and corresponding risk factors

| rating.factor | class | duration | n.claims | rels.frequency | rels.severity | rels.risk |
|---|---|---|---|---|---|---|
| Weight | 0-500 | 11956.7190 | 34 | 0.2712058 | 1.2269942 | 0.3327679 |
| Weight | 500-1000kg | 10039.1199 | 45 | 0.4310375 | 0.9871136 | 0.4254830 |
| Weight | 1000-2500kg | 12898.6416 | 119 | 1.0000000 | 1.0000000 | 1.0000000 |
| Weight | 2500-4000kg | 10866.1968 | 111 | 1.1950849 | 1.4231778 | 1.7008183 |
| Weight | >5000kg | 8670.5903 | 179 | 2.1359441 | 2.2950121 | 4.9020175 |
| Climate | Middle | 21991.9321 | 188 | 0.9630260 | 0.7297804 | 0.7027975 |
| Climate | North | 8887.6077 | 89 | 1.1602113 | 0.9262822 | 1.0746830 |
| Climate | South | 23551.7278 | 211 | 1.0000000 | 1.0000000 | 1.0000000 |
| ActivityCode | A - Agriculture, Hunting and Forestry | 9530.0000 | 106 | 1.1699222 | 0.8700959 | 1.0179445 |
| ActivityCode | C - Mining and quarrying | 1324.3479 | 13 | 1.1349216 | 0.8658600 | 0.9826832 |
| ActivityCode | F - Construction | 2504.5092 | 44 | 1.9708873 | 2.0598955 | 4.0598220 |
| ActivityCode | G - Wholesale & retail trade; repair of motor vehicles, household | 1353.1258 | 14 | 1.3280599 | 1.6849378 | 2.2376984 |
| ActivityCode | H - Hotels and restaurants | 1245.1478 | 19 | 1.6014217 | 1.1446289 | 1.8330336 |
| ActivityCode | I - Transport, storage and communication | 475.7573 | 2 | 0.6920397 | 0.8546421 | 0.5914463 |
| ActivityCode | L - Public administration and defence; compulsory social security | 5639.0372 | 48 | 1.4553348 | 0.9141597 | 1.3304085 |
| ActivityCode | M - Education | 749.2439 | 5 | 0.8098397 | 1.0885207 | 0.8815273 |
| ActivityCode | Missing | 27273.8746 | 151 | 1.0000000 | 1.0000000 | 1.0000000 |
| ActivityCode | N - Health and social work | 2661.5378 | 66 | 3.1369286 | 0.9160351 | 2.8735366 |
| ActivityCode | Other | 1674.6861 | 20 | 1.2931341 | 1.0045287 | 1.2989902 |
| VehicleAge | 01_<1years | 18995.4460 | 240 | 1.0000000 | 1.0000000 | 1.0000000 |
| VehicleAge | 01_1-4years | 17305.7762 | 143 | 0.6150547 | 1.0155823 | 0.6246386 |
| VehicleAge | apa | 18130.0454 | 105 | 0.3345154 | 0.9030982 | 0.3021003 |

Table 3: AIC for both models with and without VehicleAge$_g roup predictor$

|  | With VehicleAge__group | Without VehicleAge__group |
|---|---|---|
| Frequency model | 934.7548 | 1101.529 |
| Severity model | 9229.4179 | 9393.902 |

# 4 Conclusion

For each sub-question in the Levelling assignment we found that

1. The total expected cost for 2017 would be 170033.9 kr.

2. The total premium would be 188926.6 kr, and

3. $\gamma_0 = 238.8046$