# SF2930 Regression analysis VT2021

Project 2

The project should be done in groups of **two**.

A computer written[1] report should be handed in on Canvas no later than **2021-03-14**. Name the document "SF2930Project2-FullName1-FullName2.pdf".

## The report

The report should be handed in at the latest **2021-03-14** and should be *at most* 5 pages long. In your report, present and explain your choice of risk arguments, grouping of data and risk factors. How does this comply with Likelihood Ratio Test and different measures for goodness of fit discussed in this course? Perform at least one test to motivate your choice of model!

## Introduction

The travel insurance that is normally a part of your home insurance does not cover travel done as part of your occupation. Therefore, companies that require some travel buy a separate travel insurance for their employees. The insurance cover a range of different situations such as lost luggage, delays, medical costs etc. One important part of the insurance is the compensation that is paid out in case of cancellations.

Travel insurance has of course been of high interest the last year due to the current pandemic situation, and If has asked you to help them review their price models for this part of the insurance. They need you to create a price model that can price their commercial customers on the form

$$price = \gamma_0 \prod_{k=1}^{M} \gamma_{k,i} \tag{1}$$

where $\gamma_0$ is the base level and $\gamma_{k,i}, k = 1, ..., M$ are the risk factors corresponding to variable number $k$ and variable group number $i$. $\gamma_{k,i}$ will take different values for each individual company, depending on that company's characteristics. For example, let $k = 1$ be number of persons (NoP) insured and for one particular company the NoP is 27. Then, according to the table below, $\gamma_1 = 10$.

[1] Preferably using LaTeX

Normally, assigning the same factor regardless of if the company has 11 or 29 employees makes for a very poor price formula, but for the sake of this example we are keeping it simple. In your hand-ins you typically want to fit a continuous curve that matches the output factors you get from the regression.

| Number of Persons group $i$ | Risk factor $\gamma_{1,i}$ |
|---|---|
| 1: # <= 1 | 0.7 |
| 2: # <= 5 | 2.2 |
| 3: # <= 30 | 10 |
| 4: # = 100 | 25 |
| 5: # >= 100 | 35 |

# Material

### 1. Dataset

The file Cancellations.csv contains information on all companies with a Business Travel insurance in If P&C during 2015-2020, including claims history. The file has one row per company and Risk year, as shown in the table below.

| RiskYear | #Persons insured | Financial Rating | ... | Activity | Duration | NoOfClaims | Claim cost |
|---|---|---|---|---|---|---|---|
| 2010 | 9 | AA | ... | Construction | 0.63 | 1 | 67 099 |
| 2008 | 9 | A | ... | Missing | 0.59 | 1 | 25 850 |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ | ⋮ | ⋮ |

Here, Risk year is the year of the insurance period, #Persons insured denote the number of persons covered by the insurance, Financial rating is a measure of the company's economic **capacity** and Activity is the activity code registered on the company which defines what segment the company is active in. For each company, there is also information regarding Duration. This is the share of the risk year the company was insured. For example, a company with a one year insurance policy from 2015-07-01 to 2016-06-30 will be represented by two rows in the data; one with Risk year = 2015 and one with Risk year = 2016, both with Duration = 0.5. Finally, the number of claims and claims cost corresponding to the insurance period are denoted by NoOfClaims and ClaimCost.

See appendix A for a thorough description of the contents of the dataset.

### 2. GLM program

The template GLM.R contains a structure for a GLM analysis.

# Tasks

### 1. Grouping and risk differentiation

Perform a GLM analysis to figure out how best to describe the risk for the different companies. Use the template GLM.R. The outcome should be a multiplicative GLM

model, as described in Eq. 1, that model claims frequency and claim severity separately. Use the same variables and variable groups in both models, and propose the final risk factor $\gamma_{k,i}$, where the final risk factor is the product of the claim frequency and the claim severity.

In order to perform your GLM analysis, you will have to group some of the variables. Consider, for example, the number of persons. These cover a very wide range, as some companies want to insure only a single employee while there are companies with several hundred persons to insure. Thus, it would be impossible to analyze each individual amount alone; it is necessary to group them. When grouping a variable, there are two things to consider:

- Make each group "Risk homogeneous", meaning that you believe that the risk does not vary much within the group, with regard to the particular variable.
- Create groups with enough data to get a stable GLM analysis for each group. What is "enough" has no clear answer, bur varies, depending among other things on how many variables you use in your analysis.

Creating good groups is usually an iterative process, so try different ways to do it!

No dataset is perfect. You will find many rows with strange, missing, or incomplete data, and need to handle this. One good strategy is to put all these values in a group of its own, letting it get its own factor in the GLM analysis.

Remember to consider the Likelihood Ratio and other goodness of fit tests when testing different models. As always when modelling real data there is no ultimate solution that will capture all aspects of the risk, but there are multiple different trade offs to consider. For example having many groups for a variable might result in detailed risk explanation on the used data set but also worse results in overfitting tests.

**2. Leveling**

Having found the risk factors $\gamma_{k,i}$, determine the base level $\gamma_0$. Note that *a value* for $\gamma_0$ is estimated automatically by the GLM program, however, this value corresponds to the total claims cost of the analysis data, not the insurances that are active today. The purpose of leveling is to set $\gamma_0$ such that the price for each insurance on a *full year basis* covers its forecasted claim costs.

1. Start by estimating the claim cost for the coming year. Assume that the customers you have now would extend their insurance for a full year, what would be the claim costs for these insurances? Note that not all insurances in risk year 2020 was active for the entire year.
2. Assume that If P&C has a ratio target of 90%. The ratio target is defined as the ratio between the estimated claim cost and the total premium, $Target =$

$SUM(Claims)/SUM(Premium)$ – what should the total sum of the companies' premium be to accommodate this target?

3. Then we need to determine the corresponding total risk for the portfolio, this corresponds to the product in formula (1). First, calculate each insurance's "total risk factor" - i.e. the product of all risk factors $\gamma_{k,i}$ for that insurance then summarize it for the portfolio.

4. Now you can find the base level, $\gamma_0$, that makes the total expected premium of your portfolio match what you calculated in the previous steps.

Remember to only include the active insurances when doing the above calculations!

Good luck!

# A The data

| Variable | Description | Values |
|---|---|---|
| Activity | A column describing what the company does | A - Agricultural activities<br>B - Mining<br>C - Manufacturing<br>D - Production and distribution of electricity/power/heat<br>E - Recycling<br>F - Construction<br>G - Wholesale & Retail stores<br>H - Transport & Logistics<br>I - Restaurant & hotels<br>J - Consultants<br>K - Finance, Insurance, etc.<br>L - Property Owners<br>M - Lawyers, accountants<br>N - Leasing<br>O - Public entities<br>P - Education<br>Q - Care<br>R - Culture<br>S - Service companies (hair dressers, reparations, etc.)<br>T - Service companies (hair dressers, reparations, etc.)<br>U - Mining<br>X - Missing |
| ClaimCost | Total claim cost | |
| CompanyAge | For how long the company has been registered | |
| Dangerous areas | Describes whether travel to different dangerous areas are covered by the insurance | Excluded/Not excluded |
| Duration | The time in years that an insurance has been active for during the year | |

| Variable | Description | Values |
|---|---|---|
| DurEnd | From when the insurance starts or when the risk year starts | |
| DurStart | From when the insurance ends or when the risk year ends | |
| Financial Rating | Describes a company's credit risk. Typically geared towards the banks' risk when granting loans, but it gives some useful information about how well the company is run. | **[C - AAA]** - C- worst, AAA-best<br>**[AN]** - Newly started company, no information<br>**[IR]** - Missing |
| ID | Unique id for each row | |
| Num_of_ft_Employee | Number of full time employees registered online | |
| Number of Persons | Number of people the insurance covers | |
| NumberOfClaims | Number of registered claims | |
| RiskYear | Which year the insurance is active | 2016-2020 |
| Travelling Area | Which area of travel the insurance covers. | **Abroad (whole world)**: The entire world<br>**Abroad in Europe**: Within Europe<br>**Abroad in Nordic Country**: Within the Nordic countries<br>**In one's home country)**: Within your country of origin |