Insurance Data Science
- Assignment -

2023/2024 From 07/06/2025 to 27/06/2025

DATA:

File autodata.txt

File claimsdata.txt

PROBLEMS:

Consider you are working for the IMS Insurance Company.
You are asked to evaluate last year's claims data of the Automobile Insurance portfolio.

Part I

As a first task, please address the following points in a concise report:

1. **Exploratory Analysis of Claim Counts**
   Conduct an Exploratory Data Analysis (EDA) of the Number of Claims related to Third Party Liability in Automobile Insurance. Describe the key features of the data and identify any notable values or trends that help explain the underlying claim behavior.

2. **Descriptive Analysis of Claim Severity**
   Perform a descriptive statistical analysis of Claim Severity for Third Party Liability in Automobile Insurance. Highlight key statistics, discuss the distribution of the data, and note any relevant patterns or observations that help characterize the severity of claims.

3. **Graphical Analysis of Variable Interactions**
   Use graphical techniques to explore potential interactions between explanatory variables in relation to both Number of Claims and Claim Severity. Present your findings using appropriate visualizations (e.g., scatter plots, boxplots, interaction plots, heatmaps) and provide commentary on any patterns, relationships, or insights revealed through the analysis.

4. **Distribution Fitting for Claims Data**
   Fit appropriate probability distributions to both the Number of Claims and the Claim Severity data:

   - For the Number of Claims, remove the highest outlier before fitting. Clearly mention the removal in your report and proceed with the analysis using the cleaned dataset.
   - For Claim Severity, define an upper threshold that enables fitting a distribution from the Exponential Family. Specify the chosen threshold, the number of claims excluded, and provide justification for your decision in the report.
   - Illustrate the fitted distributions with an appropriate plot, a statistical test and include the parameter estimates on your report.
   - Calculate the mean and standard deviation of the removed claims based on the threshold applied above. Display these excluded claims using a histogram and a boxplot. Offer a brief commentary on the nature of these claims and share your view on how the insurer might account for them in the premium structure.

Part II

You are now asked to propose a **Pricing Structure** for the Automobile Insurance portfolio. Using the insights and results obtained in Part I, please address the following tasks:

1. **Modeling Claim Frequency with GLM**
   Fit a Generalized Linear Model (GLM) to the *Number of Claims* data to estimate the claim frequency for each risk profile in the portfolio.

   - Clearly state and justify your modeling assumptions and variable choices.
   - Enhance your model using appropriate statistical tests and model selection criteria.
   - Evaluate and comment on the model's overall fit.
   - Identify the **Standard Insured** profile and provide the corresponding claim frequency estimate.
   - Determine the insured profiles with the **highest** and **lowest** claim frequency risk, and estimate the claim frequency for each.

2. **Modeling Claim Severity for Common Claims with GLM**
   Fit a GLM to the *Claim Severity* data, focusing on the subset of what you define as "*common*" claims.

   - Provide a clear and reasoned definition of what constitutes a "common" claim.
   - Justify your modeling choices and assumptions.
   - Refine your model using appropriate statistical techniques.
   - Assess the model's performance and goodness of fit.
   - Identify the **Standard Insured** and the corresponding estimate of claim severity.
   - Identify the profiles with the **highest** and **lowest** claim severity risk, along with their respective severity estimates. Comment on the those features, comparing with the highest and lowest risk profiles from the Claim Frequency model.

3. **Proposing a Pricing Structure for Common Claims**
   Develop a pricing structure for the subset of *common claims*, based on your previous modeling.

   - Identify the risk profiles with the **highest** and **lowest** risk profile, and specify the corresponding premiums to be charged.
   - Provide any additional comments regarding fairness, adequacy, or implications of the proposed premium structure.

4. **Modeling and Pricing for Large Claims**
   Propose a model to incorporate *large claims* into the overall pricing structure. Consider using a Machine Learning approach to accurately estimate the probability of a large claim being reported.

   - Justify your choice of model (e.g., logistic regression, decision trees, random forests, gradient boosting, neural networks).
   - Discuss the variables used, the rationale for separating large claims from common claims, and how this model contributes to the final pricing.
   - Suggest how the output of this model could be integrated with the pricing of common claims to produce a comprehensive premium.

DELIVERABLES:

- Report (around 12 pages) with answers and comments to the insurer.
- R/Python files developed to complete the report.