Project: Modeling Insurance Claims Frequency and Severity

LACTU2150 - Statistical Analysis of Insurance Data

Introduction

In this project, you will analyze an insurance portfolio of motor third party liability (MTPL) policies from a Belgian insurer, covering data from the year 1997. Each observation represents a unique policyholder, with risk factors recorded at the beginning of the policy period.

The goal is to model the frequency and severity of claims using Generalized Linear Models (GLM) and Generalized Additive Models (GAM) and to identify the most significant features based on model selection criteria. You will explore:

1. A **Log-Poisson model** for the claim frequency (number of claims filed).
2. A **Log-Gamma model** for the average claim severity (average amount per claim).

Data Description

The dataset called `mtpl_be.rda` located in the subfolder **data** contains 163,231 policyholders and the following variables:

Variable	Description
nclaims	The number of claims filed by the policyholder.
exp	The fraction of the year 1997 during which the policyholder was exposed to the risk.
amount	The average amount per claim in Euros (defined as the total claim amount divided by the number of claims).
coverage	Type of coverage provided by the insurance policy: TPL, PO, or FO. TPL = only third-party liability, PO = partial omnium = TPL + limited material damage, FO = full omnium = TPL + comprehensive material damage.
fuel	Type of fuel of the vehicle: gasoline or diesel.
sex	Gender of the policyholder: male or female.
use	Main use of the vehicle: private or work.
fleet	The vehicle is part of a fleet: yes or no.
ageph	Age of the policyholder in years.
power	Horsepower of the vehicle in kilowatt.
agec	Age of the vehicle in years.
bm	Level in the former compulsory Belgian bonus-malus scale (0 to 22, where a higher level indicates a worse claim history).
long	Longitude coordinate of the center of the municipality where the policyholder resides.

Variable	Description
lat	Latitude coordinate of the center of the municipality where the policyholder resides.

Objectives

1. **Frequency Modeling**: Use a Poisson regression with a log-link to model `nclaims` (number of claims) as a function of other features.
2. **Severity Modeling**: Use a log-Gamma regression model for the **average amount per claim**, calculated as `amount / nclaims`, to assess the impact of different factors on claim size.

Instructions

Part 1: GLM Analysis

1. **Data Preparation**
   - Create a new variable, `avg_amount`, representing the average amount per claim: `avg_amount = amount / nclaims`.
   - For categorical variables like `coverage`, `fuel`, `sex`, `use`, and `fleet`, ensure they are set as factors.
   - Bin continuous variables such as `ageph`, `power`, and `bm` to simplify analysis.
2. **Exploratory Data Analysis**:
   - Conduct exploratory analysis to visualize the relative frequency of each feature (`ageph`, `coverage`, `fuel`, `sex`, `use`, `fleet`, `ageph`, `power`, `agec` and `bm`). Provide one plot for each feature.
3. **Frequency Model (Poisson)**
   - Fit a Poisson regression model with a log-link to predict `nclaims`, using exposure (`exp`) as an offset.
   - Exclude geographic variables (`long`, `lat`) from this initial model.
   - Select features based on:
     - **BIC**: Use BIC to assess model complexity and fit.
     - **ANOVA test**: Perform nested model comparisons to determine the significance of additional variables.
     - **Cross-validation**: Implement a k-fold cross-validation to evaluate model stability and predictive performance.
4. **Severity Model (Log-Gamma)**
   - Fit a Log-Gamma regression model to predict `avg_amount`.
   - Exclude geographic variables in this part as well.
   - Perform feature selection using BIC, ANOVA, and cross-validation as for the frequency model.
5. **Report Results**
   - Summarize the models, highlighting the most important features based on BIC, ANOVA tests, and cross-validation.
   - Include model coefficients and interpretation of the effect of each significant predictor.
   - Plot the insurance premium as a function of `ageph` and `sex`.

Part 2: GAM Analysis with Geographic Smoothing

1. **Data Preparation**
   - Include `long` and `lat` in the dataset for modeling spatial effects.
2. **Frequency Model with Smoothing**
   - Fit a Poisson-GAM model, using `long` and `lat` as smoothed covariates to capture geographic patterns.
   - Use smoothing functions for continuous variables such as `ageph`, `power`, and `bm`.
   - Evaluate the model using:
     - **BIC**: Check if the GAM model improves over the GLM by comparing BIC values.
     - **ANOVA test**: Conduct tests for the significance of smoothed terms.

  - **Cross-validation**: Use k-fold cross-validation to assess predictive accuracy.
3. **Comparison of GLM and GAM**
   - Summarize the impact of adding spatial smoothing on model accuracy and interpretation.

Submission Requirements

1. **R Markdown Report**: Submit an R Markdown (`.Rmd`) file containing your full analysis, including code, output, and interpretations. Please verify carefully that the file compiles without errors!
2. **HTML Output**: Render the R Markdown file to an HTML document to include all outputs in a single, viewable file.
3. **Zipped File**: Place both the `.Rmd` and `.html` files in a zipped folder.

Please send the zipped file on Moodle by **Sunday, January 12, 2025 at 23:59**.
