Claims frequency regression modelling case

Introduction

Insurance pricing is typically done by creating two prediction models, one for *claim frequency* and the other for *claim severity*. The pricing is then based on the *risk* which is obtained by multiplying the predicted claim frequency with the predicted claim severity. In this case you will create a linear regression model for claim frequency for vehicle insurance.

The claim frequency is defined as

$$claim\ frequency\ = number\ of\ claims\ /\ duration$$

where the duration is the amount of time the insurance policy has been active, and is measured in years.

Case - part 1

Create a regression model with the claim frequency as dependent variable using your analytical tool of choice, e.g. Python or R, and the data supplied in the data.csv file. The data includes historical vehicle insurance data. Choose maximum 3 of the 5 independent variables, owner_age, owner_gender, geo_zone, vehicle_class and vehicle_age to predict the claim frequency.

Case - part 2

Use some metric to evaluate the fit of the model and compare it with a different choice of explaining variables.

Presentation

Please prepare to present your work. Think through your choices and possible ways of improving the model.

Data

The data in data.csv includes the following columns.

- *owner_age*: the age of the owner
- *owner_gender*: the gender of the owner, M (male) or K (female).
- *geo_zone*: geographic zone numbered from 1 to 7, based on the address of the owner
- *vehicle_class*: a classification based on engine power, vehicle weight, 7 classes in total
- *vehicle_age*: vehicle age, between 0 and 99
- *duration*: the number of policy years
- *n_claims*: the number of claims
- *claim_cost*: the claim cost

The data unfortunately has some inconsistencies, please take reasonable measures to clean the data in order to use it for the modelling.