Instituto Superior Técnico
Departamento de Engenharia Civil, Arquitetura e Georrecursos
Course: Transport Demand Modeling
2020/2021

Home Assignment 2: Generalized Linear Models

due Wednesday, Dec. 9th 2020 (24hours)

1. Objectives

The objective of this HA is to expand your ability to model real phenomena using Generalized Linear Models. This, home assignment presents a case study of road accidents estimation, based on a sub-sample of a dataset collected for Ana Fernandes' PhD Thesis which aimed to develop a maintenance model based on the costs generated by road accidents. The current assignment will test your ability on formulating count data models and evaluate them.

2. Software

The reference software is R (or SPSS) and ancillary calculations may be done with basic Excel functions.

3. Data

The dataset to be used in this assignment is available at the course's Fenix webpage (DataBase_TDM_GZLM_Accidents.xlsx or DataBase_TDM_GZLM_Accidents.sav).

The dataset contains data about 157 road sections (all from highways and 1km long) and the registered road accidents occurred between 1997 and 2002.

4. Your tasks

1. Compute the 3 different specifications of count data models for this dataset, including one base model containing just the explanatory variable "IFI".

2. Perform the statistical tests for goodness-of-fit of each model and compare them (don't overlook possible overfitting).

3. Based on the complexity and goodness-of-fit of the tested models, select the best according to your criteria that you should specify.

5. Report content

Your final report should include:

1. Describe your *a priori* assumptions of the available variable on the dependent variable
2. Description and discussion of the evaluation of the different modelling specifications for the dataset.
3. Presentation of your "best" specification (it is up to you to define "best" and indicate the statistical analysis that support your criteria for deciding) and a discussion of your selection criteria.

4. A discussion of the similarities and differences between the causal inferences from your "best" specification and your a priori considerations.

6. Some Comments

1. The following criteria will be applied for grading:

   a. Your understanding of the problem (e.g. causal relationships);

   b. Your understanding of the generalized linear model’s specification principles;

   c. Your understanding of regression statistics and hypothesis testing (explain the statistics you use).

2. Remember that you must always examine and comment on your results. Computer outputs without explanations are not valuable.

3. There are no formatting rules except that you should write a concise report with 10 pages (approx.) without annexes (where you should include tables you might find useful to complete your report).

Appendix

The variables available for your specification in dataset 2 are:

Variable	Description
NumAcc9702	Number of road accidents between 1997 and 2002 (dependent variable)
IFI	International friction index
TMDAveícdia	Average daily traffic
pHeav	% heavy vehicles
AS	Average speed
pUrb	% of the segment’s length belonging to an urban area. The higher the percentage is, the lower is the circulation speed of cars and the more they are exposed to potential conflicts with vehicles or pedestrians. A higher number of emergency breaks are expected also.
pRCross	% of the segment’s length under the influence of intersections. The higher the influence of intersections (measured in terms of length of segments in the vicinity to intersections), the higher potential conflicts can be expected as intersections are obvious locations for conflicts to occur.
pTExt	Length (m) of the segment with road curves. The higher the proportion of the segment with road curves, the less favorable are the conditions for driving.
CDR	Class of the most unfavorable curve radius in the segment (a segment can have several curves with different radius). Road curve radius were categorized in 4 classes from 0 to 4: Class 0 ⇔ R>1000m; Class 1 ⇔ 750m<R<1000m; Class 2 ⇔ 500m<R<750m; Class 3 ⇔ 500m<R<300m; Class 4 ⇔ R<300m. The tighter the curve class is, the worse are driving conditions.
CI	Longitudinal inclination class. A categorical variable was created indicating if the segment has a gradient (CI=1); if it is leveled (CI=0); or if there is a mix of gradient and level parts (CI=0,5). CI =0 is better than (CI=0,5) that is better than CI=1.
AP	Average annual precipitation (mm)
AcAADT	Accumulated Annual Average Daily Traffic (10^6 vehicles)