RESEARCH PAPERS OF THE WROCŁAW UNIVERSITY OF ECONOMICS
No. 25 2008
Knowledge Acquisition and Management

Stefanka Chukova*, Dimitar Christozov**

*Victoria University of Wellington, New Zealand
stefanka@vuw.ac.nz

**American University in Bulgaria
dgc@aubg.bg

MINING AUTOMOBILE WARRANTY DATA

Abstract: Identifying the factors affecting cost of covering claims according to warranty policy is the main purpose of this study. The paper presents mining of automobile warranty data. Variability Driving Pattern was discovered as a factor having significant and sustainable impact on warranty cost.

Key words: warranty, mean cumulative cost, variability.

1. Introduction

The purpose of this paper is to share findings in mining automobile warranty data. The project is aimed to estimate the Mean Cumulative Functions for warranty cost and factors affecting this cost. Data used were recorded claims for four modifications of a given model of an automobile, produced by General Motors (GM) and sold in the US. In the automobile industry, the warranty coverage is two-dimensional {age, mileage} represented by a rectangular area, i.e., the warranty expires either when the car reaches age of 36 months or its odometer accumulates more than 36 000 miles, whichever of these events occur first.

The initial hypothesis is that higher rate in mileage accumulation rate (MAR) leads to higher warranty cost. To evaluate the cost as a function of MAR a stratification model was introduced and applied to available data. This model was applied to one of the four available data sets and all analysis and conclusions were based on findings in mining this data set. The results do not prove the initial hypothesis, but allow us to highlight an unexpected factor – driving pattern variability (DPV), which affects the warranty cost. Further, the same analytical procedures were applied to the three other data sets and the discovered pattern was confirmed.

This paper shares the way of exploring data to discover factors affecting warranty cost.

2. Backgrounds and literature review

This problem of measuring cost for covering claims under warranty agreement was studied intensively in the last years [Blischke, Murthy 1996]. Here we may list few of these studies:

- Nelson [2003] assessed the cost as standard non-parametric recurrent event mean function estimation with standard errors;
- Kalbfleisch, Lawless and Robinson [1991] use Poisson models and incorporate reporting delay in estimators;
- Hu and Lawless [1996b] and Hu, Lawless and Suzuki [1998] assessed cost by missing censoring times with supplementary information and lifetime emphasis;
- Lawless, Hu and Cao [1995] use non-parametric recurrent events, address automotive warranty, with mileage accumulation, and censored warranty policy;
- Chukova and Robinson [2005] applied linear mileage accumulation to overcome the censoring problem.

The objectives of this study were to explore data to highlight the factors affecting this cost. The measure for warranty cost, we used, was the mean-cumulative cost according to Hu and Lawless [1996a] "Robust" Estimator, where time t is measured once as age and second time as mileage and applied to the number of eligible to claim cars. The warranty cost to a given t is measured by $\hat{\lambda}(t)$ and cumulative cost by $\hat{\Lambda}(t)$.

$$\hat{\lambda}(t) = \frac{n.(t)}{M\,\overline{G}(t)} \quad \hat{\Lambda}(t) = \sum_{s=1}^{t} \hat{\lambda}(s)$$

where

$$n.(t) = \sum_{i=1}^{M} \delta_i(t) n_i(t) \quad \delta_i(t) = I(a_i \geq t)$$

M – number of vehicles,
A_i – age of vehicle i,
$N_i(t)$ – number of claims at time t for vehicle I,
$\overline{G}(t)$ – Pr (under observation at time t).

Our initial hypothesis was that higher mileage accumulation rate (MAR) leads to higher warranty cost and to test it we used a stratified model, described in Section 3.

3. Description of data sets

Data includes four samples for four modification of one model of GM car, sold in USA during the years 1998, 1999, 2000 and 2001. Every data set holds data for about 50 000 vehicles and was provided as two tables, including the following, fields:

– **Sold vehicles:**
  - Vid (vehicle id) a unique number identifying the vehicle and
  - Saledate – the date the vehicle was sold.
– **Recorded claims:**
  - Vid;
  - Warrdate – the date of registering the claim;
  - Mileage;
  - Type (we have investigated a subsystem of type specified with "P");
  - Four categories of costs (labor, part, net and other) associated with the claim.

One claim with multiple labor operations can be recorded in several claim-records.

We used the 2001 sample and at the end transfer the developed processing tools to the other sets to compare the results.

The first table gives us the exact age of every vehicle to particular date. The second table holds data about mileage at the time of the claim. Or we know the mileage only for the vehicles with registered claims and only at the time of the claim. To evaluate the warranty cost as a function of age, we can spread the cost on all vehicles, which reached this particular age. To evaluate the warranty cost according to the mileage, we need to forecast the mileage for every vehicle, based on available partial data.

The model, described in the next section was used first to distinguish between vehicles with different MAR and second to approximate the missing mileage data.

4. Model: A stratification approach

The average mileage accumulation of a given vehicle is determined by assigning the vehicle to a particular stratum (category), based on its MAR. We consider 72 strata of equal sizes. Each claim was placed in a stratum according to its ratio mileage/age. We faced the following problems in mapping vehicles to strata:

- A significant number of the vehicles have recorded more than one claim and for different claims MAR may vary.
- In previous studies only the last claim was used as representative for the largest period, but in this way the information about MAR from the other claims was not used and variability of mileage accumulation was not studied.
- We decided to incorporate all available data by measuring the variability of MAR by assigning the vehicle to a particular class, called driving pattern class (DPC).

The set of cars with claims was partitioned in non-overlapping classes – driving-pattern classes (DPC) in the following categories:

– Cars with a single claim (S);
– Cars with multiple claims;
  • Cars with all claims within 1 stratum (M1),
  • Cars with all claims within 3 strata (M3),
  • Cars with all claims within 6 strata (M6),
  • Cars with claims spread over more than 6 strata (M>6).
  • The contribution of warranty cost of a vehicle to a stratum:
– For cars with a single claim (S): the cost was assigned to the stratum;
– For cars with multiple claims
  • For cars with all claims within 1 stratum (M1): the cost was assigned to the stratum;
  • For cars with all claims within 3 strata (M3): the cost was equally spread in the three strata (see Figure 1);
  • For cars with all claims within 6 strata (M6): the cost was equally spread in the sixth strata (see Figure 1);
  • For cars with claims spread over more than 6 strata (M>6): the cost was assigned to the stratum of the last claim.

5. Data processing

Data preprocessing

Data cleaning – removing data, violating consistency rules:

- Vehicles with missing ***saledate***;
- Claims with ***vid***, not included into vehicles data set;
- Claims with ***warrdate*** earlier than the ***saledate***;

Data structuring – aggregating all records of a single claim.

Discretisation:

- The age is discretised in months forming delta-age bins;
- The mileage is discretised in 1000 miles forming the delta-miles bins.

Data analysis – first round

Data segmentation:

- We assign to **each claim** two numbers: its stratum (as a ratio mileage/age) and its delta-age;
- We assign each **vehicle with claims** to one of the classes S, M1, M3, M6 and M>6;
- We assign to **each vehicle** an age, as the difference between the given date (October 24, 2003) and its ***saledate***;
- Therefore, we can evaluate:
  – the number of vehicles per strata for each class,
  – the number of vehicles for each delta-age (a month),

– the costs (for p-claims and all claims) per delta-age bin;

- These three sets are exported to an Excel spreadsheet.

Estimating the following distributions:

- Overall strata distribution as a combination of the strata-distributions over the five classes;
- Strata - age distribution is based on strata-distribution and the age distribution of all vehicles.
- Delta-miles distribution is constructed by the miles accumulated for every cell in strata - age table. This allows to assigned vehicles to each of the delta-miles bins.
- The functions $\lambda(.)$ and $\Lambda(.)$ are calculated for every age and miles bin.

Data analysis – second round

The cost associated with particular class and strata was calculated.

Data analysis – third round

The tools developed to analyze the 2001 data set were applied to data from 1998, 1999 and 2000 and results were compared.

6. Findings

We did not find enough evidence supporting the initial hypothesis. On Fig. 1 there is shown a graph representing the average cost associated with vehicles in four categories of MAR. For 2001 data set there are a slight trend towered increasing the cost with increasing MAR, but this was not seen in the other three data sets.

The more significant trend was observed when studied the cost as a function of DPC (see Fig. 2). This representation of the relationship between cost and **variability of driving patterns** (VDP) is a bit misleading, because for the category S we have only one claim and we cannot say anything about VDP for the given vehicle. But this graph forced us to investigate whether this relationship is meaningful. The results for the four data sets are presented in Figs. 3 to 6.

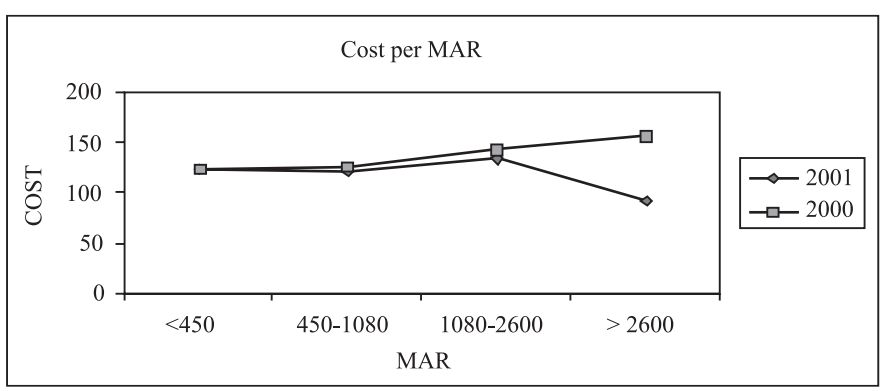

Fig. 1. Warranty cost as a function of MAR

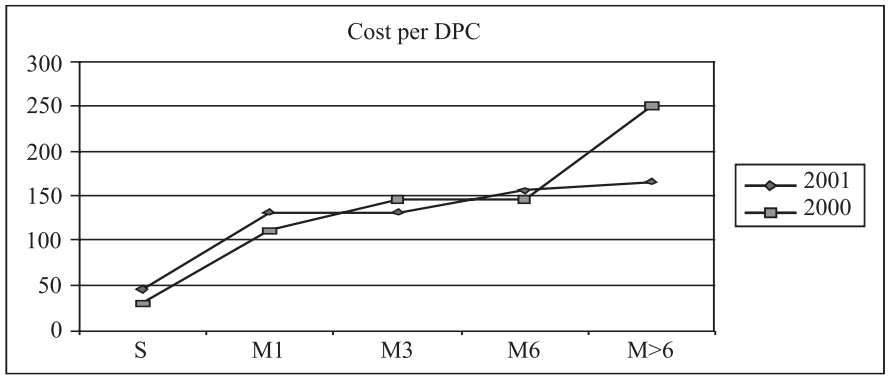

Fig. 2. Warranty cost as a function of DPC

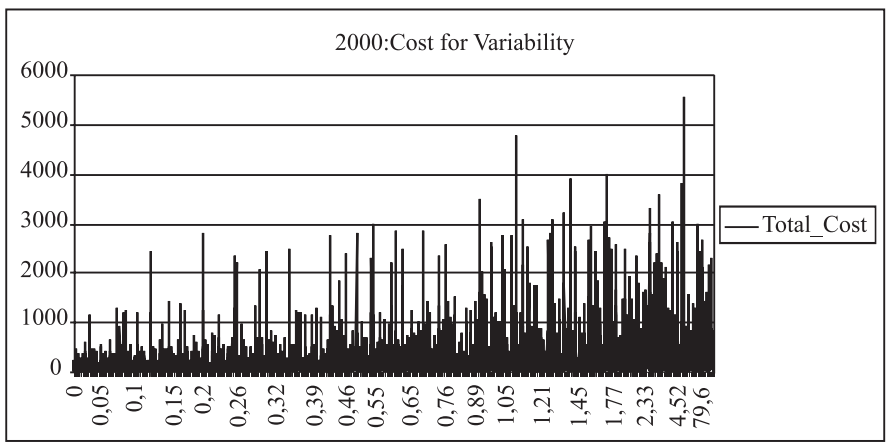

Fig. 3. Warranty cost as a function of VDP for 2001

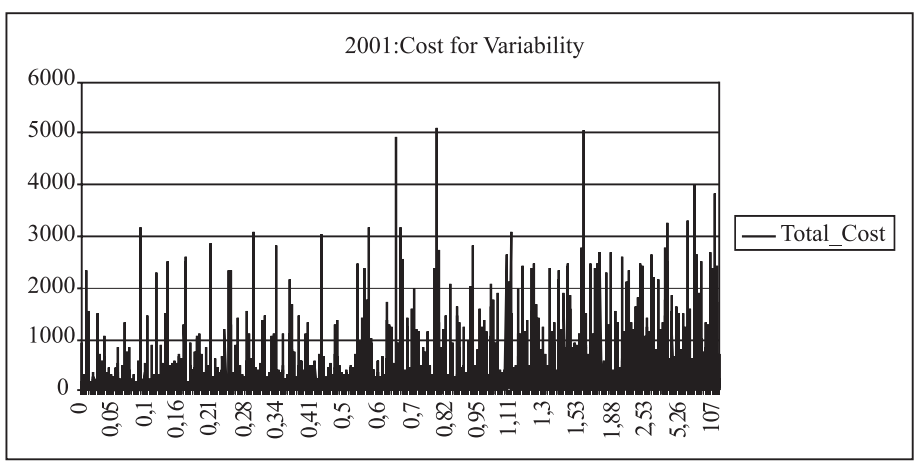

Fig. 4. Warranty cost as a function of VDP for 2000

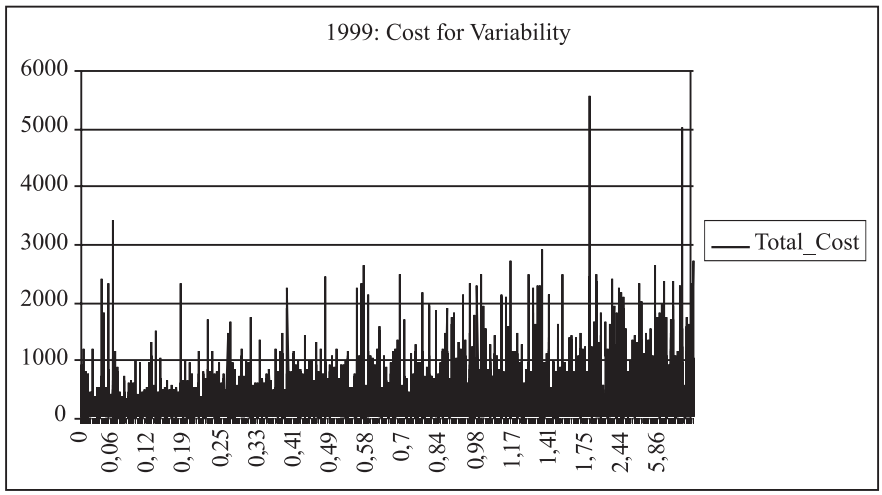

Fig. 5. Warranty cost as a function of VDP for 1999

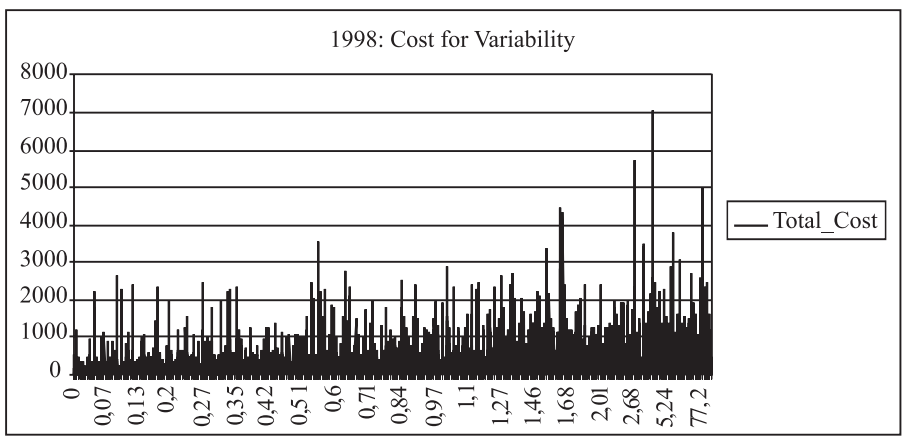

Fig. 6. Warranty cost as a function of VDP for 1998

7. Conclusion and further work

The presented results demonstrate the importance of incorporating data mining techniques in a regular practice, to increase understanding about the real factors affecting one or another aspects of the business. Availability of data is critical success factor in investigation the real world phenomena.

Comparing the presented results for the four data sets shows that dependence of the warranty cost on VDP decreases with improving the vehicle. We may assume

that a new modification of the model is an improved version of the previous one. To find evidences in support of this new hypothesis more data are needed as well as development of specific models and data processing procedures.

References

Blischke W., Murthy D. (1996). *Product Warranty Handbook*. Marcel Dekker, New York.

Chukova S., Robinson J. (2005). Estimating mean cumulative functions from truncated automotive warranty data. In: *Modern Statistical and Mathematical Methods in Reliability*. Eds. A. Wilson, N. Limnios, S. Keller-McNulty, Y. Armijo. World Scientific, Singapore.

Hu X., Lawless J. (1996a). Estimation from truncated lifetime data with supplementary information on covariates and censoring times. *Biometrika*, vol. 83, pp. 747-761.

Hu X., Lawless J. (1996b). Estimation of rate and mean functions from truncated recurrent event data. *Journal of American Statistical Association*, vol. 91.

Hu X., Lawless J., Suzuki K. (1998). *Nonparametric Estimation of a Lifetime Distribution When Censoring Times are Missing* (http://www.bisrg.uwaterloo.ca/archive/RR-96-04.pdf).

Kalbfleisch J., Lawless J., Robinson J. (1991). Methods for the analysis and prediction of warranty claims. *Technometrics*, vol. 33.

Lawless J., Hu X., Cao J.(1995). Methods for the estimation of failure. distributions and rates from automobile warranty data. *Lifetime Data Analysis*, 1.

Nelson W. (2003). Recurrent events data analysis for products repairs, *Disease Recurrences, and Other Applications*, ASA-SIAM, Philadelphia.