REVSTAT – Statistical Journal
Volume 14, Number 2, April 2016, 171–192

MODELING NON-LIFE INSURANCE PRICE FOR RISK WITHOUT HISTORICAL INFORMATION

Authors: Filipe Charters de Azevedo
– Department of Sciences and Technology, Universidade Aberta, Portugal
fchartersazevedo@hotmail.com

Teresa A. Oliveira
– Center of Statistics and Applications (CEAUL), University of Lisbon, and Department of Sciences and Technology, Universidade Aberta, Portugal
teresa.oliveira@uab.pt

Amilcar Oliveira
– Center of Statistics and Applications (CEAUL), University of Lisbon, and Department of Sciences and Technology, Universidade Aberta, Portugal
amilcar.oliveira@uab.pt

Received: October 2015 Revised: February 2016 Accepted: February 2016

Abstract:

- How should an insurer price a risk for which there is no history? This work intends to show, step by step, which main mechanisms are needed to capture the tariff model of another insurance company minimizing the risk involved. The document generally deals with the price-making mechanisms in non-life insurance through the GLM regression models — Generalized Linear Model, more precisely the Poisson, Gamma and Tweedie models. Given the complexity of the application of these models in experimental design, it is studied a simpler way to characterize the rate, namely considering the Box–Cox transformation with SUR — Seemingly Unrelated Regression. An orthogonal experimental design to collect information is also presented as well as an application of these methods in the motor industry considering different companies.

Key-Words:

- *pricing (non-life insurance); GLM; Box–Cox; optimal designs; SUR — Seemingly Unrelated Regression.*

AMS Subject Classification:

- 62J12, 62K05, 91B24, 91B30.

1. INTRODUCTION

An insurance company bases its production model in the value of a commodity with unknown cost by the time of production. Furthermore, the company "purchase" claims and "sell" safety, if a company buys the claims at a low price then it makes money; if it buys the claims at an expensive price then it loses money. In the value chain, a company can rely on the law of large numbers that mitigates volatility and market uncertainty — provides security on average.

The bottom line of a company is then how to evaluate the purchase price of claim: What is the cost of a risk (pure premium)? Usually an insurer has historical data that allow to estimate this value: based on the behavior of their customers it is reasonable to offer a premium, that is identical to the liabilities assumed (adding administrative costs, distribution and shareholder remuneration). But how should an insurer do to price a risk, for which there is no history? Should a company to "pay to view" — risking prices and their future sustainability? What should an insurer do if both — market and risk — are unknown?

From a practical point of view these questions are extremely important once the market has a strong barrier to overcome — the knowledge of the cost of raw materials. However there are solutions available in the literature. Some companies:

- Hire experienced technicians that heuristically define a charging table. Many investors are attracted to base their decisions on the information "currently available in their minds" see (Nocetti [5] and [6]). Thus, many times even when company has some historical data, experts opinions can be more plausible than the detailed analysis.
- Adopt reinsurance for (almost) 100% of the costs, transferring the risk for more experienced companies (which will draw a tariff) and that have financial muscle (to support higher risks).

In both cases (hiring experienced technicians or reinsurance) there is risk, and/or potential revenue loss. Are the companies locked to this reality? In any case, the insurer will always bear the costs of administration and distribution.

The challenge assumes more interesting contours since it is known that the *player* who first entered the market, or which has a higher market, has a strong competitive advantage: its historical references provide knowledge, which in this industry means the ability to determine more accurately the cost of the raw material. The *player* with no experience, only will get an interesting share if he gets a similar competitive advantage over the incumbent.

The aim of this work is thus to present a minimization method of pricing risk by capturing the tariff model, enabling a comparative advantage in the market to

smaller *players* (in terms of market share), with no relevant history and without financial padding to buy knowledge in a significantly way that is assuming risk. This capture method is based on the assumption that the smallest company can access a reasonable number of simulations with surgically chosen risk profiles. This collection can be performed, for example, by a mystery shopper or using their own mediators.

The methodological approach for answering to this challenge, follows the classic process of experimental design:

- Step 1: Identify the factors that define the product;
- Step 2: Identify levels that define the product;
- Step 3: Optimal Design;
- Step 4: Gathering information;
- Step 5: Analysis.

Considering the particular case of motor insurance an application will be performed in the sequence.

This work is organized in the following chapters:

- General Linear Models. In chapter 2 attention is given to changing pricing methodologies, particularly with regard to the GLM model associated to Tweedie distribution.
- Experimental design in context of a Tweedie population. The purpose of chapter 3 is to build a sample design which minimizes the field of endeavor, by using an Optimal Design and Box–Cox transformation. This is a practical solution once considering Tweedie populations, the variation component is not easily determined in an experimental design.
- Optimal Design. In chapter 4 the orthogonality concept is presented in order to gather information and allows discussing their suitability to the main objectives of the project: reducing the volume of information to be collected in order to obtain a manageable model and efficient estimates to facilitate the risk modeling. In this chapter special emphasis will be given to Seemingly Unrelated Regression — SUR — in order to maximize the predictability capacity.
- Applications. The methodologies explored in previous chapters are applied in chapter 6. We are working on a confidential real database, considering motor insurance data from a Portuguese insurance company in 2011.
- Conclusion and remarks. In the last chapter 7 emphasis should be given to the widespread conditions.

2. GLM — GENERALIZED LINEAR MODELS

How does a company know that it is expensive or cheap to pay for a policy? Going to market a company subscribes a policy and accept risks for which the real cost is unknown.

The well known expression of what usually is known as pure premium, which supports the rational of an insurance rate construction is:

$$\text{(2.1)} \qquad \textit{Pure Premium} \;=\; \textit{sinister frequency} \times \textit{average claim cost}\,(+\textit{error})\;.$$

Usually an insurer apply statistical models to estimate the frequency of an accident and their average claim cost. This problem can be seen isolated (frequency and average claim cost) or estimated jointly (Pure Premium). The concept of regression tackles this problem successfully, whatever its formulation. It should be noted that in a generalized regression model there are two components:

i) A random vector $Y = (Y_1, ..., Y_n)'$ is following a distribution with unknown parameters vector $\mu = (\mu_1, ..., \mu_n)'$;

ii) A function relation between μ and the involved parameters' vector $\beta = (\beta_1, ..., \beta_k)'$, such that $\mu = f(\beta)$, considering $f(\cdot)$ a continuous and univocal function.

Following the terminology of Jørgensen ([15]), these two components are referred respectively as the random component and the systematic component of the model. The random vector Y is designated as the response, while the random component can assume any stochastic process, including errors measurement. The function $f(\cdot)$ is designated as the regression function and the β parameters represent the regression parameters. This whole system of vectors and distributions is defined though the average for each Y_i on the conditions of μ, $E(Y_i\,|\,\mu)$. The variation associated with $E(Y_i\,|\,\mu)$ provide a measure of the adjustment quality.

An important class of regression models can be expressed as:

$$\begin{aligned} g(\mu_i) \;&=\; \eta_i \;, && i = 1, ..., n \;; \\ \eta_i \;&=\; \sum_{j=1}^{k} x_{ij}\,\beta_j \;, && i = 1, ..., n \;. \end{aligned}$$

The function $g(\cdot)$ is continuous and unequivocal and is designated as *link function*. The matrix $X = \{x_{ij}\}$ is the design matrix model and x_{ij} are the covariates or explanatory variables. A model of this form is said to be linear.

When $g(\cdot)$ is an identity function and Y distribution is homoscedastic or even normal, the simple linear regression model is considered. Usually in this

simple case parameters are estimated by the least squares error minimization or by the maximum likelihood.

In some cases it is possible to question the type of function $g(\cdot)$ assumes as well as the distribution associated with Y, which is a very convenient way to determine the heterogeneity of the data. It is typical to assume that Y distribution is defined by a Poisson probability function and that Gamma distribution is used to compute the average cost. When a claim frequency is the goal, to establish the terms in (2.1) together, the most traditional mechanism is through GLM composite models. In this case, a very convenient way to determine the heterogeneity of the data is assuming that the Y distribution is defined by an exponential distribution model (ED):

$$p(y,\theta,\lambda) \;=\; \alpha(\lambda,y)\, e^{\left[\lambda\left\{y\theta-k(\theta)\right\}\right]}\,, \qquad \text{with} \;\; y\in\mathbb{R}\,.$$

Note that $\alpha(\cdot)$ and $k(\cdot)$ represent functions, and $\lambda>0$ and θ belongs to a real domain.

Thus, let $Y\sim ED(\mu,\sigma^2)$ where $\mu=k'(\theta)$ represent the expected value of Y and $\sigma^2=\frac{1}{\lambda}$ represent the variance where ED refers to the family of exponential distributions and in Jørgensen ([15]) there is a particular case of this family distributions, characterized by $V(Y)=\sigma^2=\phi\,V(\mu)$.

The particular cases of $V(Y)=\sigma^2=\phi\,\mu^p$, to diverse p assume an important class usually associated to the Tweedie distribution model. This class can be:

- a normal data generator when $p=0$;
- a Poisson data generator when $p=1$;
- a Gamma data generator when $p=2$;
- an Inverse Gaussian when $p=3$.

Considering $1<p<2$ the Tweedie exponential distribution assumes the expression:

$$p(y,\theta,\lambda) \;=\; \sum_{n=1}^{\infty} \frac{\left\{(\lambda\omega)^{1-\alpha}\,k_\alpha\!\left(-\frac{1}{y}\right)\right\}^n}{\Gamma(-n\,\alpha)\,n!\,y}\; e^{\left[\lambda\left\{y\theta-k_\alpha(\theta)\right\}\right]}\,, \qquad \text{to} \;\; y>0\,.$$

Let $P(Y{=}0)=e^{\{\lambda\omega k_\alpha(\theta_0)\}}$ where $k_\alpha(\theta)=\frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^\alpha$, $\theta_0=\theta\,\lambda^{\frac{1}{(1-\alpha)}}$, and ω represents the weight associated to the observation exposition. As we can observe, for the Tweedie distribution the density function depends on the parameter p which relates to the variance $V(Y)=\sigma^2=\phi\,\mu^p$. This parameter p is thus defined exogenously before the estimation process, usually due to the analyst experience. These GLM models are widely recognized in the industry. Anderson *et al.* ([1]) presents it as the standard method to define motor and other lines of commercial branches tariffs. It is also indicate that these models are used by companies

in the UK, Ireland, France, Italy, the Netherlands, Scandinavia, Spain, Portugal, Belgium, Switzerland, South Africa, Israel and Australia. The referred paper also states that this model has gained popularity in Canada, Japan, Korea, Brazil, Singapore, Malaysia and Eastern European countries. More details on this distribution applied to the actuarial context can be obtained in Jørgensen and de Souza ([17]).

The type of function $g(\cdot)$ follows some rationals, as the context of analysis and distribution of Y.

3. EXPERIMENTAL DESIGN CONSIDERING TWEEDIE POPULATIONS

It has been noted that insurance companies are usually using a way of charging based on GLM that combine Poisson|Gamma model or on a composite model (known as Tweedie model). For the sample design in Tweedie regression, see Jørgensen and de Souza ([17]), it should be noted the following approaches, which are well known in the literature:

(**i**) **Sequential design**: The sequential design for binary responses has a rich history, which dates back to the 1940s, see Wald ([23]), on trying to find designs which results lead to asymptotic properties, see also Haines *et al.* ([12]), Ivanova and Wang ([14]) and Karvanen *et al.* ([18]). These authors concentrate their work on one factor designs and the challenge in our work is to extend this research to multifactorial designs. In Woods *et al.* ([24]) and Dror and Steinberg ([8]), solutions to this multifactoriality problematic are presented. However, such solutions are computationally complex, and the associated methodology is based on "ebb and flow" and trial and error, making the process complex and nonintuitive.

(**ii**) **Design based on clusters**: The Tweedie regression is based on the estimation of three parameters vectors: φ, θ and p, where p conditions affects the other two parameters computation. The design by clusters seek to find homogeneous groups of observations in order to determine p in an exogenously way. This idea is conceptually interesting, and is computationally easy to perform.
In Dror and Steinberg ([7]) is suggested an approach based on K-means cluster — since this process allows rapid exploration of various designs outperform the existing alternatives. The authors mention "given the set of location D-optimal designs, the core of the proposed method is to combine them into the set of vectors location and use K-means clustering to derive a robust design".

The possibility of finding an optimum location with this method has, however, a serious problem with respect to the other model coefficients: how to evaluate the estimated degree of accuracy? The question arises once the clusters were defined exogenously and a sample experimental design is always reduced to allow "good" experiences performance. As an avenue for improvement one can explore the use of computational simulations to ensure the best model based on different levels of p. Algorithms based on *random forest* may be an important issue to consider. However this is not the main goal of this project.

3.1. Experimental design 1: A pragmatic solution in Tweedie populations

As the GLM models Tweedie are not easily applicable in the experimental context it is necessary to find a pragmatic solution. The main problem is to have an experimental model analysis under heterocedasticity conditions or dispersion models, where Tweedie distribution fits well.

Trying to stabilize the variance, see Box and Cox ([4]), the usual method is to determine empirically or theoretically the ratio between the variance and the mean. The empirical relationship can be found by the logarithm and the average graph, or to make a transformation in the dependent variable.

For positive expected value, the well-known Box–Cox Transformation is frequently used:

$$y^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda} & \text{to } \lambda \neq 0\,, \\ \log(x) & \text{to } \lambda = 0\,. \end{cases} \tag{3.1}$$

The choice of λ however, in our days, is usually done automatically, while Osborne ([20]) and Harrison and McCabe ([13]) propose the following algorithm:

1. Splitting the key variable in 10 (or more) intervals;
2. Calculate the mean and standard deviation for each interval;
3. Design a graphic with $\log(\sigma)$ vs. $\log(\mu)$ for each of the regions;
4. Estimate the average slope of the graphic, and use the $1-\lambda$ as initial value of λ.

It is important to refer that this algorithm is not an unanimous choice for researchers and usually, as in Ripley *et al.* ([22]) it is assumed that the best way to estimate λ is the one that guarantees the maximum likelihood. Drawing the evolution of the maximum likelihood function can be useful in this case.

However this method is not prudent, since any $|\lambda|$ too high will reduce the variability of the variable goal. Therefore when you re-build the target variable of $y^{(\lambda)}$ to y the result may be an estimated variable without any variability. Some software and statistical packages (MASS in R, STATA) maintain this approach, but impose limits to $y^{(\lambda)}$, usually in $y^{(\lambda)} < 1$, $y^{(\lambda)} < 2$.

Finally, another alternative is to look for a λ value that makes sense to the analyst. A careful reading of Box and Cox ([4]) points in that direction. So it is noteworthy that the way loglin $\Longleftrightarrow \lambda = 0$ is theoretically the one that makes more sense to use for the premium model, and is easier to interpret:

1. The distribution of total / pure premium costs (i.e. Tweedie with reliable parameters) is visually close to a log normal or gamma;
2. The log-lin model has the advantage that the coefficients represent elasticities; a very meaningful concept in terms of premiums.

A very simple way to find and test the data transformation — at least between the linear form and log — is presented in Mackinnon and Davidson ([19]). Although the estimation process may seem a little complex, the test logic is very simple: if the linear model is in fact correct, the formula $e^{(\widehat{\log(y_{\text{based on log model}})})}$ will be related to the model under evaluation (so, it will be enough to use the regression and a t-test).

In short, to determine the tariff model competition in experimental design context, the Box–Cox methodology is preferable to the Tweedie regression. In addition, after the analysis of the best functional form, subsequently a Tweedie regression may be applied but using Box–Cox, for a rough indication of what value p may assume (i.e. the shape of the Tweedie), so that to overcome the already mentioned difficulties. That is why we propose this strategy to overcome the existing computational difficulties.

3.2. Experimental design 2: Box–Cox regression correction

The key variable to estimate in this case is $y = \textit{commercial premium}$ and not $y^{(\lambda)}$. When the estimation process is integrated there is the need to decompose the Box–Cox formulation in the correct formulation:

$$\widehat{y} = \begin{cases} \widehat{y^{(\lambda)}}\,\lambda + 1 & \text{to } \lambda \neq 0\,, \\ e^{(\widehat{y^{(\lambda)}})} & \text{to } \lambda = 0\,. \end{cases} \tag{3.2}$$

However, this formulation is not the most statistically efficient. In fact the application of Box–Cox expression underestimates the y expected value.

The mass point of the linear Box–Cox regression (the average of x and y on average) is not identical in both equations, in order to $\widehat{y^{(\lambda)}}$ and in order to y. In Wooldridge ([25]) is presented a solution for the case $\lambda = 0$, which can be generalized to any variant of Box–Cox regression. It is possible to obtain a corrected model by regressing y to $\widehat{y}$ without the constant component. The coefficient associated to $\widehat{y}$ gives the correction factor of the mass points. So, taking into account the correction, the final prediction for two stages estimation is:

$$\widehat{\widehat{y}} = \widehat{y} \times \textit{correction coefficient} \, . \tag{3.3}$$

4. EXPERIMENTAL DESIGN: IDENTIFICATION OF FACTORS AND LEVELS TO COLLECT

Regarding the determination of the factors, the experimental work is easy: Each customer must fill out a quotation document so that a quote can be issued. Usually, there are no data to work beyond the required (although it is known that some insurers in bancassurance partnerships use the bank behavior data, and in other countries it is known that the profile on social networks can be used). The work on validation factors in brainstorming sessions and interviews with different experts, see Barker ([2]), is dispensed as companies in the quotation indicate which factors are supposed to be investigated.

In the case of motor insurance, the factors usually considered are:

1. **Characteristics of the insured**
   - Gender: The rational of this variable is associated mainly to a different frequency of accidents according to gender. It should be noted, however, that in March 1, 2011 the European Court of Justice ruled that insurance companies which use gender as a risk factor were to disregard EU equality laws. However, in Portugal, in February 2015 (Law No. 9/2015), it became amicable to have same gender discrimination if premiums and benefits "are proportionate and justified by a risk assessment based on actuarial and statistically relevant and accurate data". The possibility of using this variable from 2015 thus became a reality.
   - Age: The rational of this variable is to measure the inexperience and risk trend of the insured. It is a variable impacting the accident frequency and, depending on the coverage, the average cost.
   - Claims History: The claim record can be consulted by the Portuguese insurers in SegurNet (a managed platform for the sector's association with the claim record).

- Age when the driver got driver license: When combined with age, it attain an instrumental variable of the driver's experience.
- Status: Rarely used in Portugal, although there is some sensitivity to point to the fact that married drivers have fewer accidents than the rest of the population.
- Usual path: Variable indicating the accident frequency — the greater the distance house-work, the greater the likelihood of an accident.
- Payment: The payment method reveals the financial pressure that the driver is subject; is a factor that correlates with the driving profile — frequency. In addition, the payment method is correlated with the insurer capital consumption and therefore to effect on the commercial premium via the administrative burdens and profit loading.

2. **Features of insurance risks**
   - Vehicle rating: The power to weight ratio increases the sinister frequency.
   - Brand and classification of vehicle: There are brands whose parts cost more than others, so this variable has an impact on the average cost. The type of construction, security features also has its influence on the average cost.
   - Using the insured object: If the object is essential for day-to-day, or for professional use, accident frequency increases while the frequency per unit of exposure (measured in km driving) decreases. Thus, having a profession and any instrumental variable of the insured object use is relevant.

3. **Regional and general contexts**
   - Weather: The loss context has been the least considered issue in the construction of a tariff. For example, if it rains, there are more accidents, but companies have rates for a given country in which implicitly rainfall rates do not vary. If there is a crisis, people use less the car, so there are fewer accidents. These context variables are linked to the evolution of times and this issue should be carefully analyzed.
   - Region: Regional variables are often neglected since everything is placed in large commercial areas and not with sufficient granularity.
   - Sector: For professional cases in certain sectors, for example transport and distribution, there is a greater exposure.

4. **Company**
   - To the mentioned factors another one should de added: the company. Presumably this factor has strong impact on the relationship among all the others: each company defines their particular pricing model.

This information should be used to determine the true market risk, since in theory all companies are measuring the same risk: frequency and average cost. Otherwise, since there is no exchange of tariff models between insurance companies, there will be as many models as the number of companies: the model is statistically different from company to company and that it is not controlled with a simple randomization. The inclusion of this information in the estimation process will be detailed in the next section. A mathematized way, and considering that x has the usual reading of exogenous variables, the model assumes the form:

$$y_i = f_{company_j}(x_i) \tag{4.1}$$

and not

$$y_i = f(company_i, x_i)\ . \tag{4.2}$$

Regarding the levels the question is different since the collection is performed continuously on some key variables. More, there may be some ratios and values derived (the calculation of the power weight is perhaps the most obvious case). It should be noted, however, that the choice of levels must be such as to minimize the information or the variance, leading to an efficient and feasible project. At this point it must be assumed that using a panel of experts, see Barker ([2]), it is possible to minimize/aggregate the number of levels, where it is emphasized that is best to arrange an experiment as a team effort and use the brainstorming technique to scope the entire problem.

5. OPTIMAL DESIGNS

The Completely Randomized Design (CRD) is the simplest form of statistical experimental design. In a CRD the treatments are randomly assigned and the model is linear. It is necessary to check a set of hypotheses, often called classical hypotheses and to estimate the generating process of tariffs by maximum likelihood, in order to obtain a centered model. With classic conditions (linearity in the parameters, random sample, absence of perfect multicollinearity) it is possible to ensure the centering of the maximum likelihood estimators. Thus, in a random sample, as is customary in regression work context, there is a random mechanism which selects the sample within all possible samples.

In experimental design context, adopting the usual notation, as in Graybill ([9]), it is possible to interpret the problem differently. If there is a data generating process (and not a random selection mechanism of samples), the data

structure remains the same, the estimator formula will be the same and centering is guaranteed (strict conditions).

It is thus possible to collect any data to ensure the $\widehat{\boldsymbol{\beta}}$ estimator centering. But centering is not the only prerequisite. Under this assumption of data generating mechanism it is also possible to go further in terms of variance test.

The second condition for a good sample design is to maximize the power of the tests. To ensure that the process is efficient, i.e. that the standard deviation associated with each of the estimates of the betas is minimal, there is the need to examine the beta estimator formulation (to see its derivation see, e.g., Gujarati ([11])):

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}\,. \tag{5.1}$$

The $\widehat{\boldsymbol{\beta}}$ variance is given by:

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\,. \tag{5.2}$$

For an efficient estimator there is the need to have enough number of cases in order to estimate regression and that each of the elements in the diagonal of matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is minimum (assuming no interaction effects). The best way is to guarantee that the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is only filled in the diagonal — there is no correlation between the different x, (cf. Cramer Rao, see [3]). In such case the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is orthogonal. So, we have:

$$\widehat{\boldsymbol{\beta}} = \mathbf{I}\mathbf{X}'\mathbf{y}\,. \tag{5.3}$$

In such case the estimator values are easily obtained and will have minimum variance. Note that if $(\mathbf{X}'\mathbf{X})^{-1} \neq \mathbf{I}$ there will be *confounding* and it will not be possible to estimate without an high error associated to the estimated coefficients.

5.1. Optimal Design — functional adjustments

When the sample design and the data analysis are performed, it is possible to obtain a simple linear regression model to determine the importance of factors (through a t-test) and the degree of criticality of their levels (again with t-test, assuming levels as dummy variables), possibly setting the best functional form.

But it is worth exploring the meaning of (4.2) and the need to have a function for company seen in the previous section. Relation (4.2) indicates that the option is to collect and model data for a single company, assuming that each company should have autonomous pricing models. If one considers only two

companies, the model can be described as:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}, \qquad \text{or moreover} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}\,, \tag{5.4}$$

where

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \mathbf{e} \sim N\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} \sigma_1 I_T & 0 \\ 0 & \sigma_2 I_T \end{bmatrix} = \mathbf{W} \right], \qquad \mathbf{e} \sim N(\mathbf{0}, \mathbf{W})\,.$$

Another issue arises: Companies are operating in the same market, is it all the information available to capture the tariff model on the market being used? Or even more directly: "The tariff models are autonomous, but are they independent?" In a more mathematized form (applying the same rational Griffiths *et al.* ([10])): And if the mistakes of the different equations, e_1 and e_2, are correlated?

Thus, consider

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \mathbf{e} \sim N\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} \sigma_{11} I_T & \sigma_{12} I_T \\ \sigma_{21} I_T & \sigma_{22} I_T \end{bmatrix} \right]. \tag{5.5}$$

The idea is that we can estimate the (4.2) per blocks in order to take different functional forms, and perhaps different explanatory variables; but considering (5.5) we can have greater accuracy in forecasting and more power in the tests. The demonstration of these statements follows below, considering just the case of two companies, although the generalization is directly (and it can be confirmed in the Annex to Sec. 17, see Griffiths *et al.* ([10]).

Considering maximum likelihood, it follows that:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left[\begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} \sigma_{11} I_T & \sigma_{12} I_T \\ \sigma_{21} I_T & \sigma_{22} I_T \end{bmatrix}^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \right]^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \end{aligned}$$

$E(\mathbf{e}\mathbf{e}') \neq \boldsymbol{\sigma}\,\mathbf{I}$, the usual case does not apply, so: $E(\mathbf{e}\mathbf{e}') = \mathbf{W}$ and $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\,\mathbf{W}^{-1}\mathbf{X})^{-1}$.

However, this estimator is not likely to be calculated, since the matrix W is not known; then it must be estimated. In other words, it is necessary to establish the following relationship: $\widehat{\sigma} = \widehat{\mathbf{e}}\,\widehat{\mathbf{e}}'$. Thus:

$$\begin{aligned} \widehat{\widehat{\boldsymbol{\beta}}} &= (\mathbf{X}'\,\widehat{\mathbf{W}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left[\begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} \widehat{\sigma_{11}} I_T & \widehat{\sigma_{12}} I_T \\ \widehat{\sigma_{21}} I_T & \widehat{\sigma_{22}} I_T \end{bmatrix}^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \right]^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \end{aligned} \tag{5.6}$$

$$= \left[\begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} \widehat{e_1'}\widehat{e_1} I_T & \widehat{e_2'}\widehat{e_1} I_T \\ \widehat{e_1'}\widehat{e_2} I_T & \widehat{e_2'}\widehat{e_2} I_T \end{bmatrix}^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \right]^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

with $\mathrm{Cov}(\widehat{\widehat{\beta}}) = (\mathbf{X}' \widehat{\mathbf{W}}^{-1} \mathbf{X})^{-1}$.

Since without other companies the model is not homoscedastic (i.e., $E(\mathbf{e}\,\mathbf{e}') \neq \boldsymbol{\sigma}\,\mathbf{I}$), the different $\widehat{\boldsymbol{\beta}}$ estimators are not centered with minimum efficiency. However $\widehat{\widehat{\boldsymbol{\beta}}}$ is homoscedastic in case of good parametrization. The analysis of t and F tests will be more precise and the experimental design gains more consistency. And in this way it can be captured which are the mechanisms that generate pure data/premiums from different companies.

Evidently the SUR adjustment should be applied before the deconstruction of objective variable and accordingly the correction of the mass point determined already indicated.

6. APPLICATIONS

To better understand the application of this concept, a database that portrays the conditions of the insurance market in 2011 for the Portuguese aggregate car liability coverage and travel assistance was studied. On the computation procedures R software ([21]) was used. The data presented were slightly retouched in order to guarantee the anonymity of the companies under review. The variable "form of payment" was excluded from the analysis in order to create one more element of non-identification of the insurers, impairing pure orthogonality.

The following subsections accompany the classic stages of experimental design.

Stage 1: Identify the factors and levels that define the product

The variables considered were not defined within this article nor by its authors. The discussion of these variables is indicated in Table 1. Each broker must deliver a quotation under a specific scenario for the minimum legal capital requirement plus a minimum coverage for travel assistance.

Stage 2: Optimal Design

The sample was drawn by mystery shopping imposing a minimum number of observations: A standard case was set up for the main factors, and variable levels were changed in five subsamples. Therefore, orthogonality was not guaranteed. The estimated model is well centered, but is not necessarily statistically efficient.

Table 1: Description of factors and levels.

Variables	**Number of Levels**	**Identification of levels**
Gender	2	Male Female
Age	7	19 23 28 35 45 57 67
Claims history	21	0 injury/10 years/15 years 0 injury/15 years/15 years 0 injury/ 2 years/ 2 years 0 injury/ 4 years/ 4 years 0 injury/ 5 years/10 years 0 injury/ 5 years/12 years 0 injury/ 5 years/ 5 years 0 injury/ 5 years/ 6 years 0 injury/ 5 years/ 7 years 0 injury/ 5 years/ 8 years 0 injury/ 7 years/ 9 years 1 injury/ 0 years/ 1 years 1 injury/ 0 years/ 2 years 1 injury/ 0 years/ 3 years 1 injury/ 0 years/ 4 years 1 injury/ 0 years/ 5 years 1 injury/ 1 years/ 5 years 1 injury/ 2 years/ 9 years 1 injury/ 3 years/ 9 years 1 injury/ 4 years/10 years no experience
motor classification	6	Picup truck Light vehicle Commercial vehicle Multi-purpose vehicle Pickup Off-road vehicle
Automobile age	13	{0,1,...,10}; 15 and 20
Region	58	58 Municipalities
Companies	7	Confidential

Stage 3: Collection of information

This database was collected at the beginning of the decade for a specific consulting project. A detailed technical specifications could allow customer identification, project objectives and market conditions. Thus, all data were carefully calibrated so as not to allow companies to determine the target or operating results.

It should be noted that for the realization of information collection resorted in non-exclusive agents. These brokers collected quotations, according to a scenario/profile structured to be simulated beforehand and subsequent recording of information on observation grids.

Care was taken to ensure that the brokers were gathering in municipalities belonging to the working regions.

Stage 4: Analysis

First of all, it should be noted that the model to estimate will run with the following steps:

1. Building a model for insurance and calculation of λ;

2. Estimation of the model by SUR;

3. Mass point correction;

4. Obtaining the best β estimator;

5. Critical analysis of the results and, optionally, repetition of the cycle.

S4.1. Building a model for insurance and calculation of λ

As discussed, the model should include all the variables collected for each company.

For choosing λ it was decided initially to obtain a graphic with the evolution of the likelihood function between -10 and 1, and the limits around zero and in order to include the maximum of each of the functions. The results can be observed in Figure 1 and indicate that the λ which maximize the objective functions are: -3.504, -1.723, -4.055, -3.469, -1.380, -3.276 and -3.024.

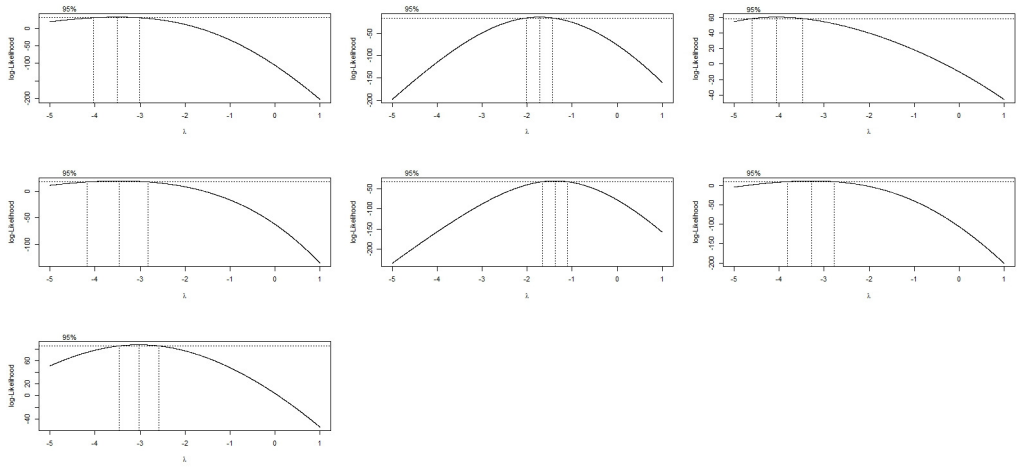

Figure 1: Evolution of the likelihood functions for each of the companies and obtaining λ.

All optimal points move away from intervals with easy interpretation. It should also be noted that apparently the companies use different tariff models reinforcing the idea of estimating each model separately, according to equation (4.2).

When using these lambda values, and estimates the data generating process by ordinary regression and reconstructs the variable ultimate goal, the result is bleak. The choice as accurate lambda ultimately eliminate all variability in the model. It's worth mention in Figure 2 the results with $\lambda < \sim -3.2$ the result is a parallel to the x-axis. The model is therefore estimated using $\lambda = 0$.

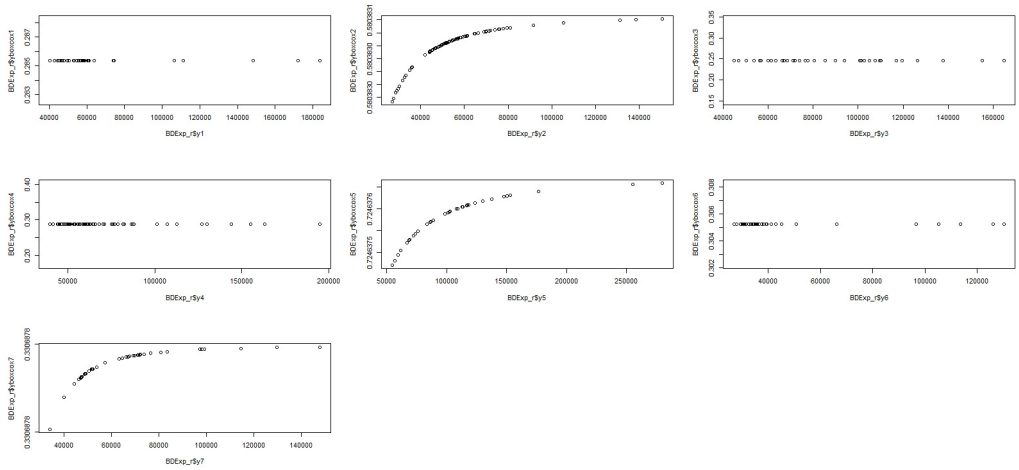

Figure 2: Function of $\widehat{y}$ with Box–Cox that maximizes the likelihood function.

S4.2. Estimation with model SUR

The model estimated by SUR, even with the transformation Box–Cox, shows a strong correlation matrix (qualitative assessment) between models of different companies. Indeed, the correlation between the estimated rates varies between 49% and 93%.

Table 2: Estimated models per enterprise — the error correlation matrix.

	eq1	eq2	eq3	eq4	eq5	eq6	eq7
eq1	1.00	0.28	0.40	0.53	0.64	0.51	0.34
eq2	0.28	1.00	0.60	0.52	0.40	0.32	0.56
eq3	0.40	0.60	1.00	0.56	0.48	0.54	0.38
eq4	0.53	0.52	0.56	1.00	0.52	0.29	0.58
eq5	0.64	0.40	0.48	0.52	1.00	0.49	0.24
eq6	0.51	0.32	0.54	0.29	0.40	1.00	0.01
eq7	0.34	0.56	0.38	0.58	0.24	0.01	1.00

S4.3. Mass point correction

Correction of the mass point has different impacts: in some cases is almost negligible, in others may require a correction of $> 4\%$. In fact, per company it is possible to find the following correction factors: 1.0427, 1.0202, 1.0053, 1.0177, 1.0263, 1.0398 and 1.0079 .

S4.4. Critical analysis of the results and, optionally, repetition of the cycle

In this work only the $R2$ between the final estimated variable and the variable goal is analyzed — indeed to the general objective of this work the main interest is in evaluating the predictive power of the models. So we have the following coefficients $R2$: 0.88021, 0.81354, 0.91175, 0.87484, 0.89378, 0.90491, and 0.8674 — very high values indicating excellent adjustment capacity.

The remaining quality indicators usually calculated on a regression analysis may also be applied. In any case it is interesting to compare the estimated model with the observed pattern. As can be observed in Figure 3, the largest deviation holds mainly thanks to the existence of outliers that were not treated/corrected. Therefore the determination of the final model will be made using the matrix W.

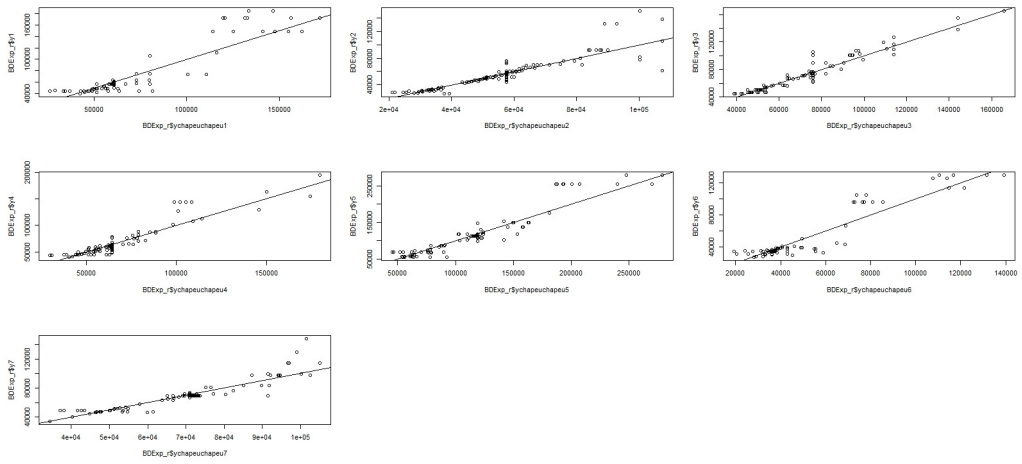

Figure 3: Comparison of SUR model with original data.

7. CONCLUSIONS AND FUTURE RESEARCH

The main purpose of this work was to present a method for collecting and capturing the tariff model for an insurance company and this goal was achieved. An approach to sample design based on the principles of orthogonality was presented as well as the linear regression model, with Box–Cox transformation and point correction for a first analysis. We presented a methodology to integrate information from more than one company and therefore increasing the efficiency of the estimators through a SUR model.

For a future work is the possibility of designing a more complex experimental design model with GLM — Tweedie. This would potential provide a greater adherence to data , specially if one can indicate how to get a rough estimate for the dispersion factor p. It will be interesting to investigate how the Box–Cox model may contribute for an efficient estimation of Tweedie on the determination of p. It will be also interesting to assess the prevalence of the SUR approach in the case of GLM.

ACKNOWLEDGMENTS

The authors acknowledge the valuable suggestions from the referees.

Teresa A. Oliveira and Amílcar Oliveira were partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal — FCT under the project UID/MAT/00006/2013 and by the sabbatical scholarships SFRH/BSAB/113669/2015 and SFRH/BSAB/113590/2015, respectively.

REFERENCES

[1] Anderson, D.; Feldblum, S.; Modlin, C.; Schirmacher, D.; Schirmacher, E. and Thandi, N. (2007). *A Practitioner's Guide to Generalized Linear Models — a foundation for theory, interpretation and application*, Third edition, CAS Discussion Paper Program.

[2] Barker, T.B. (1994). *Quality By Experimental Design*, Second edition, Chapman and Hall/CRC.

[3] Bhapkar, V.P. (2014). *Cramer-Rao Inequality*, Wiley StatsRef: Statistics Reference Online, John Wiley & Sons, Ltd.

[4] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with Discussion), *J. R. Statist. Soc. B*, **26**, 211–256.

[5] Nocetti, D. (2005). A model of mental effort and endogenous estimation, *Econ. Bull*, **4**(14), 1–10.

[6] Nocetti, D. (2006). Portfolio Selection with endogenous estimation risk, *Econ. Bull*, **7**(6), 1–9.

[7] Dror, H. and Steinberg, D.M. (2006). Robust Experimental Design for Multivariate Generalized Linear Models, *Technometrics*, **48**(4), 520–529.

[8] Dror, H. and Steinberg, D.M. (2008). Sequential Experimental Designs for Generalized Linear Models, *Journal of the American Statistical Association*, **103**, 288–298.

[9] Graybill, F.A. (1976). Theory and Application of the Linear Model, First edition, *Duxbury Classic Series*.

[10] Griffiths, W.; Hill, R.C. and Judge, G. (1993). *Learning and Practicing Econometrics*, John Wiley and Sons.

[11] Gujarati, D.N. (1995). *Basic Econometrics*, Third edition, McGraw Hill.

[12] Haines, L.M.; Perevozskaya, I. and Rosenberger, W.F. (2003). Bayesian-Optimal Designs for Phase I Clinical Trials, *Biometrics*, **59**, 591–600.

[13] Harrison, M.J. and McCabe, B.P.M. (1979). A Test for Heteroscedasticity based on Ordinary Least Squares Residuals, *Journal of the American Statistical Association*, **74**, Issue 366a, 494–499.

[14] Ivanova, A. and Wang, K. (2004). A Non-Parametric Approach to the Design and Analysis of Two-Dimensional Dose-Finding Trials, *Statistics in Medicine*, **23**, 1861–1870.

[15] Jørgensen, B. (1989). *The Theory of Exponential Dispersion Models and Analysis of Deviance*. In "Lecture notes for short course, School of Linear Models", University of São Paulo, 129 pages. Second Edition, 1992: *Monografias de Matemática*, **51**, IMPA, Rio de Janeiro.

[16] Jørgensen, B. (1987). Exponential dispersion models, *Journal of the Royal Statistical Society*, Series B, **49**(2), 127–162.

[17] Jørgensen, B. and de Souza, M.P. (1994). Fitting Tweedie's compound Poisson model to insurance claims data, *Scand. Actuarial J.*, **1994**, 69–93.

[18] Karvanen, J.; Vartiainen, J.J.; Timofeev, A. and Pekola, J. (2007). Experimental Designs for Binary Data in Switching Measurements on Superconducting Josephson Junctions, *Applied Statistics*, **56**, 167–181.

[19] Mackinnon, H.W. and Davidson, R. (1983). Tests for Model Specification in the Presence of Alternative Hypothesis; Some Further Results, *Journal of Econometrics*, **21**, 53–70.

[20] Osborne, J. (2010). Improving your data transformations: Applying the Box–Cox transformation, *Practical Assessment, Research & Evaluation*, **15**(12). `http://pareonline.net/getvn.asp?v=15&n=12`.

[21] R Development Core Team (2015). A language and environment for statistical computing, *R Foundation for Statistical Computing*, Vienna, Austria. `http://www.R-project.org`.

[22] Ripley, B.; Venables, B.; Bates, D.M.; Hornik, K.; Gebhardt, A. and Firth, D. (2015). *Package MASS*. `http://www.stats.ox.ac.uk/pub/MASS4`.

[23] Wald, A. (1947). *Sequential Analysis*, John Wiley and Sons, New York.

[24] Woods, D.C.; Lewis, S.M.; Eccleston, J.A. and Russel, K.G. (2006). Designs for Generalized Linear Models With Several Variables and Model Uncertainty, *Technometrics*, **48**, 284–292.

[25] Wooldridge, J. (2003). *Introductory Econometrics: A Modern Approach*, Second edition, Thomson South-Western.

[26] Zellner (1962). Exponential dispersion models, *Journal of the Royal Statistical Society*, Series B, **49**(2), 127–162.