Annals of Actuarial Science, Vol. 6, part 2, pp. 231–234. © Institute and Faculty of Actuaries 2012
doi:10.1017/S1748499512000115

GUEST EDITORIAL

Sexless and beautiful data: from quantity to quality

Montserrat Guillén*
Department of Econometrics, Riskcenter-IREA, University of Barcelona, Spain

Actuarial science has received an enormous influence from statistics since the early times. However, in the recent decades, the interplay between those two disciplines is somehow different compared to the past, with more emphasis on large-scale data analysis and prediction modelling based on customer historical records. In this respect, very little seems to have been said about the quality of data that are being stored in insurance companies' databases, even though data are the fuel for actuarial modelling.

The first idea that came into my mind when the Editor invited me to write this guest editorial was to address gender discrimination. The new European regulation, that is supposed to be effective at the end of 2012, imposes a ban on insurance services which are not allowed to treat men and women as different types of customers. No news, on the grounds that this is already the case in some other parts of the globe. But then I thought this topic is just a small grain of sand in the desert, compared to a much wider problem. How much data on customers is gathered by insurers? Is it worth storing billions of Megabytes? Who takes care about data quality? How dependent are we on the good quality of insurance data? Unfortunately, we all believe, and especially IT departments do, that a treasure is hidden in the databases. Something that is unknown, but extremely valuable. As a result, a significantly larger amount of budget is spent on backing up, than on looking at what is useful in the data files.

Let me first briefly comment on the European gender nightmare and then, in the second part, I will return to the importance of data quality.

Gender discrimination and lifestyle habits

Gender is a simple variable usually having at most four levels: man, woman, unknown or not applicable. The latter is used when a policyholder is not a person but a legal entity. However, let us concentrate on humans. Gender has been used by ancient actuaries, but our modern view in Europe is about to terminate the male/female dichotomy when setting insurance policy prices. Gender potentially splits the population of policyholders into two equally sized groups, so we are not talking about a small fraction of high-risk or low-risk individuals, we are talking about two big segments.

I want to emphasize that it is well known by insurance economists that banning costless categorization (like gender) is inefficient (see Crocker & Snow, 1986; Finkelstein *et al.*, 2009, and

*Correspondence to: Dr M. Guillén, Department of Econometrics, University of Barcelona, Diagonal, 690, 08034 Barcelona, Spain. Tel: +34 93 4037039. Fax: +34 93 4021821. E-mail: mguillen@ub.edu

Rothschild, 2011). As a consequence, I find it strange that in the current state of economic crisis, our regulatory bodies impose rules that are known to damage performance, rather than promoting solutions, such as partial social insurance, that are known to address categorization efficiently.

No one seems to have evaluated the aggregated impact on European consumers of banning decisions based on gender, and only a few have talked about the implications for insurers, while some corrupt the whole scene as they find this the perfect excuse to unduly inflate prices and blame others.

Let us admit that we cannot even mention gender when pricing life and non-life policies from now on in the EU. Should actuaries still work under the premise that gender is a risk factor?

The obvious answer is yes. But then, the incoming regulation clearly separates the role of actuaries from the role of marketing and sales departments. Yet another contradiction. While we advocate that risk management has to be integrated in all parts of the insurance business process, we just build a wall between actuaries, for whom gender should be maintained to assess portfolio risk, and price designers, for whom it should not.

A naive and quick solution, suggested to me by a young actuary a few months ago, is to end up with the problem and cope with the regulation. Delete that, now annoying, gender item from all customer data files. I would not do that. I would keep gender information for risk management if I think that for internal reasons it is beneficial, even if not allowed for external purposes. The content of a gender variable has much valuable information about the portfolio composition because it is a virtually costless categorization. Gender is correct and reliable information about the insured.

The alternative way out suggested by theoretical insurance economists is the creation of a varied menu of contracts from which insureds should choose, and thus reveal information. It is a valuable thought, but it may be too slow in practice. However, I find it interesting as it could re-inspire the concept of "best restaurant" awards.

So, what are the next steps?

a) Keep gender, a variable that once upon a time was a good predictor. Use it to assess risk, or potential adverse selection in the portfolio.
b) Start correlating gender information with every single piece of data for the corresponding customer. Sooner or later another indicator will prove to be both legal and about as good a predictor as gender.
c) Introduce lifestyle information in actuarial modelling, which is well known to be even more informative than gender. The main problem when replacing gender information by lifestyle data is moral hazard and the cost of verifying that a policyholder has not provided fraudulent information. And there we get to the quality issue.

Data quality

Banning the use of a variable for pricing could indicate a concern about the content of historical information. This is far from reality. No one seems to care much about the use and abuse of data storage. Data are accumulated in insurance companies, just in case they are needed, and as a brute source of information. It has become increasingly cheaper to store in the cloud, somewhere, even on a daily basis.

How do we, as actuaries, use statistical data? In general, data are assumed to be perfect, fixed and known. Model predictions are conditioned on the fact that databases are magically ideal, so-called "clean". No information on data quality, measurement error, or the like, is usually provided in research papers that show empirical work. I would advocate in favour of many more contributions with an empirical content to be published, and I would also require authors to provide extensive information and details about how they addressed missing data, imputation practices or sampling designs, whenever necessary. Something that is common in many other scientific fields should also become routine in ours. This is vital if we want to promote a balance between theoretical advances and practical implementations. We need to be very rigorous when it comes to presenting and evaluating applied actuarial modelling. Having myself been involved in the analysis of many insurance databases, I can confirm that looking into the data usually absorbs much time, but it also guides model specification in the right direction. Sound empirical contributions in our scientific journals should be required to comply with these minimum rules.

As far as practice is concerned, real-life huge data sets require data quality assessment. Measuring data quality is a concept that has six dimensions: validity, integrity, completeness, consistency, accuracy and timeliness. Validity refers to the recorded values actually lying within the range of possible values. Integrity is coherence among entities and attributes, like birth date being prior to driving license issuance date, for instance. Completeness shows that necessary data is present. Those three magnitudes are easily verifiable. The other three turn out to be an Achilles' heel. Consistency, which refers to duplication and cross-systems checks, accuracy, which indicates that data require verification, and timeliness, which means that the information should constantly be updated, are difficult in practice.

I strongly support further cooperation between actuaries, who are the end-users of data, and data managers, who are responsible for databases, when defining key performance and data quality indicators.

Expert opinion and qualitative information is still far from being optimally integrated in statistical modelling, but some signs indicate a looming revolution in this direction (Martínez Miranda *et al.*, 2012). Moreover, there has been an explosion of new methods in statistical science that are readily applicable to actuarial problems. Most of those methods rely on an intensive use of good quality data. To name only a few examples: kernel smoothing (Bolancé *et al.*, 2003), hierarchical models (Frees & Valdez, 2008), panel data analysis (Boucher & Denuit, 2006), and multi-level models (Antonio *et al.*, 2010). Those advances in applied statistics would mean a lot more to us, should the quality of insurance information be higher than it is at present.

Concluding remarks

Insurers spend money in making data servers secure. However, much of the precious information in the data is currently hidden in just a small group of variables, which are really the ones that are neat and verifiable, or in qualitative sources. These are beautiful data.

I want to thank all of you to let me share my thoughts on gender discrimination as an open window to this much richer spectrum of problems. My view is that more attention should be given to data analysis prior to actuarial modelling.

Data quality rather than data quantity is the right way to go to improve actuarial models in the future. So, while we make this future possible, let me share with you the joy of reading the current issue of Annals of Actuarial Science.

References

Antonio, K., Frees, E.W. & Valdez, E.A. (2010). A multilevel analysis of intercompany claim counts. ASTIN Bulletin, *The Journal of the International Actuarial Association*, **40**(1), 151–177.

Bolancé, C., Guillén, M. & Nielsen, J.P. (2003). Kernel density estimation of actuarial loss functions. *Insurance: Mathematics and Economics*, **32**(1), 19–34.

Boucher, J.P. & Denuit, M. (2006). Fixed versus random effects in Poisson regression models for claim counts: case study with motor insurance. *ASTIN Bulletin*, **36**, 285–301.

Crocker, K. & Snow, A. (1986). The efficiency effect of categorical discrimination in the insurance industry. *Journal of Political Economy*, **94**(2), 321–344.

Finkelstein, A., Poterba, J. & Rothschild, C. (2009). Redistribution by insurance market regulation: Analyzing a ban on gender-based retirement annuities. *Journal of Financial Economics*, **91**(1), 38–58.

Frees, E.W. & Valdez, E.A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, **103**(484), 1457–1469.

Martínez Miranda, M.D., Nielsen, J.P. & Verrall, R. (2012). Double chain ladder. *ASTIN Bulletin*, **42**(1), 59–76.

Rothschild, C. (2011). The Efficiency of Categorical Discrimination in Insurance Markets. *Journal of Risk and Insurance*, **78**, 267–285.