The underwriter and the models - solo dances or pas-de-deux?

What policy data can tell us about how underwriters use models

Stuart Armstrong, Mario Weick, Anders Sandberg, Andrew Snyder-Beattie, Nick Beckstead

msamlin.com

Executive Summary

Using a collection of data on some of the catastrophe policies written since 2006 by a major reinsurance organisation, this paper explores how tightly underwriters follow the models and under what conditions they deviate from them.

The data included underwriter premium and LE (loss estimate), and sometimes included LE from up to four different models (AIR, RMS ALM, RMS DLM, and IHM – the in-house model). We analysed the data in order to see what could be said about the relationship between model LE and premiums (as well as underwriter LEs). Mimicking a common procedure in machine learning, the data was randomly divided into a training set and a testing set, allowing many different theories to be investigated on the training set without risk of overfitting and detecting spurious connections.

Results

There were three main results, all statistically significant and with large effect sizes:

1. The models gave good predictions as to what the underwriter premium and LE would be. In a regression test, 79% of the variance in the premium, and 88% of variance in the LE, was explained by variance in the models. In fact, most of this variance was explained by the mean LE of the four models – 78% and 87% respectively, corresponding to correlations of 0.88 and 0.93.

2. As the modelled loses rose, underwriter estimates moved closer to those of the models. This was evident through a variety of measures: the underwriters would report using more models than otherwise, the underwriter LE would become more strongly correlated with the mean model LE, and they would have less extreme premium/LE ratios (meaning that the LE information was being used more to fix the premium).

It should be noted that one effect that we expected to find - that underwriters were more willing to follow the models if these were more closely bunched together - was not present in the data. It seems that the underwriter does not make much use of model spread.

The role of underwriters

As far as can be seen in the data, the underwriters premiums (and LE) were strongly correlated with the models' LE estimates. The higher the expected loss (as seen by model LE), the less likely the underwriters were to deviate from the models.

That high correlation may seem to suggest a limited role to the underwriter. However, this conclusion is premature for several reasons. Most significantly, this dataset only included policies that the reinsurance organisation had actually written, so the role of underwriters in rejecting policies could have been very important. There are also issues of yearly variation and changing market conditions (2012, for instance, seems to be an outlier in many ways). This analysis also ignores the effect of underwriters negotiating and interacting with brokers – even if every good underwriter were to sign similar kinds of deals, this does not mean the underwriters were superfluous. It is significant that the models were more predictive of underwriter LE than of premium (which would be influenced by negotiation).

Underwriters may also play an important role in correcting erroneous information in the policy, and making sure that the correct models were applied in the first place. Finally, there were no details of outcomes in the data (which policies led to pay-outs, and by how much?), limiting our ability to estimate underwriter expertise. Thus it is possible and likely that the underwriter played (or could play) a more synergistic role with models, focusing on quality control, market insights, and business relations.

Introduction

The underwriters have traditionally been the key players in the insurance industry, making the final decision on any particular policy. It is their responsibility to negotiate a price and ultimately accept or decline to insure the risk. But recent years have seen the emergence of another key insurance player: computerised models that give their own estimation of risk, exposure, and other critical features of a policy. In CAT (catastrophe) insurance, these models are now used extensively by insurers, re-insurers, and regulators, and have underpinned the risk-linked securities markets.[1]

In this new, model-centric world, what is the current role of the underwriter? More usefully, what will the role of the underwriter become? Some studies suggest that the underwriter is soon to be replaced by automation.[2] The Frey and Osborne study analysed current automation trends and concluded, based on O*NET data (an online job classification service developed for the US Department of Labor[3]), that underwriting involved none of the skills estimated to be hard to automate such as manual dexterity, strategic intelligence or socially dependent tasks[4] (see Table 1).

Computerisation bottleneck	O*NET Variable	O*NET Description
Perception and Manipulation	Finger Dexterity	The ability to make precisely coordinated movements of the fingers of one or both hands to grasp, manipulate, or assemble very small objects.
	Manual Dexterity	The ability to quickly move your hand, your hand together with your arm, or your two hands to grasp, manipulate, or assemble objects. How often does this job require working in cramped work spaces that requires getting into awkward positions?
Creative Intelligence	Originality	The ability to come up with unusual or clever ideas about a given topic or situation, or to develop creative ways to solve a problem.
	Fine Arts	Knowledge of theory and techniques required to compose, produce, and perform works of music, dance, visual arts, drama, and sculpture.
Social Intelligence	Social Perceptiveness	Being aware of others' reactions and understanding why they react as they do.
	Negotiation	Bringing others together and trying to reconcile differences.
	Persuasion	Persuading others to change their minds or behaviour.
	Assisting and Caring for Others	Providing personal assistance, medical attention, emotional support, or other personal care to others such as coworkers, customers, or patients.

Table 1 O*NET variables that are bottlenecks to automation, according to Frey and Osborne.

[1] US Gov't Accountability Office, GAO-02-941, "Catastrophe Insurance Risks: The Role of Risk-Linked Securities and Factors Affecting Their Use" (2002).

[2] Frey, Carl Benedikt, and Osborne, Michael. "The future of employment: how susceptible are jobs to computerisation?" Oxford Martin School Working Paper (2013).

[3] http://www.onetonline.org/

[4] Or that, when it did involve these skills, that they were not fundamental to the job.

This thesis can be strongly questioned. First, the social role of the underwriter should not be underestimated: they are involved in negotiations with brokers, and require perceptiveness of the opinions and behaviour of the other market actors. Secondly, when the practical role of the underwriter was analysed in detail, some of the work involved coping with poor data quality and correcting model errors.[5] Though these tasks do not appear on the list of bottlenecks to automation, they are certainly tasks that cannot be easily automated, as they represent a failure of the automation process itself. Currently (and for the foreseeable future), only humans possess the skills to apply these kind of corrections, which often involve deducing what kind of errors have occurred or what kind of extra data could improve the situation. Thirdly, underwriters may be using strategic intelligence to select models and to maintain the sort of overall vision that could resolve larger systemic risks.[6]

Therefore it would be incorrect to see the underwriters as necessarily in direct competition with the models, but as occupying (potentially) different and complementary roles. This was one of the approaches advocated in the "autopilot problem" paper: change the "pilot's" (or underwriter's) role.[7] Merely relying on effective automation technology can both weaken the skills of the human and make the whole endeavour vulnerable to situations where the automation fails for some reason. However, making use of particular human abilities to control or complement the automation allows for both greater performance and better robustness. For instance, though simple linear models outperform expert predictions in many domains, these simple linear models could only be constructed thanks to expert knowledge of the important factors.[8] Similarly, experienced underwriters are potentially aware of what information is relevant in a particular type of case – and whether such a case is at hand.

The quality of expertise is highly dependent on features of the task,[9] rather than on features of the expert, so it is important that the task the underwriter performs be well designed to make best use of their human qualities (especially as expertise tends to be quite specific to the task performed[10]). If this can be achieved, the underwriter and models of the future will amplify each other.

We thus need to understand how underwriters currently interact with models. To this end, a major reinsurance organisation has made available several years' worth of records on policies priced by its underwriters and by its models so that a comparison could be made and the role of the underwriter teased out. The most useful features of this data were the premiums that were actually charged, the underwriters' Loss Estimate (LE), and the same LEs as given by the models. Using this data, this paper explores how tightly underwriters follow the models and under what conditions they deviate from them.

[5] This was the judgement formed by three of the paper's authors during several periods of immersion at the reinsurance organisation, which involved detailed conversations with, and questioning of, employees in various roles within the company.

[6] Sandberg, Anders. "Defining Systemic Risk" in "Systemic Risk of Modelling." Joint Future of Humanity Institute-MS Amlin White Paper 1 (2014).

[7] Armstrong, Stuart. "The Autopilot Problem" in "Systemic Risk of Modelling." Joint Future of Humanity Institute-MS Amlin White Paper 2 (2014).

[8] Dawes, Robyn M. "The robust beauty of improper linear models in decision making." American psychologist 34.7 (1979): 571.

[9] Shanteau, James. "Competence in experts: The role of task characteristics." Organizational behavior and human decision processes 53.2 (1992): 252-266.

[10] Weiss, David J, and James Shanteau. "Decloaking the privileged expert." Journal of Management and Organization 18 (2012): 300-310.

Data and Methods

The reinsurance organisation provided a large collection of policies with appropriate premium and LE information. Data was taken from two sources: a reinsurance system used to record catastrophe model outputs; and an underwriting system used to record class of business and other risk details. The data were compiled over several years since 2006, with modellers recording model outputs and underwriting teams recording risk details as business was transacted. The data included premium charged, underwriter LE, general location, year, and potential LEs from up to four different models. LE is the loss estimate – the mean amount of money that the underwriter expects their company will pay out on that policy. The more usual "loss on line" is the LE for a particular "layer" of insurance, divided by the limit for that layer. The four models were AIR Catrader (AIR), RMS RiskLink Aggregate level model (RMS ALM), RMS RiskLink Detailed level model (RMS DLM), and the organisation's in-house model (IHM).

After cleaning the data and restricting to US/Canada policies (where the models are the most reliable), we were left with a collection of 660 policies where all four models were used[11] – no further selection was applied to this set.

It was decided to split the data into a training and a testing set. This allows "cross validation", where hypotheses are formed on the training set and tested on the testing set.[12] This prevents overfitting, in which hypotheses are tailored to narrowly to the data, modelling noise rather than signal.[13] The training set would be thoroughly analysed to generate hypotheses; these would then be tested for statistical significance on the testing set. The ideal size of the testing set is 1/3 of the original set.[14] As the splitting into testing and training was done before restricting down to the 660 policies, this subset ended up divided into 457 policies in the training set, and 203 in the testing set.

A total of 32 hypotheses were formed on the training set, which were then tested on the testing set, and all were found to be significant at the 5% level[15], even when accounting for multiple comparisons[16]. Most of these hypotheses were linear regressions, but other comparisons were made as well (see detail of results).

[11] Out of a collection of 6138 US policies which used four or less models.

[12] Picard, Richard R, and R Dennis Cook. "Cross-validation of regression models." Journal of the American Statistical Association 79.387 (1984): 575-583.

[13] Hawkins, Douglas M. "The problem of overfitting." Journal of chemical information and computer sciences 44.1 (2004): 1-12.

[14] Dobbin, Kevin K, and Richard M Simon. "Optimally splitting cases for training and testing high dimensional classifiers." BMC medical genomics 4.1 (2011): 31.

[15] And all but two were significant at the 1% level.

[16] Benjamini, Yoav. "Simultaneous and selective inference: current successes and future challenges." Biometrical Journal 52.6 (2010): 708-721.

Results

Three primary results were prominent in this initial analysis. First, variance in the model's LE explained the majority of the variance in premium (and in underwriter's LE). Second, underwriters tend to be conservative in estimating losses, more often setting expected losses above those of the models rather than below them. Last, the premiums (and underwriters' LE) moved closer to the models for more expensive policies.

The causes and implications of these results are still uncertain – this preliminary analysis is only capable of identifying correlations, not causations.

Result 1: Heavy Use of Quantitative Models

A variety of regressions supported the notion that variations in model LEs explain most of the variance in premium (and underwriter LEs). Regressing[17] premium against the four model's LE resulted in an R^2 of 0.79, while regressing[18] premium against mean LE (of all four models) resulted in an R^2 of 0.78 (implying that 79% and 78% of the variance in premium is explained by variance in model LE and mean model LE, respectively).

Note that the high R^2 need not mean that the underwriters are explicitly using the mean LE in their estimates – the models are highly correlated with each other, as can be seen in Table 2. Thus many linear or quasi-linear combinations of models will be highly correlated with the models, with their mean, and hence with premium.

Indeed, the correlation between premium and models is quite comparable with the correlation the models have with each other. The premium is actually more highly correlated with the other models than the in-house model (IHM) is. Or, put another way, premium deviation with model LE is comparable to the noise[19] in the model LEs.

	Premium	AIR LE	RMS ALM LE	RMS DLM LE	IHM LE	Mean LE
Premium	1	0.825	0.808	0.796	0.852	0.881
AIR LE	0.825	1	0.909	0.864	0.878	0.966
RMS ALM LE	0.808	0.909	1	0.854	0.797	0.930
RMS DLM LE	0.796	0.864	0.854	1	0.758	0.915
IHM LE	0.852	0.878	0.797	0.758	1	0.937
Mean LE	0.881	0.966	0.930	0.915	0.937	1

Table 2 Correlations between premium, model LEs, and mean model LE.

[17] Of the form $U_i = \beta_0 + \beta_1 M_{AIRi} + \beta_2 M_{RMS\text{-}ALMi} + \beta_3 M_{RMS\text{-}DLMi} + \beta_{4\ MGMi} + \varepsilon_i$ where U is the underwriters' LE, M_X is the LE of the X'th model, and ε_i is the error term for the i'th policy.

[18] Of the form $U_i = \beta_0 + \beta_1 M_{MEANi} + \varepsilon_i$ where U is the underwriters' LE, M_i is the average LE of all four models, and ε_i is the error term for the i'th policy.

[19] Using model LE variation as an informal measure of noise.

This strong correlation could be a sign of an autopilot problem[20], if underwriters put too much trust in the models.

In the meantime, we can analyse how these correlations change over time. For example, Table 3 demonstrates how in 2012, the fit between premium and model LE declined, as compared to surrounding years. This was around the period when the controversial "RMS 11" windstorm model was in use. Though the controversy was mainly around European windstorms, not US/Canada ones, this could have had an impact on underwriter trust in models. Alternatively, the large losses in 2011 could have played a similar role.[21]

Year	# of policies	All Model LE to Premium R2	Mean Model LE to Premium R2	All Model LE to Underwriter LE R2	Mean Model LE to Underwriter LE R2
2010	74	91.32%	76.08%	91.65%	89.87%
2011	345	84.94%	82.01%	91.65%	90.94%
2012	145	75.69%	69.56%	80.13%	75.58%
2013	92	95.15%	83.22%	94.83%	88.18%
Mean		86.78%	77.72%	89.57%	86.14%
Combined	656	78.80%	77.55%	87.63%	86.95%

Table 3 R^2 between premiums and Underwriter LE with models, by year.

Result 2: Conservative Pricing

Underwriters tended to set their LEs somewhat conservatively. Underwriter LE were above the minimum model LE 94% of the time (as compared to below the maximum model 83% of the time). See Table 5 in next section for breakdown of this into highest and lowest quartiles.[22]

This was especially interesting as the maximum value and the mean were both better predictors of underwriter LE than the minimum (see Figure 1 and Table 4). Thus, it seems the minimum value was a rough lower bound for underwriters, but that they used little information from it beyond this.

	Mean	Max	Min
Underwriter LE	0.933	0.913	0.863

Table 4 Correlation coefficients between underwriter LE and the mean, maximum, and minimum of the model LEs.

[20] Armstrong, Stuart. "The Autopilot Problem" in "Systemic Risk of Modelling." Joint Future of Humanity Institute-MS Amlin White Paper 2 (2014).

[21] Aon Benfield. "Reinsurance Market Outlook." Aon Benfield Analytics (2013).

[22] Note that we are using quartile in the sense of a set of data representing a quarter of the policies, not in the sense of the three values (lower quartiles, median, upper quartile) that divide the ranked data into those four sets.

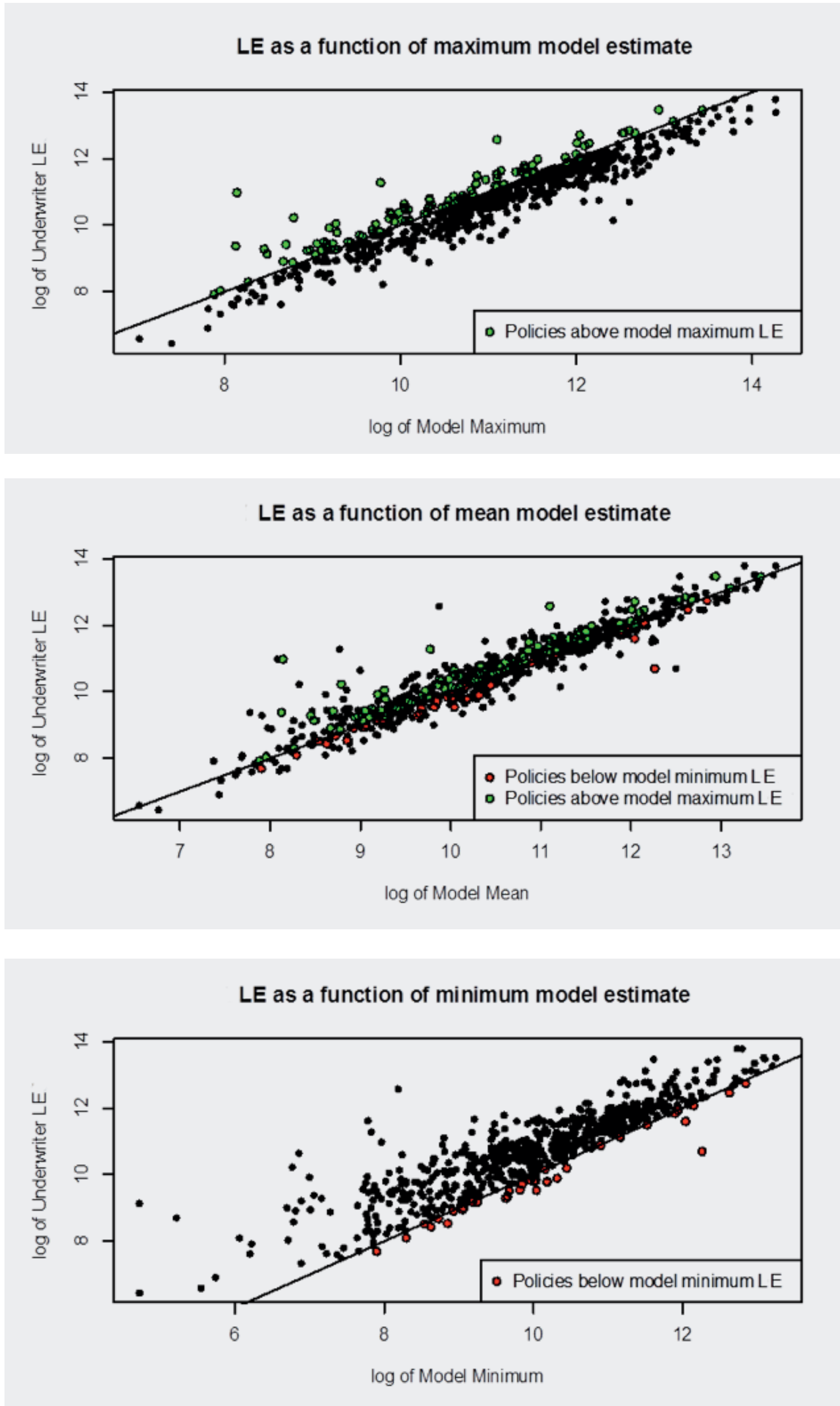

Figure 1 - Underwriter's LE as a function of cat model LE. With a set of models, underwriters are unlikely to set LE below the lowest model, and tend to stick fairly close to the mean model estimate.

Result 3: Model Dependency for Biggest Risks

There were many indications that the underwriters tended to move closer to the model estimate as the LE moved higher.

First of all was the simple fact that they used (or at least recorded the use of) more models. The full US/ Canada data (6138 policies) was decomposed into categories dependent on the number of models recorded[23], and the median premium and underwriter LE of each category was computed (see Figure 2). For illustration, the interquartile distance was added in the case of zero and four models. Though the ranges overlap, those policies which recorded two or less models clearly had lower LEs and premiums than those that recorded three models or four.

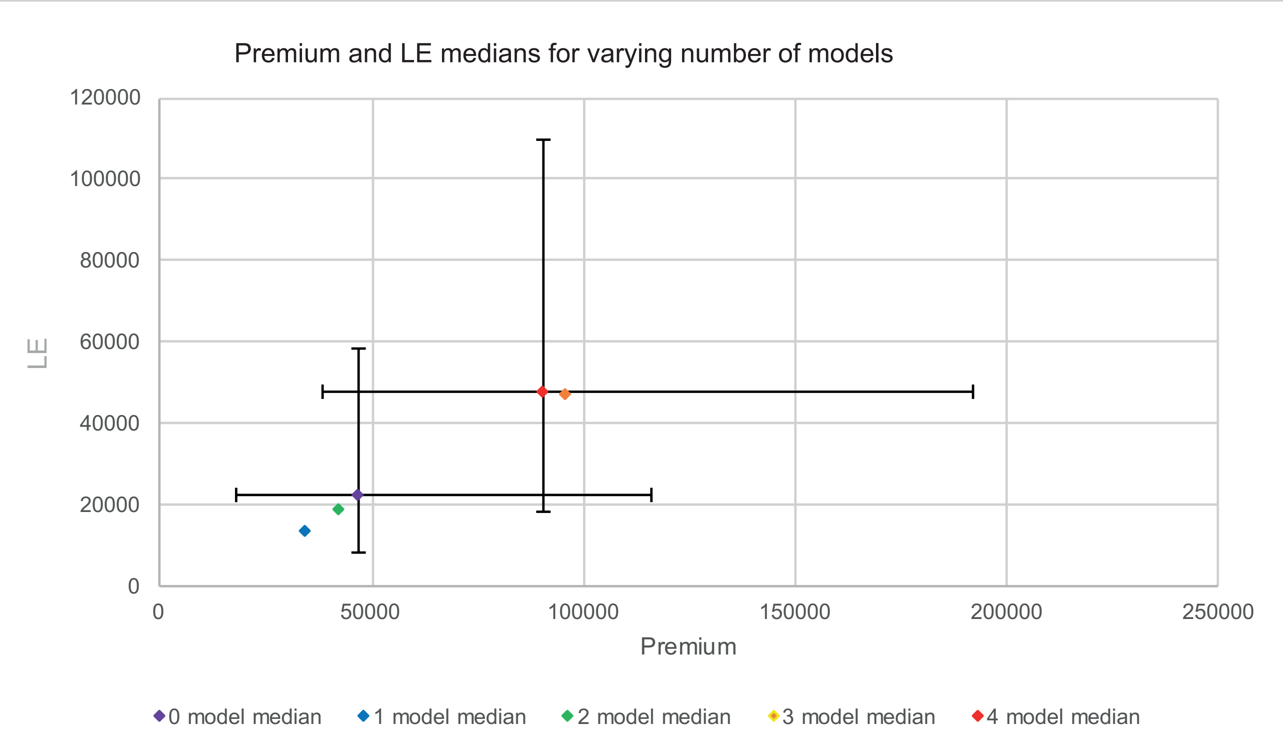

Figure 2 - Median underwriter LE and premiums, depending on how many models were reported for that policy. The inter-quartile range for 4 models and 0 models is also plotted.

[23] 5071 policies did not formally record a modelled loss, 27 had one, 85 had two, 295 had 3 and 660 had all four models.

Furthermore, the correlation between underwriter LE and model LE increased for higher model LE. To see this, the 660 policies with four models were separated into quartiles.[24] The correlation between the underwriter LE and mean, minimum, and maximum model LE was computed in each quartile (Table 5). The data suggests underwriters hewed closer to the model information in these quartiles.

	1st quartile	2nd quartile	3rd quartile	4th quartile
Mean	0.449	0.233	0.598	0.876
Max	0.486	0.387	0.363	0.847
Min	0.231	-0.035	0.329	0.747

Table 5 Correlation coefficients between underwriter LE and mean model LE, maximum model LE, and minimum model LE, separated onto the four quartiles.

A last way of looking at the difference in behaviour, is to consider what happens when the underwriter has established their LE, and compare this with the final premium. The data indicates that for high underwriter LE[25], the premium/LE ratio is likely to be less extreme that for low ones. The linear anti-correlation is weak (-11%), but the effect is clearer graphically (see Figure 3). Thus the premium is more influenced by LE estimates when these are large.

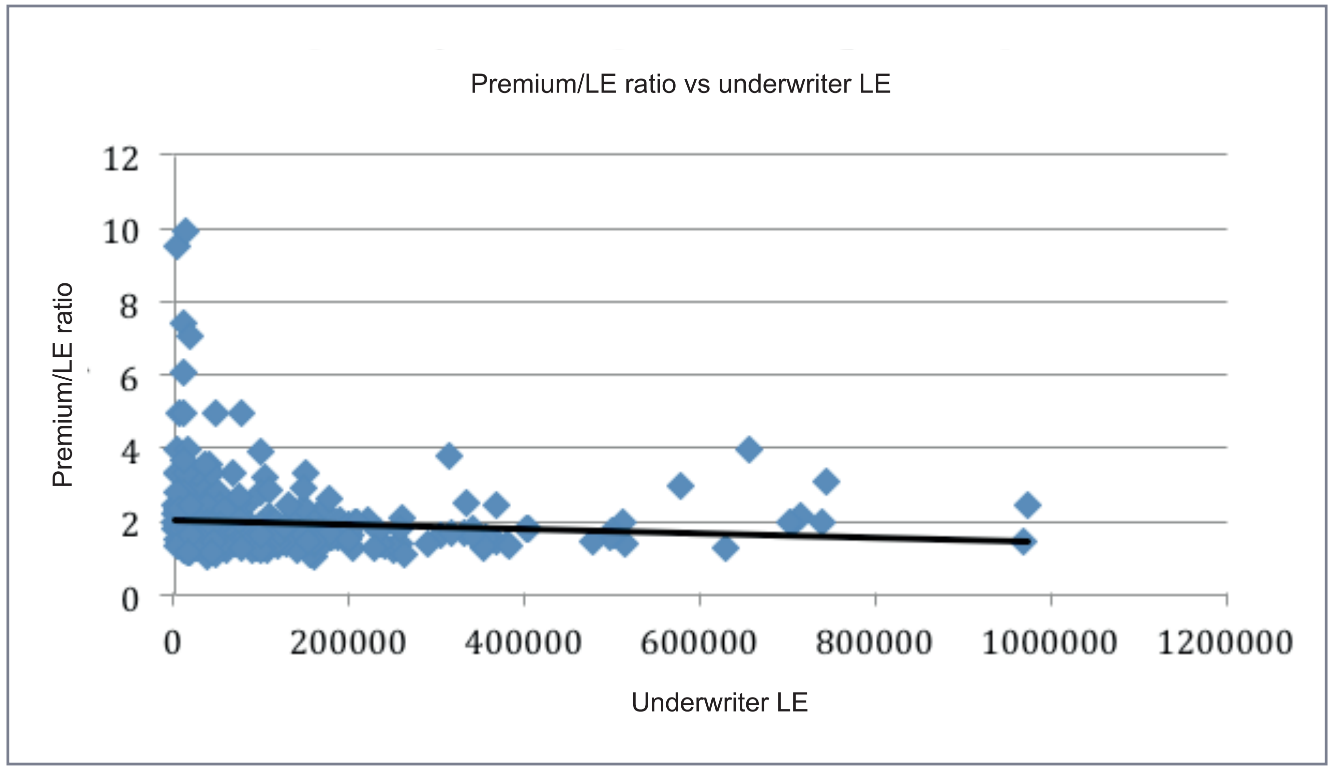

Figure 3 - Ratio of premium to underwriter LE plotted against underwriter LE.

[24] That is, the policies were ranked according to their mean LE, and they were split into four sets at lower quartile, the median, and the upper quartile. By an abuse of nomenclature, these four sets are also called quartiles.

[25] Since the underwriter would set their own LE before setting a premium, it is valid to use underwriter LE on the x-axis at this point.

Statistical testing and significance

To establish the results overleaf, a total of 32 hypotheses were tested. The significance of each result was established using the testing set, with each p-value adjusted – increased[26] – using the Holm-Bonferroni method. See table 6 (where “HBp-value” is the Holm-Bonferroni[27] adjusted p-value). As can be seen, all hypotheses were significant at the 5% level, and all but two were significant at the 1% level (these significance levels were marked in bold).

Hypothesis	p-value	HBp-value	Hypothesis	p-value	HBp-value
LE model mean correlation	2.20E-16	**7.04E-15**	LE all model regression 2013	3.37E-12	**5.39E-11**
LE model mean correlation 2010-2013	2.20E-16	**6.82E-15**	Premium all model regression 2013	7.18E-12	**1.08E-10**
LE model mean correlation 2011	2.20E-16	**6.60E-15**	Premium correlation model mean 2010	7.90E-12	**1.11E-10**
LE model mean correlation 2012	2.20E-16	**6.38E-15**	LE model mean correlation 2010	3.11E-11	**4.04E-10**
LE all model regression	2.20E-16	**6.16E-15**	mean quartile difference (2-tailed)	2.60E-09	**3.12E-08**
LE all model regression 2010-2013	2.20E-16	**5.94E-15**	LE all model regression 2010	4.99E-09	**5.49E-08**
LE all model regression 2011	2.20E-16	**5.72E-15**	Premium all model regression 2010	5.02E-09	**5.02E-08**
Premium all model regression 2010-2013	2.20E-16	**5.50E-15**	Underwriter LE difference for number of models	2.94E-06	**2.64E-05**
Premium all model regression 2011	2.20E-16	**5.28E-15**	max quartile difference (2-tailed)	6.08E-06	**4.87E-05**
Premium correlation model mean 2010-2013	2.20E-16	**5.06E-15**	equivalence between mean and full regression	6.08E-06	**4.26E-05**
Premium correlation model mean 2011	2.20E-16	**4.84E-15**	min quartile difference (2-tailed)	9.20E-05	**5.52E-04**
Premium correlation model mean 2012	2.20E-16	**4.62E-15**	mean more important than min (1-tailed)	1.86E-04	**7.45E-04**
LE all model regression 2012	6.93E-16	**1.39E-14**	Premium difference for multiple models	2.05E-04	**8.22E-04**
Premium all model regression 2012	1.00E-15	**1.90E-14**	correlation between underwriter LE and premium ratio	9.04E-04	**2.71E-03**
LE model mean correlation 2013	9.80E-15	**1.76E-13**	max more important than min (1-tailed)	1.59E-02	**3.18E-02**
Premium correlation model mean 2013	9.34E-14	**1.59E-12**	correlation between model mean and absolute deviation %	4.85E-02	**4.85E-02**

Table 6 p-values and Holm-Bonferroni corrected p-values (for multiple comparisons) for all 32 hypotheses considered.

[26] p-values of the individual hypotheses were calculated separately, then the hypotheses were ordered by increasing p-value. These p-values were then multiplied by (32+1-r), where r was the rank of the hypothesis on the list. Thus the hypothesis with the lowest p-value had this value multiplied by 32, that with the next-lowest had its value multiplied by 31, all the way down to the one with the highest p-value, which was multiplied by 1. As long as these adjusted p-values were below the criteria of significance, then the familywise error rate would be below that.

[27] Holm, Sture. “A simple sequentially rejective multiple test procedure.” Scandinavian journal of statistics (1979): 65-70.

Of these hypotheses, 22 were regression/correlations with premium or underwriter LE as dependent variable, and either all model LEs or mean model LE as independent variables. We should expect these regressions to show scale invariance to some extent: if all the model LEs double, then, say, the underwriter LE should also double as well. This means that we expect the residuals (the deviations of the underwriter LE from its “theoretic value” as predicted by the regression model) to double as well. Thus we expect residuals to be higher for high expected losses and lower for low expected losses: the data should be heteroscedastic. And indeed it is (see Figure 4).

Excessive heteroscedasticity precludes the use of the standard F-test to determine the p-values for the model. Instead, we took the logarithms of all the variables, expecting that this would remove the scale variations in the residuals. The residuals that resulted were much closer to being homoscedastic (see Figure 4).

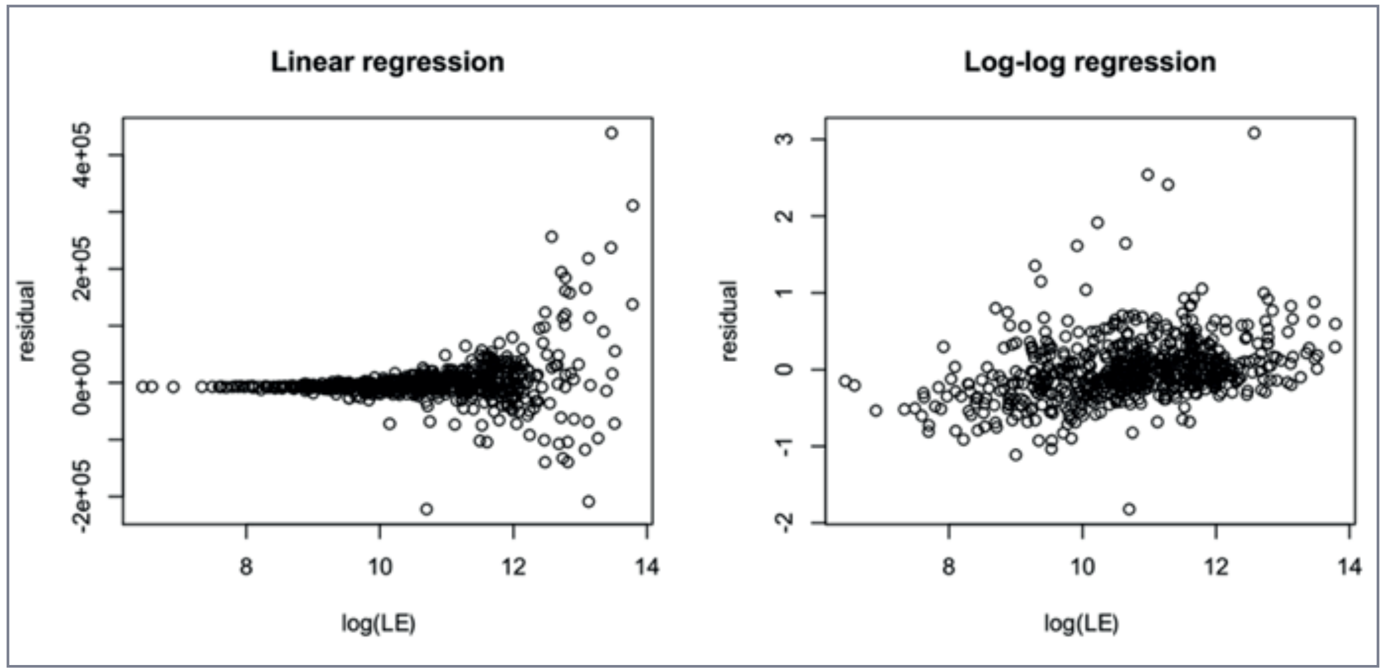

Figure 4 - Residuals for linear and log-log regression for underwriter LE versus all four model LEs.

Thus the p-values of these regressions and correlations were calculated using the log-log regression rather than the standard linear regression. These p-values were sufficient to show that all regression models were significantly different from the null hypothesis. All regression data given in this paper (the R2 and the correlation coefficients), however, came from a standard linear regression, as a linear model is more likely to be closer to the underwriters’ behaviour than a logarithmic one. Generally, the logarithmic model had slightly lower R2 than the corresponding linear model, but the values were very close (and on occasion the logarithmic model had a slightly better fit than the linear model).

Discussion

The high correlation between model-estimated LE and premium (and underwriter LE) may seem to suggest a limited role to the underwriter. However, this conclusion is premature for several reasons. For a start, this dataset is likely incomplete, as it had to be put together specifically, and many policies failed to record model estimates.[28] Most importantly, it only included policies that had actually been written; the role of underwriters in rejecting policies could have been very important. The overall result ignores the fact of yearly variation: in particular, 2012 fits very poorly into the general analysis. It is likely that underwriters were aware of changing market conditions (or changes to the models themselves) and were able to react to them accordingly in that year.

This analysis also ignores the effect of underwriters negotiating and interacting with brokers. It is significant that the models were more predictive of underwriter LE than of premium[29] (which would be influenced by negotiation). Underwriters may also play an important role in correcting erroneous information in the policy, and making sure that the correct models were applied in the first place[30]. Finally, there were no details of outcomes in the data (which policies led to payouts, and by how much?), limiting our ability to estimate underwriter expertise.[31]

Conversely, there could have been a lot of wasted effort on the underwriter's part. The four models are so highly correlated (see Table 4) that any attempts by the underwriters to strike a fine balance between them would likely have made little impact.

Given all those caveats, the story that the data presents is clear. If this sample is taken to be representative, then there seems little difference between using models to estimate loss, and having the underwriters do the same. More importantly, a similar pattern is true to a lesser extent in premium, where around 80% of the variance in premiums is explained by the models in a completely linear fashion.

The remaining 20% is unlikely to represent perfect performance on the part of the underwriter, free of noise and bias. Underwriters, like all humans, are subject to general biases that interfere with their performance[32], some of which could have a specific impact on their job.[33] The important question is how biases balance against expertise within this 20%. We have done some preliminary exploration of these biases and counteracting expertise with an experimental pilot study.[34] In that study, the underwriters appeared less model-bound than in this paper: the actual conditions of work may play a large role in decisionmaking. A full analysis of the underwriters' work is needed if the aim is to develop expertise that complements models in a robust way.

We believe that data analysis such as this paper, an understanding of the cognition of underwriting, and a systemic perspective on the inherent risks of outsourced cognition – whether from autopilot biases or model-induced correlations across the insurance market – can help develop practices that both reduce systemic risk and amplify human capacity. It will become more and more vital for insurance companies to record their own data as this one has, and to analyse it intelligently.

[28] In many cases, it is likely that models were used but the data wasn't recorded.

[29] The models explain 88% of the underwriter LE variance in the 2010-2013 period, but only 79% of the premium variance.

[30] From conversations with people in the industry, this last effect is more likely to be a factor in insurance modelling than in reinsurance modelling, which seems to be a more mechanical process.

[31] And even loss data in the short term is not enough to estimate true underwriter expertise, as many risk as of low probability/high return period, and wouldn't show up in the data.

[32] The bias literature is vast, but Kahneman, Daniel. Thinking, fast and slow. Macmillan, 2011 provides a good overview, Gigerenzer, Gerd. "How to make cognitive illusions disappear: Beyond "heuristics and biases"." European review of social psychology 2.1 (1991): 83-115 provides a good critique, and Kahneman, Daniel, and Gary Klein. "Conditions for intuitive expertise: a failure to disagree." American Psychologist 64.6 (2009): 515 gives a good synthesis of some opposed views on the subject.

[33] Beckstead, Nick. "Biased error search as a risk of modelling in insurance" in "Systemic Risk of Modelling." Joint Future of Humanity Institute-MS Amlin White Paper 3 (2014)

[34] Tilli, Cecilia, and Sandberg, Anders. "Pilot study of underwriter cognitive bias." Joint FHI-MS Amlin White Paper, forthcoming 3 (2016).