Correlations between insurance lines of business: An illusion or a real phenomenon? Some methodological considerations

Benjamin Avanzi[a,b], Greg Taylor[a,*], Bernard Wong[a]

[a] *School of Risk and Actuarial Studies, UNSW Australia Business School*
UNSW Sydney NSW 2052, Australia
[b] *Département de Mathématiques et de Statistique, Université de Montréal*
Montréal QC H3T 1J4, Canada

Abstract

This paper is concerned with dependency between business segments in the non-life insurance industry. When considering the business of an insurance company at the aggregate level, dependence structures can have a major impact in several areas of Enterprise Risk Management, such as in claims reserving and capital modelling. The accurate estimation of the diversification benefits related to the dependence structures between lines of business ("LoBs") is crucial for (i) capital efficiency, as one should avoid holding unnecessarily high levels of capital, and (ii) solvency of the insurance company, as an underestimation, on the other hand, may lead to insufficient capitalisation and safety.

There seems to be a great deal of preconception as to how dependent insurance claims should be. Often, presence of dependence is taken as a given and rarely discussed or challenged, perhaps because of the lack of extensive data sets to be publicly analysed. In this paper, we take a different approach, and consider how much correlation some real data sets actually display (the Meyers-Shi dataset from the USA, and the AUSI dataset from Australia). We develop a simple theoretical framework that enables us to explain how and why correlations can be illusory (and what we mean by that). We show with some real examples that, sometimes, most (if not all) of the correlation can be 'explained' by an appropriate methodology. Two major conclusions stem from our analysis:

1. In any attempt to measure cross-LoB correlations, careful modelling of the data needs to be the order of the day. The exercise will not be well served by rough modelling, such as the use of simple chain ladders, and may indeed result in the prescription of excessive risk margins and/or capital margins.
2. Such empirical evidence as examined in the paper reveals cross-LoB correlations that vary only in the range zero to very modest. There is little evidence in favour of the high correlation assumed in some jurisdictions. The evidence suggests that these assumptions derived from either poor modelling or a misconception of the cross-LoB dependencies relevant to the purpose to which they are applied.

Keywords: Actuarial Models, Dependence, Capital margin, Correlation, Real data, Reserving, Risk margin
JEL codes: C52, C55, G22

1. Introduction

1.1. Motivation

This paper is concerned with **dependency between business segments** in the non-life insurance industry. Particularly relevant examples would be lines or sub-lines of business ("LoBs"). At the broadest

*Corresponding author.
Email addresses: b.avanzi@unsw.edu.au (Benjamin Avanzi), greg.taylor@unsw.edu.au (Greg Taylor), bernard.wong@unsw.edu.au (Bernard Wong)

non-technical level, dependency between any collection of business segments is defined as the phenomenon whereby the experience of one segment varies in sympathy with the experience of one or more of the others (see, e.g. Joe, 1997). This may occur because a common process affects the experiences of the different segments. For example, a hail storm might generate claims in both the Motor and Home LoBs; a legal precedent might cause similar increases in claim costs in the Workers Compensation and Compulsory Third Party LoBs. Independence between business segments means, of course, that the experience of one segment does not change as that of other segments varies.

When considering the business of an insurance company at the aggregate level, dependence structures **can have a major impact in several areas of Enterprise Risk Management**, including

— determining actuarial reserves for outstanding claims: legislations around the world typically require reserves to correspond to a certain quantile of the fair estimate of the distribution of future payments (for instance, 75% in Australia). Alternatively, the reserve might be required to correspond to a central estimate (the mean), to which a certain margin is added, which depends on the variability of the aforementioned distribution;

— determining a risk based capital for solvency assessment purposes (such as required in Solvency II in the EU or the Swiss Solvency Test in Switzerland): this typically corresponds to a quantile of the overall net loss of the insurance company, under some realistic (adverse) scenarios and over varying time horizons (in Australia, APRA GPS 310 requires a 99.5% probability of adequacy over a 1-year horizon).

It is well known that independence between business segments may create **risk diversification** in the sense that the risk associated with a collection of segments is less than the sum of their individual risks. If risk is measured in terms of standard deviation, this diversification result may be represented quantitatively as follows:

$$\sigma_{\text{total}} = \left[\sigma_1^2 + \sigma_2^2 + \ldots\right]^{1/2} < \sigma_1 + \sigma_2 + \ldots, \tag{1.1}$$

where σ_1, σ_2 are the standard deviations of segments 1, 2, etc, and σ_{total} is the standard deviation of the totality (sum) of these segments. Dependency can, in principle, increase or decrease diversification. In insurance, it usually decreases it. An extreme form of dependence called co-monotonicity can eliminate diversification altogether.

The **accurate estimation** of the diversification benefits related to the dependence structures between LoBs is crucial for (i) **capital efficiency**, as one should avoid holding unnecessarily high levels of capital, and (ii) **solvency** of the insurance company, as an underestimation, on the other hand, may lead to insufficient capitalisation and safety.

Dependence and correlation are terms that are often (inaccurately) used interchangeably. The quantity usually referred to as correlation, or correlation coefficient, is in fact the Pearson correlation coefficient. There are a number of others. The Pearson correlation coefficient between two stochastic quantities is a measure of the dependency between them that is well adapted to the case in which the two quantities are linearly related and both can be assumed normally distributed (De Gooijer, 2006). This assumption is often false in insurance applications, sometimes spectacularly so. It may be grossly misleading as a measure of dependency in the tails of the variables (Embrechts et al., 2002). This is of especial relevance to capital management, which is wholly concerned with the tail of the net assets variable.

Nevertheless, in this paper, we will primarily **focus on correlation to specify dependence structures**. Correlation is the predominant tool in actuarial practice. Furthermore, the materiality of non-normality is generally low near the centre of the distributions of the subject variables, where loss reserves at moderate percentiles (as required in some jurisdictions) are located. But also importantly, correlation is familiar and sophisticated enough for us to be able to make our point, and this is a clear and understandable fashion.

1.2. Aims and structure of this paper

While there is theoretical literature in risk theory about claims processes and the modelling of dependence between them, few authors have yet adapted and validated those models directly to the needs of the insurance industry. There seems to be a great deal of preconception as to how dependent insurance datasets should be. An increasing number of academic papers are developing innovative methods for taking into account those

dependence effects. Often, presence of dependence is taken as a given and rarely discussed or challenged, perhaps because of the lack of extensive dataset to be publicly analysed.

In this paper, we take a different approach, and consider how much correlation some real data sets actually display. Sometimes, most correlation can be removed by incorporating some significant effects (such as weather). We develop a simple theoretical framework that enables us to explain rigorously how and why correlations can be illusory (and what we mean by that). This has significant implications as explained above.

Section 2 describes Australian practice, which serves as an insightful case study, and further motivates our work. A careful exposition of our notation and correlation framework is developed in Sections 3 and 4, respectively. In Section 5, we use the aforementioned framework to discuss and decompose the effects of modelling on the estimation of dependencies. We define and decompose errors into three components in order to make our point clear. This is illustrated in Section 5.5 with simulated data, and in Section 6 using US and Australian reserving real data (triangles), whose nature and origin are described in Appendix A. Section 7 discusses why caution should be used when interpreting our findings in the context of forecasting. Section 8 concludes.

Remark 1.1. *We use a stochastic reserving framework to illustrate our arguments. The literature on stochastic reserving has not received a lot of attention until only recently, when new prudential requirements were introduced around the world. Surprisingly, while the underlying problem of obtaining a quantile with dependence is similar in the area of stochastic reserving (see Taylor, 2000; Wüthrich and Merz, 2008, for general references on reserving) to that in capital modelling, very few papers use flexible tools to model dependence between LoBs (such as in Shi and Frees, 2011, with copulas). While we believe this is an exciting research area, this paper does not claim a direct contribution to it. It rather aims at reminding the modeller that some careful consideration of actual dependence (beyond systematic effects) is warranted, as correlations displayed in a dataset may be illusory.*

2. Case study: Australian practice

The co-authors of this paper are engaged in collaborative research with three major Australian non-life insurers (see also Appendix A), and as such, have access to specific insight about the way dependence (and correlation) is modelled on the Australian market. This section offers a description of the Australian situation, which may be of interest to some readers. It also presents the advantage of further motivating the paper, and of providing some support to some of our claims, namely, that correlation is a predominant tool, and that practictioners are at risk of routinely overstating these correlations.

2.1. Legislative framework

In Australia, the non-life insurance industry is governed by *Insurance Act 1973* (Cwlth) ("the Insurance Act"). This Act appoints the Australian Prudential Regulation Authority ("APRA") as the industry regulator (Section 8) and empowers it to issue prudential standards (Section 32). APRA has issued a number of these standards, each with a title of the form GPS[number].

GPS 320: "Actuarial and Related Matters" requires that an insurer's insurance liabilities include a risk margin in addition to the central estimate of those liabilities (paragraph 14). The risk margin is essentially insurer capital. Paragraph 29 permits the total of the individual risk margins across all lines of business to be reduced in allowance for the diversification of risk provided by those lines. Any assessed diversification benefit must have due regard to dependencies between the LoBs. Any mis-estimation of those dependencies will carry a penalty in terms of economic efficiency:

— To the extent that they are under-estimated, the risk margin is also likely to be under-stated, implying more lenient insurance regulation than intended, i.e. a greater likelihood of insurer failure than targeted by APRA;

— To the extent that they are over-estimated, the risk margin requirement is also likely to be over-stated, in which case insurers will be required to over-capitalise, i.e. retain idle capital surplus to their risk requirements, with the consequent opportunity cost.

GPS 113 "Capital Adequacy: Internal Model-based Method" sets out the standard according to which APRA will permit an insurer's capital requirement for regulatory purposes to be determined internally by the insurer. It requires that the insurer have an Economic Capital Model ("ECM") (paragraph 7(a)) and the model is subject to the following especially significant requirements (paragraph 20):

> For each component of risk, the ECM must utilise a distribution with appropriate shape and tail characteristics. In combining components of risk, the model must make appropriate allowance for correlation between risks, particularly correlations in the tail of distributions. A regulated institution wishing to incorporate diversification assumptions in respect of operational risk must demonstrate an adequate process for estimating dependencies (particularly for extreme losses) and must apply conservatism in its assumptions that is commensurate with the uncertainty of those assumptions.

Of the major Australian insurers, one has obtained APRA's approval of its ECM, and all others are at various stages of ECM development. It is evident from the above quotation that dependencies between risk components are central to any ECM and are liable to effect an insurer's capital requirements substantially. Any mis-estimation of those dependencies will carry a penalty in terms of economic efficiency similar to that discussed in the two points above. Furthermore, information on the relation between risk and return is essential to rational economic decision-making, and is enabled only by the existence of an ECM.

2.2. Correlations in Australian practice

In Australia, stochastic modelling received little impetus before 2001, after which the Insurance Act required reserving and capitalisation to a certain quantile, as mentioned above, taking account of the dependency problem. After 2001 Australian insurers have responded with a number of devices for estimation of liability risk margins; and the use of dynamic financial analysis ("DFA") (sometimes called dynamic risk modelling) for the investigation of capital issues (see Britt and Johnstone, 2001; Mulvey et al., 2008, for a review of the key features and applications of DFA models). Perhaps the most commonly used devices are: (i) bootstrapping for the construction of risk margins; (ii) correlation matrices for the evaluation of diversification, and (iii) a balanced scorecard approach to model error and external systemic error.

Specification of the pair-wise correlations between a number of LoBs requires a correlation matrix. For risk margins, for example, the industry frequently relies on the correlation matrices given by Bateup and Reed (2001, see Tables 2.1 and 2.2) or Collings and White (2001). The correlations contained there are based largely on the judgement of a small number of actuaries, with limited under-pinning in data (in one case none). A number of the correlations are of considerable magnitude, with little factual evidence to support them, and are possibly overstated, resulting in deliberately conservative risk margins (overstated). The Risk Margins Task Force (2008) of the Institute of Actuaries of Australia proposed a procedure for the formulation of such correlation matrices, including allowance for model error (note that their work is largely based on O'Dowd et al., 2005). The procedure incorporates many components of risk, and is heavily judgement based. While there is much of value in the Task Force proposal, comparatively little guidance is given on the selection of an algebraic structure that will integrate the components of risk into a matrix in a manner that is logical and succeeds in generating a positive definite matrix[1].

Remark 2.1. *The capital modellers are aware of the limitations of correlation for determining tail-related quantities. There is not much documentation to refer to, but we believe tail dependence measures are often used for capital modelling purposes. This raises the issue of inconsistency of the methodology that is being used in reserving and capital modelling. Further discussion of this is outside the scope of this paper.*

[1]For some potential methods to tackle the problem of non-positive definite matrices, see, for example, Rousseeuw and Molenberghs (1993) and Lindskog (2000).

	ABI	Workers Comp	Prof Ind	Inwards Re	Fire/ ISR	APD	Home	Other
Liab	0.25	0.25	0.25	0.25	0	0	0	0
ABI		0.35	0.25	0.25	0	0.25	0	0
Workers Comp			0.25	0.25	0	0	0	0
Prof Ind				0.25	0	0	0	0
Inwards Re					0.05	0.05	0.05	0.05
Fire/ISR						0.10	0.10	0.05
APD							0.20	0.10
Home								0.10

'APD' refers to Auto Property Damage (Motor in Australia)
'ABI' refers to Auto Bodily Injury (CTP in Australia)

Table 2.1: Assumed Total Variance Correlation Matrix between Lines of Business from Bateup and Reed (2001, Table 7.5)

	ABI	Workers Comp	Prof Ind	Inwards Re	Fire/ ISR	APD	Home	Other
Liab	0.35	0.40	0.45	0.45	0	0	0	0
ABI		0.50	0.40	0.40	0	0.55	0	0
Workers Comp			0.45	0.50	0	0	0	0
Prof Ind				0.55	0	0	0	0
Inwards Re					0.15	0.15	0.15	0.15
Fire/ISR						0.40	0.35	0.10
APD							0.75	0.25
Home								0.20

Table 2.2: Derived Correlation Matrix Applicable to the Systemic Variance Only (i.e. excluding process error) of the net Outstanding Claims Liability from Bateup and Reed (2001, Table 7.6)

2.3. Australian perceptions

Table 2.3 is the result of a workshop organised in March 2014, where the co-authors asked industry representatives (all senior actuaries) to collaboratively specify and qualify the strength of dependencies they expected to arise between different LoBs on the Australian market. They were asked to name major LoBs (columns), risks inducing dependence (rows) and the expected strength of dependence arising from those sources. The latter was to be qualified as minor (correlation significantly more than 0, say 0-20%), major (correlation of about 20-40%) or critical (correlation significantly above 40%).

3. Framework and Notation

The following analysis will be concerned with conventional triangles of incremental paid losses for various insurers and various lines of business. The following notation will be used:

— k indexes accident periods, $k = 1, 2, \ldots, J$;
— j indexes development periods, $j = 1, 2, \ldots, J$;
— n indexes LoBs, $n = 1, 2, \ldots, N$;
— $T = \frac{1}{2}J(J-1)$ is the number of observations in a triangle;
— $Y_{kj}^{(n)}$ is the amount of paid losses in development period j of accident period k for LoB n, assumed to be a realisation of a random variable;

Risk \ LoB	APD	ABI	WC	Liability	Property
Idiosyncratic					
Catastrophic	2		1		3
Inflation		2	2	2	
Social	1	3	3	3	
Fraud		2	1	2	
Internal factors	2	2	2	2	2
Weather	2	2			2

'Social' includes: economic, medical, societal, judicial, political, technology, ...
'Internal factors' includes: same actuary, same manager, data systems, ...

Table 2.3: Heat map of expected sources of dependence and their strength, according to a group of Australian experts.

— $Y^{(n)} = \left\{ Y_{kj}^{(n)} : k = 1, 2, \ldots, J;\ j = 1, 2, \ldots, J-k+1 \right\}$ is the triangle of incremental paid losses for LoB n;
— $P\left(Y_{kj}^{(n)}; x_{kj}^{(n)}, \theta^{(n)}\right)$ is the distribution function of $Y_{kj}^{(n)}$, dependent on parameter vector $\theta^{(n)}$, and a vector $x_{kj}^{(n)}$ of deterministic measurements (covariates) associated with the (k, j) cell;
— $p = \dim \theta^{(n)}$ is the number of parameters defining the distribution of each $Y_{kj}^{(n)}$;
— $\mu_{kj}^{(n)} = \mathbb{E}\left[Y_{kj}^{(n)}\right] = g_\mu\left(x_{kj}^{(n)}, \theta^{(n)}\right)$ for some function g_μ, possibly non-linear; and
— $\sigma_{kj}^{2(n)} = \mathrm{Var}\left[Y_{kj}^{(n)}\right] = g_{\rho^2}\left(x_{kj}^{(n)}, \theta^{(n)}\right)$ for some function g_{ρ^2}.

The notation $\sum_{k,j\in\Delta}(\bullet)$ will be used as an abbreviation for $\sum_{k=1}^{J}\sum_{j=1}^{J-k+1}(\bullet)$, the sum over the entire triangle of (k, j) cells.

Accident periods and development periods will sometimes be years, sometimes quarters. Their definitions will be given at the time of each usage.

It is assumed that a model is fitted to each triangle $Y^{(n)}$. The form of the model is left unspecified at this stage. It will be supposed only that the model, consisting of $\theta^{(n)}$, P, g_μ, g_{ρ^2} has been applied to obtain parameter estimates. It is assumed that p, P, g_μ, g_{ρ^2} are all independent of k, j, and n. This leads to the following further notation:

— $\hat{\theta}^{(n)}$ is the estimate of parameter vector $\theta^{(n)}$;
— $\hat{\mu}_{kj}^{(n)} = g_\mu\left(x_{kj}^{(n)}, \hat{\theta}^{(n)}\right)$ is the fitted value associated with $Y_{kj}^{(n)}$;
— $\hat{\sigma}_{kj}^{2(n)} = g_{\rho^2}\left(x_{kj}^{(n)}, \hat{\theta}^{(n)}\right)$ is the estimate of $\mathrm{Var}\left[Y_{kj}^{(n)}\right]$;
— $R_{kj}^{(n)} = \left(Y_{kj}^{(n)} - \hat{\mu}_{kj}^{(n)}\right)/\hat{\sigma}_{kj}^{(n)}$ is the standardised Pearson residual associated with $Y_{kj}^{(n)}$;
— $R^{(n)} = \left\{ R_{kj}^{(n)} : k = 1, 2, \ldots, J;\ j = 1, 2, \ldots, J-k+1 \right\}$ is the triangle of standardised Pearson residuals for LoB n.

4. Measurement of dependency between lines of business

In this section, we define the measures of dependence we will use illustrate our arguments in the remainder of the paper.

4.1. Unweighted correlation

Let

$$\bar{R}^{(n)} = T^{-1} \sum_{k,\, j \in \Delta} R_{kj}^{(n)} \tag{4.1}$$

$$S^{(n)} = \left\{ T^{-1} \sum_{k,\, j \in \Delta} \left[R_{kj}^{(n)} - \bar{R}^{(n)} \right]^2 \right\}^{\frac{1}{2}} \tag{4.2}$$

be the mean and standard deviation[2] of the triangle of residuals $R^{(n)}$. Now, let us define the correlation coefficient of residuals across LoBs n_1 and n_2 as

$$r^{(n_1,\, n_2)} = \frac{T^{-1} \sum_{k,\, j \in \Delta} \left[\left(R_{kj}^{(n_1)} - \bar{R}^{(n_1)} \right) \left(R_{kj}^{(n_2)} - \bar{R}^{(n_2)} \right) \right]}{S^{(n_1)} S^{(n_2)}}. \tag{4.3}$$

This is the correlation between cells of different LoBs, but in the same positions within triangles $R^{(n_1)}$and $R^{(n_2)}$. If observations in $Y^{(n_2)}$ tend to be high (or low) when the corresponding observations in $Y^{(n_1)}$ are high (or low), then a high correlation $r^{(n_1,n_2)}$ will be returned. This correlation will therefore be accepted (subject to section 4.2 below) as an indicator of dependency between LoBs n_1 and n_2. This is also consistent with current practice (for risk margins), as described in Section 2.

4.2. Weighted correlation

The correlation coefficient (4.3) is the conventional one. All cells (k, j) are equally weighted. This will usually be satisfactory for LoBs with extended development, but less so when development is short.

For example, a characteristic of the Home insurance LoB considered later in Section 6.1 is that the majority of an accident quarter's claim cost is paid in the first three development quarters, and so any measure of dependency of this and another LoB should rely largely on the residuals associated with that limited range of development quarters. Correlation between the LoBs for these quarters might be small but, for some reason, larger for the higher development quarters, which are of little financial significance. Retention of (4.3) could then yield a misleadingly high value.

For this reason, a weighted form of correlation coefficient is introduced, defined as follows:

$$r^{w(n_1,\, n_2)} = \frac{\sum_{k,\, j \in \Delta} w_{kj}^{(n_1,\, n_2)} \left[\left(R_{kj}^{(n_1)} - \bar{R}^{w(n_1)} \right) \left(R_{kj}^{(n_2)} - \bar{R}^{w(n_2)} \right) \right]}{S^{w(n_1)} S^{w(n_2)}}, \tag{4.4}$$

where the weights are

$$w_{kj}^{(n_1,\, n_2)} = \frac{\left(\hat{\mu}_{kj}^{(n_1)} \hat{\mu}_{kj}^{(n_2)} \right)^{\frac{1}{2}}}{\left(\sum_{k,\, j \in \Delta} \hat{\mu}_{kj}^{(n_1)} \right)^{\frac{1}{2}} \left(\sum_{k,\, j \in \Delta} \hat{\mu}_{kj}^{(n_2)} \right)^{\frac{1}{2}}}, \tag{4.5}$$

and where the sample means and standard deviations are also weighted accordingly:

$$\bar{R}^{w(n)} = \frac{\sum_{k,\, j \in \Delta} \hat{\mu}_{kj}^{(n)} R_{kj}^{(n)}}{\sum_{k,\, j \in \Delta} \hat{\mu}_{kj}^{(n)}}, \text{ and} \tag{4.6}$$

$$S^{w(n)} = \left\{ \frac{\sum_{k,\, j \in \Delta} \hat{\mu}_{kj}^{(n)} \left[R_{kj}^{(n)} - \bar{R}^{w(n)} \right]^2}{\sum_{k,\, j \in \Delta} \hat{\mu}_{kj}^{(n)}} \right\}^{\frac{1}{2}}. \tag{4.7}$$

[2]These measures include slight biases. For example, some models, including the chain ladder, generate identically zero residuals in the top right and bottom left corners of the triangle, in which case $(T - 2)$ would be a preferable divisor in (4.1). Even in the absence of this effect, $S^{(n)}$ is the biased form of standard deviation estimator. Any downward bias in $S^{(n)}$ will cause a slight upward bias in any correlations in whose denominators it appears. However, these effects are small, and considered inconsequential for the data analysis that follows.

An equivalent form of (4.4) is

$$r^{w(n_1,\,n_2)} = \frac{\sum_{k,\,j\in\Delta}\left(\hat{\mu}_{kj}^{(n_1)}\hat{\mu}_{kj}^{(n_2)}\right)^{\frac{1}{2}}\left[\left(R_{kj}^{(n_1)}-\bar{R}^{w(n_1)}\right)\left(R_{kj}^{(n_2)}-\bar{R}^{w(n_2)}\right)\right]}{\left\{\sum_{k,\,j\in\Delta}\hat{\mu}_{kj}^{(n_1)}\left[R_{kj}^{(n_1)}-\bar{R}^{w(n_1)}\right]^2\right\}^{\frac{1}{2}}\left\{\sum_{k,\,j\in\Delta}\hat{\mu}_{kj}^{(n_2)}\left[R_{kj}^{(n_2)}-\bar{R}^{w(n_2)}\right]^2\right\}^{\frac{1}{2}}}. \tag{4.8}$$

The correlation coefficient (4.3) arises as a special case of (4.4) when all fitted values are equal within each triangle, i.e. $\hat{\mu}_{kj}^{(n)}$ is constant over k, j, and n.

It may be checked by an application of the Cauchy-Schwarz inequality that (4.8) that $-1 \le r^{w(n_1,\,n_2)} \le +1$. However, it is noteworthy that a requirement of $r^{w(n_1,\,n_2)} = 1$ is equivalent to requiring that the quantities $\left(\hat{\mu}_{kj}^{(n_1)}\right)^{1/2}\left[R_{kj}^{(n_1)}-\bar{R}^{w(n_1)}\right]$ and $\left(\hat{\mu}_{kj}^{(n_2)}\right)^{1/2}\left[R_{kj}^{(n_2)}-\bar{R}^{w(n_2)}\right]$ to be perfectly correlated. Thus, even when $R_{kj}^{(n_1)}-\bar{R}^{w(n_1)} = R_{kj}^{(n_2)}-\bar{R}^{w(n_2)}$ for all k, j, one will find $r^{w(n_1,\,n_2)} < 1$ unless the $\hat{\mu}_{kj}^{(n_1)}$ are a constant multiple of $\hat{\mu}_{kj}^{(n_2)}$ over k, j.

In view of these remarks, it must be said that the weighted form of correlation has no particular theoretical foundation, and is no more than a heuristic device whose objective is to address situations of the type described at the beginning of the present sub-section in which an unweighted (standard) correlation might be statistically correct, but financially misleading. As it happens, weighted and unweighted versions of correlation are not dramatically different in the numerical results reported in Section 6.

5. The role of modelling in forecasting the estimation of cross-LoB dependency

In this section, we discuss and decompose the effects of modelling on the estimation of dependencies. Systematic effects, if not appropriately captured, can make dependencies appear higher than they are; see Section 5.1. Furthermore, a careful decomposition of errors into three components permits us to unravel the effects of such errors on apparent correlation when considering one (Section 5.2) or multiple (Section 5.3) lines of business.

5.1. Modelling of systematic effects

It is an elementary principle that all time series should be de-trended before estimation of correlation between them (e.g., Shumway and Stoffer, 2011). Otherwise, one risks estimating high (or low) correlation simply because the subject series trend in the same direction (or opposite directions). There are numerous examples to illustrate this. A simple, though absurd, example follows in Figure 5.1, obtained from Vigen (2015). The correlation coefficient calculated directly from the plotted observations is 0.947. Although this is extremely high, one would do best to display caution before inferring any causal relationship between cheese consumption and death by bedsheet tangling. The fact is that both time series increase steadily over time, probably in consequence of some third factor, unseen here.

For many applications, correlation between two variates relates to the **stochastic noise** contained in them. These are the components of the variates that are, by their nature, not capable of being modelled. In such cases, one should model all systematic (non-stochastic) effects that are identifiable in the observations, remove these effects, and correlate the remainders of the observations, i.e. the residuals, as in Section 4.

The appearance of sympathetic trends in the subject variates is a simple and obvious feature that requires removal before estimation of correlation. However, much subtler and insidious examples are possible. Section 6.1 describes a case in which the paid losses of two respective LoBs are sympathetically affected by a less obvious third variate.

5.2. Prediction of a single line of business

5.2.1. Prediction error

The end purpose of modelling is often the forecast of future observations. This is certainly the nature of loss reserving and assessment of the uncertainty associated with a loss reserve is concerned with the properties of the prediction error of the model that provided the loss reserve forecast.

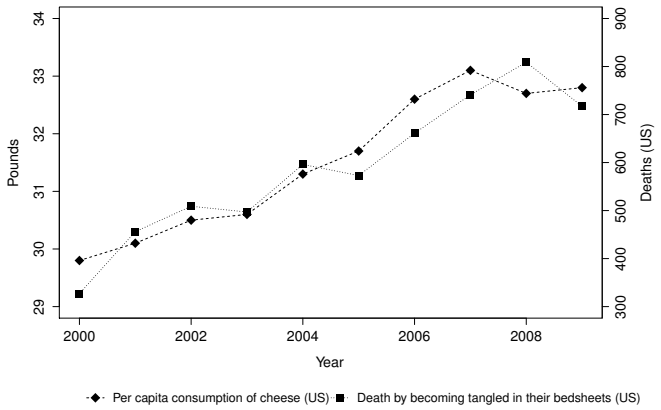


Figure 5.1: Example of spurious correlation (Vigen, 2015)

Where the loss reserve relates to multiple LoBs, the issue of dependency between those LoBs will arise. Hence one will be concerned with the modelling of such dependencies and its impact on prediction error.

Consider the application of the model described in Section 3 to forecasting. The observations requiring forecast will be $Y_{kj}^{(n)}$ for some **future cells** $(k,\, j)\,,\ k+j > J+1$. The means of these cells will be taken as

$$\mu_{kj}^{\mathrm{mod}(n)} := \mathbb{E}\left[Y_{kj}^{(n)}\right] = g_\mu\left(x_{kj}^{(n)},\, \theta^{(n)}\right), \tag{5.1}$$

and their forecasts as

$$\hat{\mu}_{kj}^{(n)} = g_\mu\left(\hat{x}_{kj}^{(n)},\, \hat{\theta}^{(n)}\right), \tag{5.2}$$

just as for fitted values associated with past cells, but where the $\hat{x}_{kj}^{(n)}$ are forecasts of the covariates $x_{kj}^{(n)}$ for the future cells $(k,\, j)$. The forecast of these quantities will be discussed later in Section 5.4.

Linearise the array of these future observations as a vector designated $Y^{*(n)}$, and construct similar vectors $\mu^{*\mathrm{mod}(n)}$ and $\hat{\mu}^{*(n)}$. The prediction error associated with the forecast $Y^{*(n)}$ is

$$\left[Y^{*(n)} - \mu^{*\mathrm{mod}(n)}\right] + \left[\mu^{*\mathrm{mod}(n)} - \hat{\mu}^{*(n)}\right] \tag{5.3}$$

It is advisable at this point to recognise that the model $g_\mu\,(.)$ that has been used to model the data and produce forecasts of future observations will almost certainly be an over-simplified representation of reality, and that the true means of those future observations will be $\mu^{*true(n)}$ say rather than $\mu^{*\mathrm{mod}(n)}$. Then the decomposition (5.3) can be extended as follows:

$$e^{(n)} = \underbrace{\left[Y^{*(n)} - \mu^{*\mathrm{true}(n)}\right]}_{\text{Process Error}} + \underbrace{\left[\mu^{*\mathrm{mod}(n)} - \hat{\mu}^{*(n)}\right]}_{\text{Parameter Error}} + \underbrace{\left[\mu^{*\mathrm{true}(n)} - \mu^{*\mathrm{mod}(n)}\right]}_{\text{Model Error}} \tag{5.4}$$

This may be conveniently written in the form:

$$e^{(n)} = e_{\mathrm{process}}^{(n)} + e_{\mathrm{parameter}}^{(n)} + e_{\mathrm{model}}^{(n)} \tag{5.5}$$

where the components on the right label those in (5.4) in the obvious way. The components of the decomposition have been labelled:

Process error: This is the component of forecast error arising from the random noise in the process. It is supposed that the observations are realisations of a random process and that, even if one were able to forecast the true mean accurately, the relevant observation would depart from it, simply as a result of its randomness.

Parameter error: This is the component of forecast error arising from error in calibration of the chosen $g_\mu(.)$. This occurs as a result of the limited sample size of the data set. The parameter estimates are functions of the data, and therefore random variables, which differ from the values they would take in the presence of an indefinitely large sample.

Model error: This is the component of forecast error arising from any inaccuracy in the choice of model.

The process error arises entirely from future observations, whereas the parameter error arises entirely from past observations. It is common for models to assume stochastic independence between past and future observations (possibly between all observations), in which case process and parameter error are stochastically independent. Since the model is selected before the occurrence of future observations, it cannot be influenced by them. If past and future observations are assumed independent, as above, then model error and process error must be stochastically independent. It then follows from (5.5) that the **mean square error of prediction ("MSEP")** of the forecast $\hat{\mu}^{*(n)}$ is given by

$$\begin{aligned} \text{MSEP}\left[\hat{\mu}^{*(n)}\right] &= \text{Var}\left[e^{(n)}\right] \\ &= \text{Var}\left[e^{(n)}_{\text{process}}\right] + \text{Var}\left[e^{(n)}_{\text{parameter}}\right] + \text{Var}\left[e^{(n)}_{\text{model}}\right] + \text{Cov}\left[e^{(n)}_{\text{parameter}}, e^{(n)}_{\text{model}}\right] \end{aligned} \tag{5.6}$$

The question of dependence between model error and parameter error is more difficult. It is possible that past observations have affected model selection, as well as its parameter estimation (and therefore parameter error). However, the linkage will be unclear if the model selection is subjective. It is often reasoned that this linkage is tenuous, and that the assumption of approximate stochastic independence between model error and parameter error is therefore reasonable. In this case, all three components in (5.5) become approximately stochastically independent, and so the MSEP of $\hat{\mu}^{*(n)}$ is given approximately by

$$\text{MSEP}\left[\hat{\mu}^{*(n)}\right] = \text{Var}\left[e^{(n)}_{\text{process}}\right] + \text{Var}\left[e^{(n)}_{\text{parameter}}\right] + \text{Var}\left[e^{(n)}_{\text{model}}\right]. \tag{5.7}$$

We assume here unbiasedness of the estimators (so that MSEP is the same as variance).

5.2.2. Impact of the omission of certain explanatory variables

Consider the model of Section 3, expressed in the form:

$$Y^{(n)} = g_\mu\left(x^{(n)}, \theta^{(n)}\right) + \varepsilon^{(n)} \tag{5.8}$$

where $x^{(n)}$ is the vector with components $x^{(n)}_{kj}$, and $\varepsilon^{(n)}$ is a centered (zero mean) random error term. Suppose that the parameter vector $\theta^{(n)}$ may be decomposed as follows:

$$\theta^{(n)} = \left(\phi^{(n)}, \xi\right), \tag{5.9}$$

where ξ is a sub-vector that does not depend on n, e.g. a set of inflation, or superimposed inflation, parameters that is common to the models of all LoBs.

Consider now a diminished model that omits ξ from its parameter vector, and naturally also omits the associated covariates within $x^{(n)}$. Express this model in the form:

$$Y^{(n)} = g_{\mu(-)}\left(x^{(n)}_{(-)}, \phi^{(n)}\right) + \eta^{(n)} \tag{5.10}$$

for some new centered random error term $\eta^{(n)}$. From (5.8) and (5.10),

$$\eta^{(n)} = g_\mu\left(x^{(n)},\, \phi^{(n)},\, \xi\right) - g_{\mu(-)}\left(x^{(n)}_{(-)},\, \phi^{(n)}\right) + \varepsilon^{(n)} \tag{5.11}$$

$$= G^{(n)}_{\mu(-)}\xi + \ldots(\text{higher order terms}) + \varepsilon^{(n)} \tag{5.12}$$

where the difference $g_\mu - g_{\mu(-)}$ has been expanded to first order in ξ, with the matrix $G^{(n)}_{\mu(-)}$ denoting the functional derivative of $g_{\mu(-)}$ in the direction of $g_\mu - g_{\mu(-)}$. Working to first order only now yields

$$\eta^{(n)} \approx G^{(n)}_{\mu(-)}\xi + \varepsilon^{(n)}. \tag{5.13}$$

It is seen that the move from model (5.8) to the inferior form (5.10) has the effect of shifting a part of the signal that can be modelled, and is therefore predictable, into the term that is regarded as just noise. In engineering terms, the signal-to-noise ratio has been reduced.

In prediction terms, it will change the model error in (5.4) from

$$\mu^{*\text{true}(n)} - g_{\mu(-)}\left(x^{(n)}_{(-)},\, \phi^{(n)}\right) \tag{5.14}$$

to

$$\mu^{*\text{true}(n)} - g_{\mu(-)}\left(x^{(n)}_{(-)},\, \phi^{(n)}\right) + G^{(n)}_{\mu(-)}\xi. \tag{5.15}$$

Model error has been augmented by the additional component $G^{(n)}_{\mu(-)}\xi$, and so will increase except in the highly serendipitous circumstance that this additional component happens to offset model error in the more complete model (5.8).

In practice, model error will rarely be measurable with any accuracy, and so a consideration of how poor modelling will manifest itself practically is worthwhile. The true mean will be inaccessible to the estimation of process error as set out in (5.4) and, as a practical measure, one might re-cast (5.4) in the slightly different form:

$$\underbrace{\left[Y^{*(n)} - \mu^{*\text{mod}(n)}\right]}_{\text{Process Error}} - \underbrace{\left[\hat{\mu}^{*(n)} - \mu^{*\text{mod}(n)}\right]}_{\text{Parameter Error}} + \underbrace{\left[\hat{\mu}^{*(n)} - \mu^{*\text{true}(n)}\right]}_{\text{Model Error}}. \tag{5.16}$$

In this case, future process error may be estimated from its analogue from the past, $Y^{(n)} - \mu^{\text{mod}(n)}$, i.e. from residuals such as $R^{(n)}_{kj}$. The shift from model (5.8) to the inferior form (5.10) now changes the process error by the term $G^{(n)}_{\mu(-)}\xi$, and the measured process error from the past is increased, and then the increase extrapolated into the future. That is, one's own assessment of forecast error is larger than need be. It will be based on goodness-of-fit to past data which, in this case, is poor on account of a poor model.

5.3. Prediction of multiple lines of business

We now extend the discussion developped in Section 5.2 to multiple lines of business. This allows us to describe neatly in mathematical terms how the omission of certain covariates leads to apparent correlation in Section 5.3.2.

5.3.1. Prediction Error

Consider the total forecast across all LoBs. Denote this

$$\hat{\mu}^{*(\bullet)} = \sum_{n=1}^{N} \hat{\mu}^{*(n)}. \tag{5.17}$$

The associated MSEP is

$$\text{MSEP}\left[\hat{\mu}^{*(\bullet)}\right] = \text{Var}\left[e^{(\bullet)}\right], \tag{5.18}$$

where, by (5.5),

$$e^{(\bullet)} = \sum_{n=1}^{N} e^{(n)} = \sum_{n=1}^{N} \left[e^{(n)}_{\text{process}} + e^{(n)}_{\text{parameter}} + e^{(n)}_{\text{model}} \right] \tag{5.19}$$

$$= e^{(\bullet)}_{\text{process}} + e^{(\bullet)}_{\text{parameter}} + e^{(\bullet)}_{\text{model}}. \tag{5.20}$$

Now, just as one argued for the stochastic independence of forecast error components in (5.5), one might equally argue for independence in (5.20), in which case (5.18) and (5.20) yield

$$\text{MSEP}\left[\hat{\mu}^{*(\bullet)}\right] = \text{Var}\left[e^{(\bullet)}_{\text{process}}\right] + \text{Var}\left[e^{(\bullet)}_{\text{parameter}}\right] + \text{Var}\left[e^{(\bullet)}_{\text{model}}\right] \tag{5.21}$$

It is now necessary to consider any dependencies between the sub-components of the components $e^{(\bullet)}_{\#}$, $\# = \{\text{process, parameter, or model}\}$. In general, (5.21) may be expanded as follows:

$$\text{MSEP}\left[\hat{\mu}^{*(\bullet)}\right] = \sum_{\#} \sum_{n_1,\, n_2=1}^{N} \rho^{(n_1 n_2)}_{\#} \sigma^{(n_1)}_{\#} \sigma^{(n_2)}_{\#}, \tag{5.22}$$

where

$$\sigma^{2(n)}_{\#} = \text{Var}\left[e^{(n)}_{\#}\right] \tag{5.23}$$

$$\rho^{(n_1 n_2)}_{\#} = \text{Corr}\left[e^{(n_1)}_{\#}, e^{(n_2)}_{\#}\right] \tag{5.24}$$

There is usually little reason to expect any dependency between the parameter errors of different LoBs, so one might assume

$$\rho^{(n_1 n_2)}_{\text{parameter}} = 0 \tag{5.25}$$

in (5.22), which produces some simplification.

However, there is not necessarily any reason to expect $\rho^{(n_1 n_2)}_{\text{process}} = 0$ or $\rho^{(n_1 n_2)}_{\text{model}} = 0$. Indeed, there will often be good reason to expect $\rho^{(n_1 n_2)}_{\text{model}} > 0$. If, for example, the legal environment were to switch from a state of quiescence to one of judicial activism, the effects of increased awards might be felt simultaneously across a number of Casualty LoBs. If industrial activity were to cause a sudden surge in wage levels, both Property and Casualty claim sizes might escalate. The situation is illustrated in Figure 5.2.

Thus the dependency issues surrounding the prediction error associated with an aggregate forecast across a number of LoBs reduces to the estimation of $\rho^{(n_1 n_2)}_{\text{process}}$ and $\rho^{(n_1 n_2)}_{\text{model}}$. The former of these is, in principle, estimable from past observations, but the latter, and even estimation of $\sigma^{2(n)}_{\text{model}}$, is problematic.

Model error is essentially unobservable from past observations. If errors in the adopted model were identifiable, then the model could be modified and the errors modelled away. The estimation of any properties of model error therefore require a methodology fundamentally different from that applied to process error (see, e.g. O'Dowd et al., 2005).

For this reason, the quantification of model error, and its associated dependencies with respect to LoB will be left for future research, and attention will be concentrated on process error.

5.3.2. Process error

Consider once again the model (5.8), and its inferior form (5.10). Section 5.2.2 investigated the effect of the inferior model on the prediction of each LoB in isolation. There are, however, additional effects when multiple LoBs are considered.

Consider the hypothetical situation in which there are no stochastic dependencies between LoBs. In this case,

$$\text{Cov}\left[\varepsilon^{(n_1)}, \varepsilon^{(n_2)}\right] = 0 \tag{5.26}$$

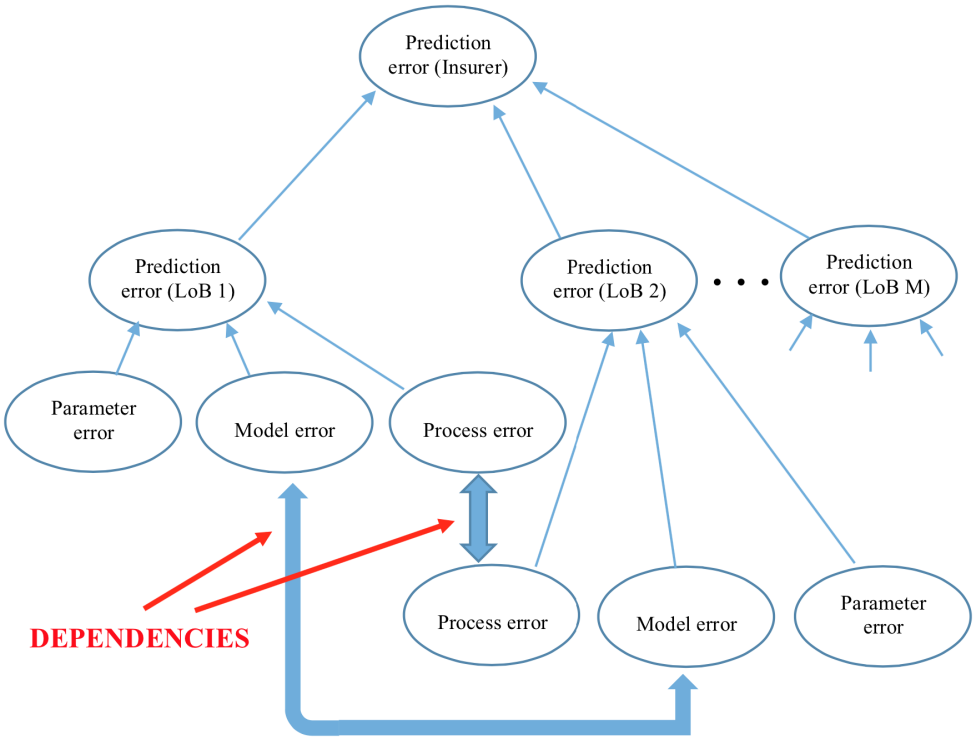


Figure 5.2: Structure of dependencies between components of prediction error

According to the formulation of the more complete model (5.8), the quantity $\zeta^{(n)}$ is partly stochastic $\left(\varepsilon^{(n)}\right)$, and partly deterministic $\left(G^{(n)}_{\mu(-)}\xi\right)$. However, (5.10) views the entirety of $\zeta^{(n)}$ as a stochastic error term. So, for two LoBs, the terms $\zeta^{(n_1)}, \zeta^{(n_2)}$ will contain contributions from the common influence ξ.

If, for example, $G^{(n)}_{\mu(-)}$ is regarded as deterministic, and ξ as stochastic, within this model, then

$$\text{Cov}\left[\zeta^{(n_1)}, \zeta^{(n_2)}\right] = G^{(n_1)}_{\mu(-)} \text{Var}\left[\xi\right] \left[G^{(n_2)}_{\mu(-)}\right]^{\text{T}} \neq 0 \tag{5.27}$$

where the upper 'T' denotes matrix transposition.

A comparison of (5.26) with (5.27) shows that poor modelling that omits influential covariates from the modelling of individual LoBs can create the appearance of dependency between those LoBs where there is in fact none. Since the source of the dependency between $\zeta^{(n_1)}$ and $\zeta^{(n_2)}$ arises from the fact that each is a linear transformation of ξ, it can be considered as akin to a **common shock** (Lindskog and McNeil, 2003).

It was remarked in Section 5.2.1 that the omission of influential covariates from a model design has the effect of changing model error. This is but one form of model mis-specification, of which many other forms are possible. These include mis-specification of the response variate's distribution, of the algebraic form of its relation with the covariates, etc. Any of these might result in biased parameter estimates, mis-estimation of dispersion or dependencies, or other model properties. The present paper is concerned with just the effects on correlation of the specific form of mis-specification that omits influential covariates from a model.

5.4. Covariate selection for forecasting

The inclusion in a model of additional explanatory variables with predictive power will improve the model. However, whether or not this leads to improved model forecasts depends on whether or not those variables are capable of accurate extrapolation into the future.

Taylor et al. (2008) classify explanatory variables (or covariates) into three categories; which are (with slight change of terminology):

1. **Static covariates**: These are variates whose values do not change over the life of a claim, e.g. vehicle make and model under an Auto insurance.
2. **Dynamic covariates**: These are variates whose values do change over the life of a claim. There are two sub-categories:
    1. **Deterministic**: These are variates which are directly related to calendar time, and whose future values are therefore predictable, e.g. development period.
    2. **Stochastic**: These are variates which change over time in a manner that is not predictable with certainty, e.g. case estimate (variously called the manual estimate or physical estimate) of ultimate incurred loss associated with a claim.

The inclusion of static or deterministic dynamic covariates in a forecast will be straightforward. The future values of stochastic dynamic covariates, on the other hand, will be unknown. The predictive power of the model to be used for forecasts may have been shown to be enhanced (i.e. prediction error reduced) in the event of inclusion of the stochastic dynamic covariates. However, this conclusion will rest on an assumption that accurate future values of those covariates are available.

Forecast of these values may be possible, but with forecast errors of their own, and the effects of these on the main forecasts must be taken into account. The issue may be considered analytically as follows. Consider once again the more and less complete models (5.8) and (5.10) respectively, but this time suppose that the covariates in $x^{(n)}$ that are omitted from $x_{(-)}^{(n)}$ are stochastic dynamic covariates. Suppose further that forecasts $\hat{x}^{*(n)}$ of the future values of $x^{*(n)}$ (including the stochastic dynamic covariates) have prediction error $\omega^{(n)}$ according to the following relation:

$$\hat{x}^{*(n)} = x^{*(n)} + \omega^{(n)} \tag{5.28}$$

One may now forecast on the basis of the either the more or the less complete model, i.e according to either Model (5.8) or Model (5.10), which we will now compare.

Model (5.10) yields a forecast

$$Y^{*(n)} = g_{\mu(-)}\left(x_{(-)}^{(n)}, \hat{\phi}^{(n)}\right) \qquad \text{by (5.10)} \tag{5.29}$$

with prediction error

$$\text{MSEP}\left[\zeta^{(n)}\right] = \text{Var}\left[G_{\mu(-)}^{(n)}\xi + \varepsilon^{*(n)}\right]. \tag{5.30}$$

Alternatively, Model (5.8)—the more complete model—yields the forecast

$$Y^{*(n)} = g_{\mu}\left(x^{*(n)}, \hat{\theta}^{(n)}\right) \tag{5.31}$$

with prediction error

$$\text{MSEP}\left[\zeta^{(n)}\right] = \text{Var}\left[\left[g_{\mu}\left(\hat{x}^{*(n)}, \hat{\theta}^{(n)}\right) - g_{\mu}\left(x^{*(n)}, \hat{\theta}^{(n)}\right)\right] + \varepsilon^{*(n)}\right] \tag{5.32}$$

$$= \text{Var}\left[\left[g_{\mu}\left(x^{*(n)} + \omega^{(n)}, \hat{\theta}^{(n)}\right) - g_{\mu}\left(x^{*(n)}, \hat{\theta}^{(n)}\right)\right] + \varepsilon^{*(n)}\right]. \tag{5.33}$$

by (5.28). Comparison of the contending MSEPs (5.30) and (5.33) is inconclusive. Each consists of the variance of a shifted version of the centered error $\varepsilon^{*(n)}$. The relative magnitudes of the MSEPs will depend on those of the respective shifts.

The inclusion of the stochastic dynamic covariates in the model, augmenting the covariate vector from $x_{(-)}^{*(n)}$to $x^{*(n)}$ eliminates the shift $G_{\mu(-)}^{(n)}\xi$ that occurs in (5.30), but introduces a new shift

$$g_{\mu}\left(x^{*(n)} + \omega^{(n)}, \hat{\theta}^{(n)}\right) - g_{\mu}\left(x^{*(n)}, \hat{\theta}^{(n)}\right), \tag{5.34}$$

as in (5.33).

Even if the additional covariates are highly predictive in the modelling of past observations, a high degree of uncertainty in the forecast of their future values may increase MSEP in such a way that it more than offsets the penalty incurred in the omission of these covariates from the forecast model.

5.5. Hypothetical example of spurious correlation

Sections 5.2.2 and 5.3 demonstrate how the omission of essential terms from models of paid loss triangles for multiple LoBs can generate spurious correlations between those LoBs. The present section illustrates this phenomenon numerically on the basis of simulated paid loss triangles.

A paid loss triangle is simulated for each of two LoBs, Home and Motor. In each case, observations are simulated according to an augmented form of the simple chain ladder model set out in (6.1)-(6.2), specifically

$$Y_{kj}^{(n)} \sim \text{Poisson}\left(\mu_{kj}^{(n)}, \phi^{(n)}\right), \qquad n = 1, 2; \tag{5.35}$$

$$\mu_{kj}^{(n)} = \exp\left\{r_k^{(n)} + s_j^{(n)} + t_{k+j-1}^{(n)}\right\}, \tag{5.36}$$

where the term $t_{k+j-1}^{(n)}$ relates to the $k+j-1$-th diagonal, containing the (k, j) cell. This additional model term affects all cells in a diagonally uniformly. It therefore resembles an inflationary effect. In fact,

$$t_{k+j-1}^{(n)} = (k+j-1)t^{(n)} \tag{5.37}$$

would be equivalent to an inflation rate of $\exp\left(t^{(n)}\right) - 1$ per period. Equation (5.36) is the age-period-cohort model, containing row, column and diagonal effects, as discussed in Kuang et al. (2008b,a, 2009).

The paid loss triangles are subject to $J = 41$, which is in fact the dimension of those that will appear in Section 6.2. The values of $r_k^{(n)}, s_j^{(n)}, \phi^{(n)}$ are also those estimated by application of the simple chain ladder model (6.1)-(6.2) to the Home and Motor data sets used in the same section. Three separate scenarios are assumed for the values of $t_{k+j-1}^{(n)}$:

— **Scenario 1:** paid loss inflation at the quarterly equivalent of an annual rate of 10% in diagonals 17 to 28, and 3% in other diagonals;

— **Scenario 2:** paid loss inflation at the quarterly equivalent of an annual rate of 3% in diagonals 1 to 20, and 10% thereafter;

— **Scenario 3:** paid loss inflation at the quarterly equivalent of an annual rate of 1% in diagonals 1 to 4, increasing by 1% in diagonals 5,9, etc., to attain 11% in diagonal 41;

Crucially, all observations in these triangles are generated as stochastically independent, whether within or between triangles. Any cross-LoB correlation should be measured as zero.

Thus, six paid loss triangles were created, three each for the Home and Motor LoBs. Simulation produced 1,000 replicates of each of the six triangles. Each of the 6,000 resulting triangles was modelled according to the simple chain ladder model (6.1)-(6.2), i.e. excluding the inflationary effects, and thereby inducing model error.

For each replicate within each scenario, the correlation between Home and Motor residuals was then calculated in accordance with (4.3) and (4.4). The results are set out in Table 5.4, where averages have been taken over the 1,000 replicates. It is seen that, for each scenario, the model error has induced a substantial cross-LoB correlation, despite the known stochastic independence between LoBs.

Scenario	Correlation between Home and Motor residuals	
	Unweighted	Weighted
1	0.20	0.27
2	0.27	0.32
3	0.17	0.25

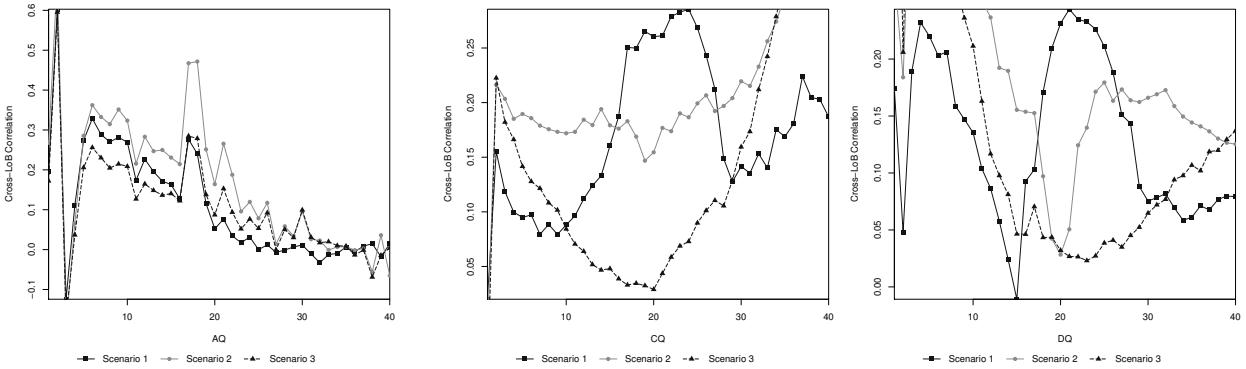
Table 5.4: Home-Motor correlation on the basis of a simulated chain ladder model with inflation

It is of interest to study this example further by decomposition of the measured correlations into more specific components. Thus, rather than calculate correlations of residuals, taken over all cells of the Home and Motor triangles, one might take residuals over just a specific row. Alternatively, over a specific column or diagonal.

Figures 5.3 (a) to 5.3 (c) plot the less dramatic unweighted correlations, taken over rows, columns and diagonals respectively. These figures illustrate how increased model error induces increased correlation.

Consider Scenario 3, for example. This is marked by inflation that steadily increases over diagonals. However, the model fitted ignores inflation. It is well known that, for the simple chain ladder, this is equivalent to assuming a constant inflation rate across diagonals. In effect, the model will be equivalent to one which assumes a constant inflation rate of the same order as the average actually included in the data, i.e. about 5-6% per annum.

This is, in fact, the rate that occurred over diagonals 17 to 24. Model error is therefore least in the region of the middle diagonals, and greatest in the regions of the lowest and highest diagonals. There is no surprise, therefore in the finding, evident in Figure 5.3 (c), that the falsely induced correlation of Scenario 3 is largely attributable to the highest and lowest diagonals. Other features of Figures 5.3 (a) to 5.3 (c) can be explained similarly.



5.3 (a) by AQ

5.3 (b) by CQ

5.3 (c) by DQ

Figure 5.3: Average Simulated Cross-LoB Correlation

6. Empirical investigation

6.1. An example of correlation illusion - Australian Home and Motor

The previous section explained how inadequate modelling could induce apparent correlation between LoBs where in fact none exists. Similarly, poor modelling might produce the illusion of high correlation, where some exists but only at a low level. The present sub-section provides an example of this phenomenon, using Australian Home and Motor data.

6.1.1. Data set

The present example investigates dependency between Home and Motor claim experiences, and so triangles of incremental paid claims have been extracted from the AUSI data set; refer to Appendix A.1. These LoBs are both short tailed. The great majority of claim cost is paid within 9 months of claim occurrence in the case of Home, and 2 years in the case of Motor. It is therefore beneficial to construct the claim triangles in terms of **accident and development quarters**.

The triangles prepared had dimension 40×40 quarters. They covered accident quarters from the 10-year period 2004 to 2013 inclusive. As mentioned above, the data have little significance beyond a limited number of development quarters.

6.1.2. Simple chain ladder model

Cross-LoB dependency is measured by means of the correlations (unweighted and weighted) defined in (4.3) and (4.4). These correlate residuals with respect to some model. Initially, the model is deliberately chosen to be simple, in fact a simple chain ladder. This will enable the examination of the effect on correlation of progressive model refinement.

In order to facilitate later refinements, the chain ladder model is defined in a fully stochastic form, technically the Generalized Linear Model representation of the over-dispersed Poisson (**"ODP"**) cross-classified chain ladder (England and Verrall, 1999; Mack and Venter, 2000; Taylor, 2011), as follows:

$$Y_{kj}^{(n)} \text{ is Poisson}(\mu_{kj}^{(n)}, \phi_{kj}^{(n)}), \tag{6.1}$$

where

$$\mu_{kj}^{(n)} = \alpha_k^{(n)} \beta_j^{(n)} = \exp\left\{ r_k^{(n)} + s_j^{(n)} \right\}, \text{ with } \left\{ \begin{array}{l} r_k^{(n)} = \ln \alpha_k^{(n)} \\ s_j^{(n)} = \ln \beta_j^{(n)} \end{array} \right. . \tag{6.2}$$

The $Y_{kj}^{(n)}$ are mutually stochastically independent (within a triangle), and $\phi_{kj}^{(n)}$ is the dispersion parameter associated with the ODP distribution. Relations (6.1) and (6.2) can be recognised as corresponding to the framework set out in Section 3, with the following correspondences:

$$\begin{array}{rcl} P & \leftrightarrow & \text{ODP;} \\ g_\mu & \leftrightarrow & \exp; \\ \theta^{(n)} & \leftrightarrow & [r_k^{(n)}, s_j^{(n)} : \text{all relevant } k, j]^{\mathrm{T}}; \\ x_{kj}^{(n)} & \leftrightarrow & \text{vector of binary variates that selects parameters } r_k^{(n)}, s_j^{(n)} \text{ from } \theta^{(n)}. \end{array}$$

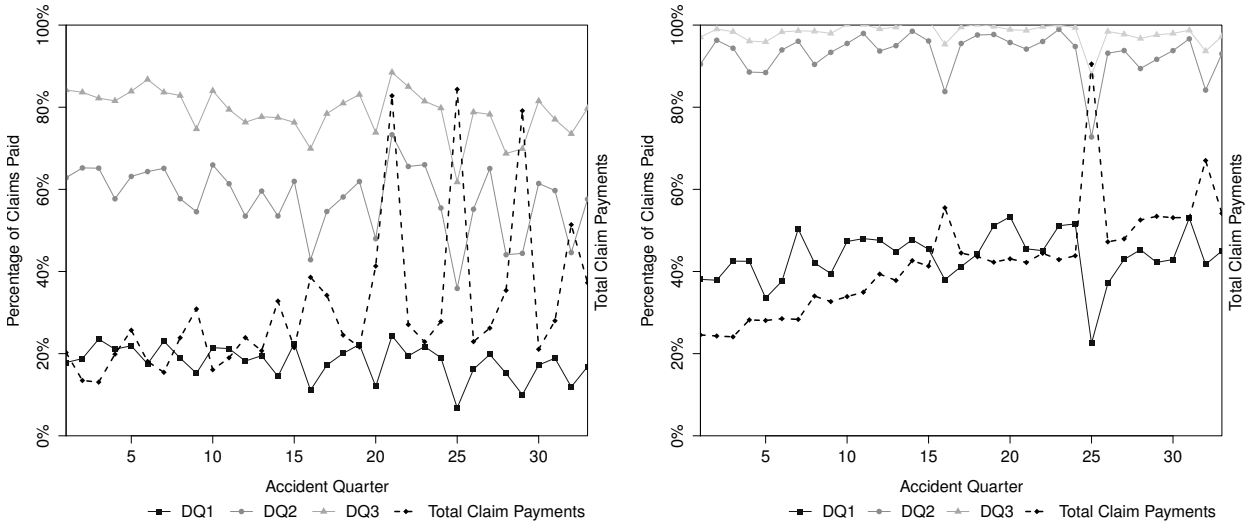
A model of this form was fitted to each of the Home and Motor triangles with $\phi_{kj}^{(n)} = \phi$, unknown but independent of k, j. Pearson residuals were calculated, as defined in Section 3, and these residuals correlated as defined in Section 4. The result appears in Table 6.5. Since the measured correlation is large, it is also re-calculated with its coverage restricted to the individual accident quarters that contribute most heavily to the total.

Accident quarters included	Correlation*	
	Unweighted	Weighted
2	0.72	0.72
$9-11$	> 0.31	> 0.78
16	0.78	0.73
$19, 21, 22$	> 0.44	> 0.71
28	0.52	0.77
30	0.39	0.82
32	0.97	0.91
All	**0.59**	**0.60**

Table 6.5: Home-Motor correlation on the basis of chain ladder model
*Lower bounds are shown when a row relates to more than one accident quarter.

6.1.3. The effect of major weather events

The briefest examination of the data indicates substantial differences between accident periods in the total volume of claims. Figure 6.4 (a) charts, for each of development quarters 1 to 3, the time series, by accident quarter, of the cumulative proportion paid of Home loss payments made in those quarters. Figure 6.4 (b) does the same for Motor claims. These charts also display, by accident quarter, the total of Home



6.4 (a) Home

6.4 (b) Motor

Figure 6.4: Volume of claims paid by accident quarter and effect on paid loss development

and Motor claim payments made in the first 8 development quarters[3]. These plots are characterized by occasional very marked peaks, which can be confidently attributed to major weather events. It may be noted that some peaks affect both LoBs (accident quarters 16, 25 and 32). These will have generated large positive residuals in both LoBs, increasing measured correlation. They will also have created a tendency in both LoBs toward negative residuals in most of the other accident quarters, increasing correlation further.

It is often the case that claims arising from major events are tagged thus, and can be eliminated from an analysis of attritional claims. This would be desirable in a study of correlation such as the present one. This is not to say that major events do not cause correlation between the claims experiences of different LoBs. They do, indeed. However, greater clarity will be achieved in correlation measurement if attritional and major event claims are analysed separately. Typically, the correlation of claims from major events will be estimated from a CAT model, as distinct from the above type of statistical modelling, which may be applied to the attritional claims.

The AUSI data base currently does not identify claims from major events. These are best dealt with by the elimination of the affected accident quarters (16, 21, 25, 29 and 32) from the modelling on which correlation estimation is based. When this is done, the correlation of 0.60 appearing in Table 6.5 is reduced to 0.14. This is a major reduction, but the revised correlation is still material, a matter that is pursued in the following sub-section.

6.1.4. Seasonal effects

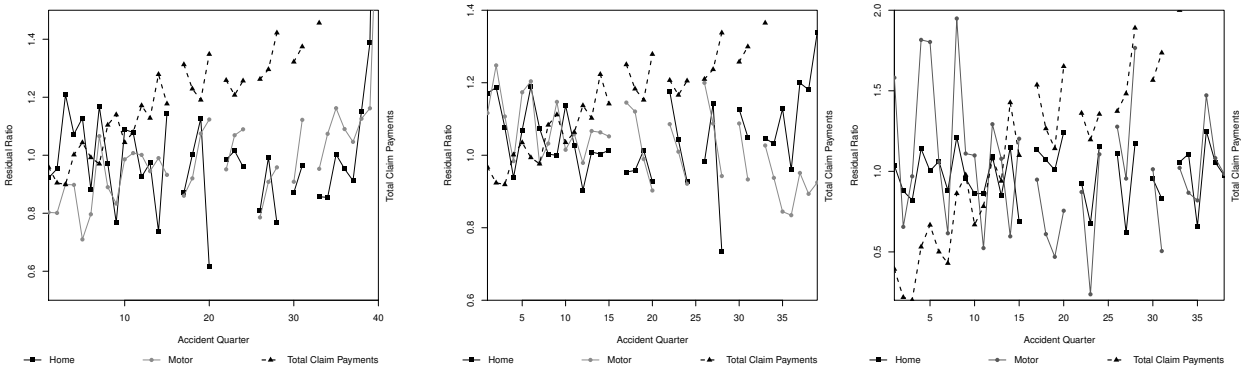
Figure 6.4 also suggests seasonal influences. There is a hint of slower payment of claims in every fourth accident quarter, specifically quarters numbered 4, 8, etc. It is also noticeable that these accident quarters are marked by generally higher claim payments.

This effect is somewhat confounded by the occurrence of major events, whose accident quarters are marked by especially high claim payments and slow rates of claim payment. It is evident that the claims of any particular accident quarter are likely to be paid more slowly (or rapidly) according as the volume of claims impacting the accident quarter is high (or low).

[3]Note that the relevant axis for this series of data, in $, would normally appear on the right hand side of the plots. We had to redact it for confidentiality reasons.

One might therefore infer a systematic effect according to which each fourth accident quarter (actually the fourth accident quarter of the calendar year, late spring and early summer) coincides with increased claims activity, probably weather related, and retarded processing of claims. In fact, a closer examination of the data indicates that a similar effect occurs in the first quarter of the calendar year (most of summer and early autumn).

Figure 6.5 illustrates the same point from a different perspective. The residual ratios of the Home and Motor LoBs are plotted when the accident quarters nominated as affected by major events are excluded from the modelling. Here, the residual ratio associated with cell (k, j) is defined as $Y_{kj}^{(n)}/\hat{\mu}_{kj}^{(n)}$, i.e. the ratio of actual to fitted value. The charts omit the accident quarters affected by major events. The sympathetic



6.5 (a) Development quarter 1

6.5 (b) Development quarter 2

6.5 (c) Development quarter 3

Figure 6.5: Residual ratios for Home and Motor

movement in the Home and Motor residual ratios in the "normal" accident quarters is evident, and this will clearly generate substantial correlations between the LoBs.

It is of interest to note that smaller peaks and troughs in the residual ratios still occur, with a hint that these are negatively related to variations in the "total claim payments" that also appear in the figures. These "total claim payments" are defined as the total of both Home and Motor payments in development quarters 1 to 8.

One might speculate on the possibility that a heavier load on the Domestic lines (Home and Motor) claims administration during the summer months, as appears to occur, might lead to slightly slower payment of claims. Thus, in modelling the (k, j) cell of the claims triangle, one might introduce a variate that distinguishes between some summer quarters (Q1 and Q4) and winter quarters (Q2 and Q3). Separate such indicators might be introduced for each of development quarters 1 to 3, since the seasonal effect might differ over these development quarters. On the basis of the evidence just cited, the Summer indicator would be expected to be associated with slower development of paid losses. A little investigation reveals that one should include separate seasonal effects as follows:

Home :

— Development quarter 1 for events with origin in calendar quarter 1;
— Development quarter 1 for events with origin in calendar quarter 4;
— Development quarter 2 for events with origin in calendar quarter 4.

Motor :

— Development quarter 1 for events with origin in summer (calendar quarters 1 and 4);
— Development quarter 2 for events with origin in calendar quarter 1;
— Development quarter 2 for events with origin in calendar quarter 4.

Note that for development quarter 1 of the Motor LoB, there is simply a "summer" effect, i.e. no distinction between Q1 and Q4 in terms of claims activity. On the other hand, for Development quarter 1 of the Home LoB, Q1 from summer is indistinguishable from the winter quarters. For development quarter 2 of either LoB, the Q1 and Q4 effects differ. The effects of the summer quarter have decayed to insignifiance by development quarter 3.

The seasonal effect, and the effect of major events discussed in Section 6.1.3, are taken into account by extending model (6.1)-(6.2) to the following:

$$\mu_{kj}^{(n)} = \exp\left\{ r_k^{(n)} + s_j^{(n)} + \sum_{j=1}^{2} \sum_i Y_{ji}^{(n)} I\left((t \text{ modulo } 4) \in Q_{ji}^{(n)} \right) \right\} \tag{6.3}$$

where $I(c) = 1$ if condition c is satisfied, and is equal to zero otherwise; where the $Q_{ji}^{(n)}$ are subsets of the set 1,4; and where the $\gamma_{ji}^{(n)}$, $j = 1, 2$, are coefficients to be estimated. In the present case, as an example, $Q_{11}^{(n)} = \{1, 4\}, Q_{21}^{(n)} = \{1\}, Q_{22}^{(n)} = \{4\}$ for the Motor LoB. The dispersion parameter $\phi_{kj}^{(n)}$ was set as follows:

$$\phi_{kj}^{(n)} = \begin{cases} \infty & \text{whenever } k \text{ is one of the accident quarters subject to a major event;} \\ \phi, & \text{unknown but independent of } k, j, \text{ otherwise.} \end{cases} \tag{6.4}$$

The setting of $\phi_{kj}^{(n)} = \infty$ amounts to an assumption of infinite variance in the (k, j) cell, and this is equivalent to exclusion of this cell from the data set in the model fitting.

When this model is fitted to the data, the correlation of **0.60** appearing in Table 6.5 (already reduced to 0.14 in Section 6.1.3) is reduced further to **-0.05**. All coefficients $\gamma_j^{(n)}$ were indeed negative, as expected on the basis of the exploration of the data in Figure 6.4 to Figure 6.5.

This modelling, and the associated reduction in correlation, is of particular interest. The effect modelled, rate of claims settlement, is of precisely the form countenanced in the theoretical argument leading to (5.27).

Changing the rate of settlement changes the distribution of an accident quarter's total claim cost over its development quarters, but does not change the total cost itself. The costs of a specific accident quarter in Home and Motor respectively may be of zero or low correlation, but failure to recognize and model the rate of settlement effect will induce a substantial increase in **apparent** correlation.

6.1.5. Summary of results

Table 6.6 summarizes the results derived in Sections 6.1.2 to 6.1.4. As one works down the table, the completeness of the modelling increases, and is seen to be accompanied by a steady reduction in measured Home-Motor Pearson correlation.

Model features	Measured correlation		
	Pearson correlation		Rank
	Unweighted	Weighted	correlation
Simple chain ladder	+0.59	+0.60	+0.30
Effects of major events excluded	+0.11	+0.14	+0.10
Controlled for rate of settlement	−0.01	−0.05	+0.01

Table 6.6: Home-Motor correlation on the basis of chain ladder model

Indeed, after both augmentations of the chain ladder model, the unweighted form of correlation is zero to two decimal places, and the weighted form actually turns negative. The conclusion appears to be that there is no inherent correlation between the Home and Motor claims experiences, but that major weather events and seasonal changes in claim processing volumes produce apparent correlation when not accounted for in the modelling of these LoBs.

As has been explained earlier, Pearson correlation is expected to be a satisfactory measure of cross-LoB dependency for the purpose of estimating VaRs at moderate percentiles. However, just in case this form

of correlation might be misleading for some unforeseen reason, Table 6.6 also includes rank correlation, a non-parametric measure of dependency. No attempt has been made to produce a weighted form of rank correlation.

Although the unmodified chain ladder generates more muted rank correlation (+0.30) than its Pearson counterpart, the effects of the additional modeling on the two forms are similar: a steady reduction as the model is expanded to recognise effects that are extraneous to the chain ladder, and a final rank correlation that is close to zero when both corrections to the chain ladder are implemeted.

6.2. Other cross-LoB dependencies

6.2.1. Australian evidence

The AUSI data set described in Appendix A.1 includes the following LoBs:

Insurer A :

— Home;
— Motor;
— CTP;
— Public Liability;

Insurer B :

— CTP;
— Public Liability.

The analysis of Insurer A Home and Motor is described in Section 6.1. As noted in Section 6.1.2, the analysis commences with simple chain ladder modelling, and was then modified for a couple of specific effects. Simple chain ladder modelling was carried out in relation to the other two LoBs listed just above, and cross-LoB correlation calculated for each within-insurer pair.

No pair produced a sample correlation numerically greater than 0.10, even in the presence of this simple modelling. Table 6.7 displays the matrix of within-insurer correlations (PL = Public Liability). Only weighted correlations are given here; the unweighted version is similar. The evidence for any correlation between LoBs, positive or negative, is extremely weak.

		Weighted correlation					
		Insurer A				Insurer B	
		Home	Motor	CTP	PL	CTP	PL
Insurer A	Home	1	−0.05	+0.02	+0.05		
	Motor		1	+0.01	+0.01		
	CTP			1	−0.01		
	PL				1		
Insurer B	CTP					1	−0.08
	PL						1

Table 6.7: Cross-LoB correlations on the basis of chain ladder model (modified in the cases of Home and Motor)

6.2.2. US evidence

Only the first three and the last of the six listed LoBs of the Meyers-Shi dataset have been used for the current project; refer to Appendix A.2. Of those, Insurer/LoB combinations have been discarded if premium has changed (upward or downward) by a factor of more than 4 over the 10-year period 1988-1997. This filters out entries and exits to the market, as well as other cases of eccentric growth or decline. The end result of the filtering is as in Table 6.8.

LoB	Number of insurers
PPA	63
CA	68
WC	34
OL	103

Table 6.8: Numbers of insurers included in study

Some triangles (118 in number) contain isolated negative incremental paid losses. This is problematic for the form of chain ladder modelling by GLM described in Section 4.1.2.1, since a Poisson variate is non-negative. For this reason, any negative entry has been replaced by a zero.

The effect of this has been checked by calculating age-to-age factors from unadjusted data, using conventional chain ladder calculations, and comparing them with the corresponding age-to-age factors obtained from the chain ladder GLM applied to the adjusted data. It was found that, of the age-to-age factors affected, just under half were affected by less than 0.01%, and just under two-thirds by less than 0.1%. Less than 4% of the factors were affected by more than 5%. In summary, the effect of the data adjustment on the results of the modelling was minor.

Each of the paid loss triangles qualifying for inclusion in the study has been subjected to the simple chain ladder analysis described in Section 6.1.2. All cross-LoB correlations have then been calculated as described in the same section. Not all insurers underwrite all LoBs. The number of insurers contributing to each LoB pair was as in Table 6.9.

LoB pair	Number of insurers
PPA-CA	42
PPA-WC	9
PPA-OL	33
CA-WC	16
CA-OL	37
WC-OL	13

Table 6.9: Numbers of insurers included in study by LoB pair

For each LoB pair, correlation is calculated separately for each insurer, and the averages of these correlations are set out in Table 6.10. The triangles included in the Meyer-Shi data set are generally considerably longer tailed than those in the AUSI data set, and so the difference between weighted and unweighted correlations is less of an issue. Nonetheless, both have been calculated. There is little difference between them and, for brevity, only the unweighted version is given in the table.

	Unweighted correlation			
	PPA	CA	WC	OL
PPA	1	+0.07	+0.01	+0.06
CA		1	+0.08	+0.00
WC			1	+0.02
OL				1

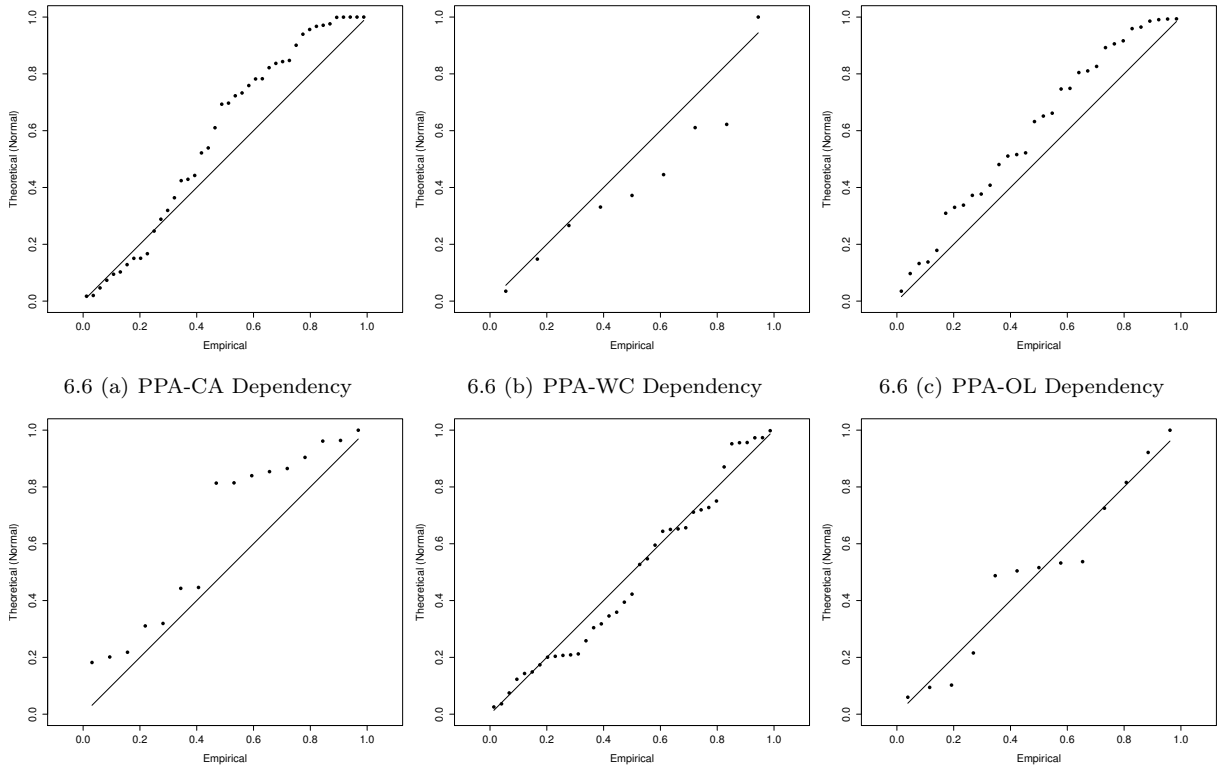
Table 6.10: Average Cross-LoB correlations from Meyers-Shi data set

As Table 6.10 contains only summary statistics from the sample of correlations, it is interesting to examine the distributional properties of the sample. Figure 6.6 therefore displays P-P plots for the six respective LoB pairs. Each of these plots the "theoretical" cumulative d.f. read-outs of the Fisher transforms of the sample of correlations for the pair against the empirical cumulative d.f. The Fisher transform is defined in

the usual way, as

$$z = \frac{1}{2} \ln \frac{1+r}{1-r} \tag{6.5}$$

for measured correlation r, and the "theoretical" d.f. is $\mathrm{N}(0, (q-3)^{-1/2})$ for sample size q. The P-P plot examines consistency of the sample of observed correlation coefficients with the hypothesis that their population mean is zero.



6.6 (a) PPA-CA Dependency

6.6 (b) PPA-WC Dependency

6.6 (c) PPA-OL Dependency

6.6 (d) CA-WC Dependency

6.6 (e) CA-OL Dependency

6.6 (f) WC-OL Dependency

Figure 6.6: PP-plots of cross-LoB Fisher transforms

An empirical curve lying above the diagonal indicates a tendency for the population mean to exceed zero. One might draw the following conclusions from Figure 6.6:

A. 3 LoB (PPA-WC, CA-OL, WC-OL) pairs give no indication of non-zero correlation;

B. The other 3 (PPA-CA, PPA-OL, CA-WC) indicate a tendency to positive correlation, but only slight.

These conclusions are consistent with Table 6.10.

It should be noted that these results derive from the simplest possible modelling of the data, mere application of the simple chain ladder to each claim triangle. One might observe, on the basis of the findings of Section 6.1, that the observed correlations, already low, might be further reduced by more refined modelling. In short, the data indicate low cross-LoB dependency for the LoBs studied.

7. Measurement of past dependencies: implications for the future

Section 6 has demonstrated an absence of empirical correlations across LoBs in the past data of a couple of specific cases. It is important, however, to observe that this does **not** necessarily imply that one should assume zero correlation between the same LoBs in all forecasts. A simple hypothetical example of why this would be erroneous in this assumption might be useful.

Consider the special case of the model (6.1)-(6.2) in which $\alpha_k^{(n)} = \alpha_k$, independent of n, i.e. the row effects of different triangles are identical. Suppose there are no other dependencies between data cells, neither within triangles nor between them.

Now the residuals of two of these triangles will display little dependency after application of model (6.1)-(6.2) to each of them. the reason is that the identical row effects will be detected in the modelling (up to stochastic variation), and most or all of any dependency that might have been induced by this source modelled away.

Now suppose that one wishes to forecast loss experience of the **next** five accident years, for which there is as yet no information. It might be reasonable to assume zero correlation between the two triangles over these future years, but only if several conditions are satisfied:

1. one has been sufficiently clever to recognize that the trends in row effects are identical as between the triangles;
2. one also knows the trends over the future years (they might continue to be identical, or might become independent);
3. one forecasts according to these known trends.

This is a great deal to ask. Suppose, for example, that the trends associated with the different triangles continued to vary identically, but the forecast assumed no variation over future years. This would create a model error in each triangle in the sense of (5.27). Moreover, the model errors in the different triangles would be perfectly correlated. It then follows from (5.4) that the prediction errors of the different triangles will be correlated.

The example given is an extreme one in that row effects are identical as between triangles. The actuary might detect this when reviewing the models of the triangles, and might then be in a position to extrapolate that relationship into the future (if indeed this is correct). In practical cases, however, more subtle relationships will be encountered, more difficult in their recognition. The result is that they may be overlooked, or incorrectly measured, inducing model error in future experience in a manner illustrated by the above example. And this might apply not only to future (unobserved) accident years, but also to the future experience of past accident years.

It is interesting to consider how this reasoning would apply to the case study in Section 6.1, where claim data were de-trended with respect to two trends, due respectively to weather events and seasonal effects. The second of these, once modelled in past data, requires no further action. Any sympathetic trends in the cash flows of separate LoBs will be automatically contained in the model forecasts without resort to the use of correlation.

The effects of weather events require different treatment. These have been removed from the final Home and Motor claims models, and so would be excluded from those models' forecasts. They would typically be restored by means of catastrophe model forecasts, applied simultaneously across LoBs, and so restoring dependency between them. Note, however, that the degree of dependency in those forecasts might differ from that observed in the past if the modelled data set were atypical with respect to major events.

Forecasts will also be subject to other components of model error, classified by O'Dowd et al. (2005) as **internal** and **external systemic errors** respectively, e.g variations of superimposed inflation from model predictions. Each of these components might or might not introduce correlation between forecasts of different LoBs. Each would require specific consideration.

In summary, spurious correlations between triangles arising from inadequate modelling of past data may possibly be eliminated by improved modelling. However, the effects generating these correlations may persist into the future, where it may not be possible to model them away. Even when models of past data indicate an absence of correlation between triangles, it may be appropriate to assume non-zero correlation between the forecasts of same triangles into the future.

8. Conclusion

Risk margins on insurance liability reserves, and capital margins are often set as VaR measures on the total insurance portfolio, across all LoBs. The greater the estimated prediction error in these reserves,

the greater the margins required. The calculation of VaRs also takes account of cross-LoB dependencies. Usually, the greater the assumed dependency, the greater the required risk margin or capital margin. In some jurisdictions (e.g. Australia) draconian cross-LoB dependencies are assumed; recall Section 2.2 and Tables 2.1-2.2. If unjustified, these may demand that insurers hold unnecessarily large, and therefore lazy, amounts of capital.

The present paper has examined dependency in the form of correlation. This may or may not be of relevance to high-percentile VaRs, but it will usually be of relevance to moderate-percentile (e.g.75%) VaRs on which risk margins liability reserves are often based.

Whenever correlations are to be considered, it is necessary to state with care the precise subjects of the correlations. A reference to an Auto-Public Liability correlation, for example, is totally meaningless in isolation.

Correlation is a measure of two stochastic variates. In any data analysis, the non-stochastic components of observations must be removed, leaving the stochastic components, before correlation is calculated. The removal of the non-stochastic components amounts to the modelling of the data, and the remaining stochastic components are the residuals relative to the model adopted. Thus, a calculated correlation will acquire meaning only when cited relative to such a model.

Since the form of model applied to the data will affect the residuals, it also follows from this discussion that the measured correlation will be model-dependent. In short, poor modelling is likely to produce unreliable estimates of correlation. The effect of quality of modelling is explored in detail in Section 5, and is found to be important in a number of respects.

First, inferior modelling is likely to lead to excessive prediction error. This, in itself, is likely to generate excessive risk and/or capital margins in respect of individual LoBs.

Second, and more subtly, is the risk that poor modelling will generate apparent cross-LoB correlations where none in fact exist. False positive correlations will generate even greater excess in margins. Section 6.1 provides a case study into how such false correlations can occur, and how they can be modelled away with more refined modelling.

From a theoretical viewpoint, one could perhaps note that the opposite result might occur, i.e. the omission of influential covariates from a model in the presence of genuine correlations between observations might lead to cancellation of those correlations in paramater estimates. That is to say, the apparent absence of correlation where it actually exists. One would expect, however, that this type of distortion would be less common than the type discussed in this paper. A considerable degree of fortuitousness would be required for the model mis-specification to be just such as to conceal a genuine correlation, whereas spurious correlations, as discussed in this paper, can be easily induced.

Section 6 also carries out an empirical study of cross-LoB correlations in two jurisdictions, namely Australia and the US. Results of the study may be found there, but a very brief summary is as follows:

Australia :

- Of the seven correlations estimated for six LoB pairs, all but one were close to zero.
- The one exception related to the Home-Motor LoB pair and, in the presence of very simple modelling, appeared extremely high (of the order 60%).
- However, it was found that much of the apparent correlation related to major weather events that affected both LoBs, and sympathetic changes in the rate of settlement of claims in the two LoBs.
- When the models of paid losses were controlled for these extraneous effects, measured correlation was vastly reduced.

US :

- Of the six correlations estimated for six LoB pairs, averaged over a substantial number of insurers, three were essentially zero.
- The remaining three LoB pairs showed a tendency to positive correlation, but only a slight one, with no sample correlation exceeding 0.10.

Thus, there are two major conclusions to be drawn from this paper:

1. In any attempt to measure cross-LoB correlations, careful modelling of the data needs to be the order of the day. The exercise will not be well served by rough modelling, such as the use of simple chain ladders, and may indeed result in the prescription of excessive risk margins and/or capital margins.
2. Such empirical evidence as examined here reveals cross-LoB correlations that vary only in the range zero to very modest. There is little evidence in favour of the high correlation that is assumed in some jurisdictions. The evidence suggests that these assumptions derived from either poor modelling or a misconception of the cross-LoB dependencies relevant to the purpose to which they are applied.

The modelling carried out in Section 6 is not claimed to be remarkable in any way. Indeed, it follows a perfectly routine procedure in which residuals are scanned for indications of any influential covariates omitted from the model. When identified, these covariates are included in the model, and this process is continued until no further signal can be found in the residuals. All perfectly routine but, unfortunately, a routine that is not always observed fully in practice. The purpose of the paper is to illustrate how, when it is not observed, grossly misleading conclusions can be reached.

Section 7 is careful to note a fundamental difference in the required treatments of past and future data respectively. It is noted there that, while past data may be sufficient for the identification of specific effects contained within it, and this may be sufficient for any apparent cross-LoB correlations to be modelled away, the same is not necessarily true of future (i.e. unobserved) data. Even if past observations are free of correlation, this may provide only a benchmark in relation to future data, in which case it may be appropriate to allow for non-zero correlations.

This paper also lights the way toward further work on the same subject. One of its themes, noted just above, is that measured correlation is model-dependent. Ample scope therefore exists for the investigation of models other than the relatively simple ones considered here.

The treatment here of major events is necessarily crude, because the data set does not identify individual claims arising from such events. It would be of interest to re-work the results in the presence of such additional data. This would enable the removal from the data of just the claims associated with these events, rather than the blunt instrument of removing whole accident quarters.

And, of course, it should never be forgotten that this paper considers only correlation as a measure of dependency, and so is largely focused on reserving at mid-range VaRs. Consideration of capital margins at more extreme VaRs opens up the question of tail dependency, and a whole new field of exploration.

Acknowledgments

This research was supported under Australian Research Council's Linkage Projects funding scheme (project number LP130100723, with funding partners Allianz Australia Insurance Ltd, Insurance Australia Group Ltd, and Suncorp Metway Ltd). The views expressed herein are those of the authors and are not necessarily those of the supporting organisations. The authors are grateful to anonymous reviewers for comments that led to significant improvements of the paper, to James Basman, Stephen Britt, Yusuf Cakan, and David Koob for very fruitful discussions, as well as to Kevin Lam for his excellent research assistance.

A. Description of the datasets

A.1. The AUSI dataset

The Australian dataset was developed as part of a Linkage Project grant awarded by the Australian Research Council (ARC) until 2016 for a project titled *Modelling claim dependencies for the general insurance industry with economic capital in view: an innovative approach with stochastic processes.* It is be referred to as the **AUSI data set**, an acronym of the names of the project partners (**A**llianz Australia Insurance Ltd, **U**NSW Australia, **S**uncorp Metway Ltd, and **I**nsurance Australia Group Ltd).

The data set currently includes data from two insurers. LoBs covered comprise two property and two casualty lines, as follows, referred to initially in Australian nomenclature, and then US:

— Home (Homeowners), including both Buildings and Contents coverages;

— Private Motor (Private Auto Property Damage);

— Compulsory Third Party ("CTP") (Auto Bodily Injury);
— Public Liability.

Data are provided in respect of a defined period ("the investigation period") in essentially the standard format, consisting of:

— A **policy file**, containing detail of each policy underwritten during the investigation period, e.g. dates of inception and expiry, sum insured, etc.;
— A **claim header file**, containing static information on each claim notified during the investigation period, e.g. dates of claim occurrence and notification, finalisation date (if finalised), claim state, etc.;
— A **claim transaction file**, containing detail of each claim transaction occurring during the investigation period, e.g. transaction date, type of transaction (claim payment or case estimate adjustment), amount of transaction, payment type (e.g. peril or head of damage under which payment is made), claim status (open/closed) after the transaction, etc.

It is evident that the detailed nature of the data set enables analysis to be conducted at a more granular level than the claim triangle, and such analyses are planned for the future. However, for the present simple analysis, claim triangles are adequate. After some preliminary analysis, we removed records containing either blank or invalid information in one or more fields to ensure the integrity of the data. The transactional data were then aggregated according to accident quarter and development quarter.

A.2. The Meyers-Shi dataset

The Meyers-Shi data set has been used for an analysis of US experience. This data set is described by Meyers and Shi (2011).

This data set contains 1010 ($J = 10$) triangles, reporting the claims history as at 31 December 1997 in respect of the 10 accident years 1988-1997. As explained by Meyers & Shi, they are extracted from Schedule P of the data base maintained by the US National Association of Insurance Commissioners.

The Meyers-Shi data base contains premium and paid loss histories in respect of six LoBs, namely:

— Private passenger auto (**"PPA"**);
— Commercial auto (**"CA"**);
— Workers compensation (**"WC"**);
— Medical malpractice;
— Products liability;
— Other liability (**"OL"**).

In each case, a triangle is provided for each of a large number of insurance companies. The data base also contains much other data, not required for present purposes.

References

Bateup, R., Reed, I., 2001. Research and data analysis relevant to the development of standards and guidelines on liability valuation for general insurance. Tech. rep., The Institute of Actuaries of Australia and Tilinghast - Towers Perrin.

Britt, S., Johnstone, 2001. The ABCs of DFA. In: Institute of Actuaries of Australia (Ed.), XIIIth General Insurance Seminar.

Collings, S., White, G., 2001. Apra risk margin analysis. In: Institute of Actuaries of Australia (Ed.), XIIIth General Insurance Seminar.

De Gooijer, J. G., 2006. Detecting change-points in multidimensional stochastic processes. Computational Statistics and Data Analysis 51, 1892–1903.

Embrechts, P., McNeil, A. J., Straumann, D., 2002. Correlation and dependency in risk management: properties and pitfalls. Cambridge University Press, Cambridge.

England, P., Verrall, R., 1999. Analytic and bootstrap estimates of prediction errors in claims reserving. Insurance: Mathematics and Economics 25 (3), 281–293.

Joe, H., 1997. Multivariate Models and Dependence Concepts. Chapman & Hall, London.

Kuang, D., Nielsen, B., Nielsen, J., 2008a. Forecasting with the age-period-cohort model and the extended chain-ladder model. Biometrika.

Kuang, D., Nielsen, B., Nielsen, J., 2008b. Identification of the age-period-cohort model and the extended chain-ladder model. Biometrika 95 (4), 979–986.

Kuang, D., Nielsen, B., Nielsen, J., 2009. Chain-ladder as maximum likelihood revisited. Annals of Actuarial Science 4 (01), 105–121.

Lindskog, F., 2000. Liner correlation estimation. Tech. rep.
Lindskog, F., McNeil, A. J., 2003. Common poisson shock models: Applications to insurance and credit risk modelling. Astin Bulletin 33 (2), 209–238.
Mack, T., Venter, G., 2000. A comparison of stochastic models that reproduce chain ladder reserve estimates. Insurance: Mathematics and Economics 26 (1), 101–107.
Meyers, G., Shi, P., September 2011. Loss Reserving Data Pulled From NAIC Schedule P. http://www.casact.org/research/index.cfm?fa=loss_reserves_data.
Mulvey, Pauling, Britt, S., Morin, 2008. Dynamic Financial Analaysis for Multinational Insurance Companies. Handbook of Asset and Liability Management. Elsevier.
O'Dowd, C., Smith, A., Hardy, P., 2005. A framework for estimating uncertainty in insurance claims cost. In: Institute of Actuaries of Australia (Ed.), XVth General Insurance Seminar.
Risk Margins Task Force, 2008. A framework for assessing risk margins. In: Institute of Actuaries of Australia (Ed.), XVIth General Insurance Seminar.
Rousseeuw, P., Molenberghs, G., 1993. Transformation of non positive semidefinite correlation matrices. Communications in Statistics - Theory and Methods 22 (4), 965–984.
Shi, P., Frees, E. W., 2011. Dependent Loss Reserving Using Copulas. ASTIN Bulletin 41 (2), 449–486.
Shumway, R. H., Stoffer, D. S., 2011. Time Series Analysis and Its Applications, 3rd Edition. Springer Texts in Statistics. Springer New York Dordrecht Heidelberg London.
Taylor, G., 2000. Loss Reserving: An Actuarial Perspective. Huebner International Series on Risk, Insurance and Economic Security. Kluwer Academic Publishers.
Taylor, G., 2011. Maximum likelihood and estimation efficiency of the chain ladder. ASTIN Bulletin 41 (1), 131–155.
Taylor, G., McGuire, G., Sullivan, J., 2008. Individual Claim Loss Reserving Conditioned by Case Estimates. Annals of Actuarial Science 3 (1-2), 215–256.
Vigen, T., 2015. Spurious correlations (last accessed on 18 march 2015 on http://www.tylervigen.com).
Wüthrich, M., Merz, M., 2008. Stochastic claims reserving methods in insurance. John Wiley & Sons.