Improving annuity pricing with address data

Socioeconomic group is an important factor for determining life expectancy, and actuaries have traditionally used pension or benefit size as a proxy for this. This article shows the weaknesses of relying solely on benefit size or amount, and explains how using postcode or address data can improve the assessment of life expectancy

BY STEPHEN J RICHARDS

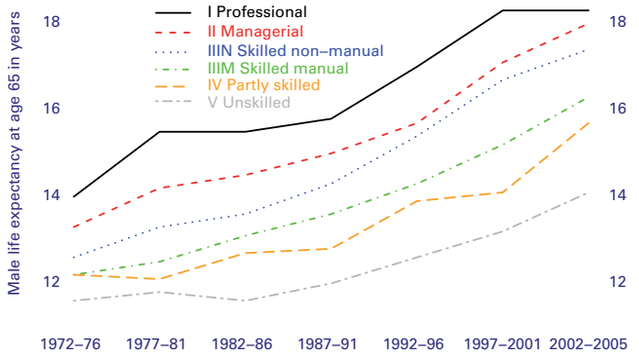
WHEN BOND YIELDS WERE HIGH, accurate assessment of mortality was only of modest importance for pension reserving and annuity pricing. However, now that bond yields are low, accurate estimation of mortality rates has become much more important. With a backdrop of lower long-term interest rates, life insurers need to ensure that they rate policyholders accurately for their mortality and longevity to remain competitive and avoid anti-selection. One particularly important factor is socioeconomic group, which was traditionally assessed purely on the basis of benefit size.

People belonging to more upmarket socioeconomic groups tend to live longer, as shown in Figure 1. It therefore costs more to provide pensions to such people, and life insurers need to charge them more as a result. The drawback of socioeconomic group is that it is determined by occupation. This is a semi-subjective categorisation, which is prone to misclassification or exaggeration, is hard to verify, and is seldom available for existing annuitants of life insurers. What actuaries need for underwriting social group is simple data that reliably predicts higher or lower mortality. One particularly helpful piece of data, the postcode or address, is already collected during the new business process. This article describes how actuaries can use this to improve their procedures for rating for socioeconomic group.

Pension size as proxy

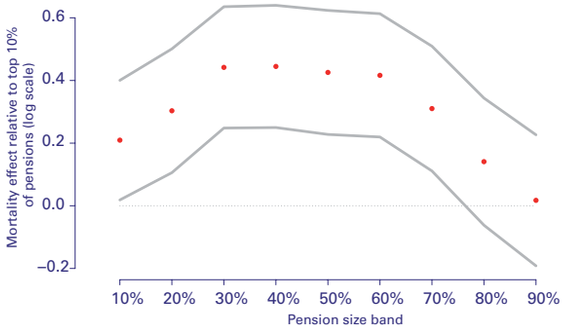
The traditional actuarial solution to addressing socioeconomic differentials has been to use benefit size as a proxy for socioeconomic group: the larger the pension or sum assured, the more up-market the policyholder is likely to be. Pension size is also convenient and not falsifiable. While it is broadly true that mortality decreases by pension size, this does not mean that pension size is a perfect proxy – long-retired people tend to have smaller pensions than recently retired ones, thus conflating birth cohort with pension size; and wealthy people may have a number of pensions where the size of any one of them is not indicative of the individual pensioner's overall income. Furthermore, surviving spouses often have pensions of half or two-thirds of the main pension, thus making pension size an indirect proxy for marital status, too. The result is that pension size is not the smooth, monotonic indicator of mortality by socioeconomic group one would ideally like.

Figure 1. Male life expectancy at age 65 by socio-economic group. Socioeconomic differentials have proved very durable in the UK over the past 30 years



Source: ONS Longitudinal Study. The study is a complete set of census records for a sample of the population of England and Wales. The sample comprises people born on one of four selected dates and was initiated at the time of the 1971 Census, and updated at the 1981, 1991 and 2001 Censuses and in routine event registrations. New study members enter through birth and immigration, and existing members leave through death and emigration. The study represents a continuous sample of the population of England and Wales, rather than a sample taken at one time point only, and now includes records for over 950,000 people. More details can be found at www.celsius.lshtm.ac.uk

Figure 2. Mortality effects by pension size-band (•) after allowing for age and gender with 95% confidence intervals (grey lines)



Source: Longevitas Ltd. UK annuity portfolio observed between 2004–2006, with around 750,000 life-years of exposure and around 19,000 deaths. Mortality is measured against that of the decile of largest pensions. The fact that the effect is not monotonic from left to right shows that pension size is not an infallible guide to socioeconomic group. Note that the effect above is relative to the top decile – that is, a positive value represents an increase in mortality relative to the largest 10% of pensions. The mortality effect is on a natural log scale, so a value of 0.69 would represent a doubling in the basic risk.

The partial reliability of pension size as a proxy for socioeconomic group is shown in Figure 2, where pension size is perhaps a reliable indicator of socioeconomic group for very large pensions, but less so for middling sizes and actually misleading for smaller ones. Indeed, pricing by fund size can change customer behaviour – if you charge higher annuity rates for larger funds, there is a clear incentive for policyholders to split their pension fund to benefit from better rates. Pricing by fund size alone means not only under-pricing the socioeconomic risk, but, to rub salt in the wound, it can mean incurring extra processing costs as well.

Pricing by fund size can change customer behaviour – if you charge higher annuity rates for larger funds, there is a clear incentive for policyholders to split their pension fund to benefit from better rates

A better proxy might be salary, but this is typically not available, or would require an expensive additional verification step for insurers. Salary also has its own weaknesses: surviving spouses don't have one, and people retiring in stages might have an official salary which is unrepresentative of their lifetime earnings. Salary may also be quite hard to define: pensionable salary or payroll salary? With or without annual bonus or commissions? And how to handle the question of non-cash items such as share options? Quite apart from all this is the practical point that people will resent having to provide this information. With all these points in mind, there is a need for some other way to counter the weaknesses of using pension size on its own. The remainder of this article describes how geodemographic profiling can make substantial improvements in risk analysis and pricing for insurance companies, with only modest additional effort and cost.

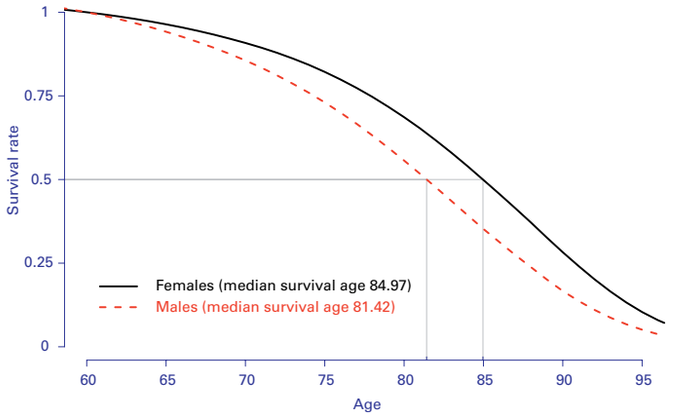
Postcodes and geodemographic models

Walking down a street will usually give quite a good impression of the socioeconomic status of the people who live there. Very broadly, a postcode covers a single street (or less) in the UK, so postcodes should be useful potential indicators of socioeconomic status. As a result, actuaries in the United Kingdom are increasingly making use of so-called geodemographic models of mortality, primarily driven by the use of postcodes. The market for bulk annuities has long been driven by postcode for rating socioeconomic group, and now annuity products marketed directly to individual consumers are priced using postcodes (Legal and General, 2007, and Norwich Union, 2008).

A geodemographic model is one where a person's address or postcode is used to profile their socioeconomic or lifestyle status. Such models have been used for mortality studies before. For example, McLoone (2000) published so-called *Carstairs Scores* for deprivation using Scottish postcode sectors from the 1991 Census. However, a key drawback of this approach lies in the use of the postcode sector rather than the full postcode itself (a postcode sector would be the EH4 2 part of the full EH4 2AB postcode). As noted in McLoone (2000), a postcode sector includes approximately 3,000 households and is only a good predictor of socioeconomic status where the postcode sector is relatively homogeneous.

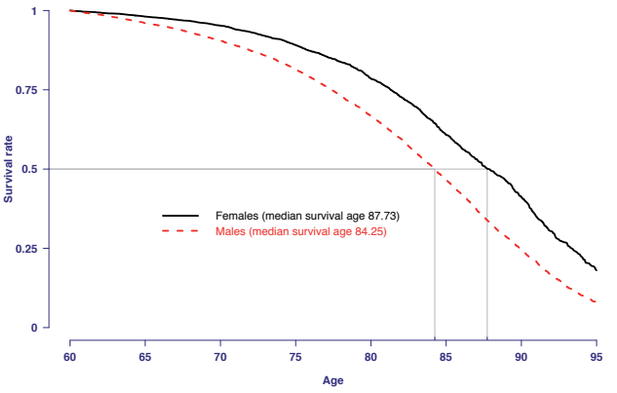
In contrast, in this article we will derive groups from geodemographic

Figure 3. Survival curves for males and females in United Kingdom in 2004–2006



Source: Government Actuary's Department. UK population data between 2004–2006, ages 60–95. Lower female mortality is expressed in a higher proportion surviving to each age. Survival curves are widely used in medical studies, where the standard is to note the age at which 50% remain – that is, the median age at death, which is marked above for both males and females.

Figure 4. Kaplan-Meier survival curves for males and females



Source: Longevitas Ltd . Lower female mortality is expressed in a higher proportion surviving to each age. UK annuity portfolio observed between 2004–2006, with around 750,000 life-years of exposure and around 19,000 deaths.

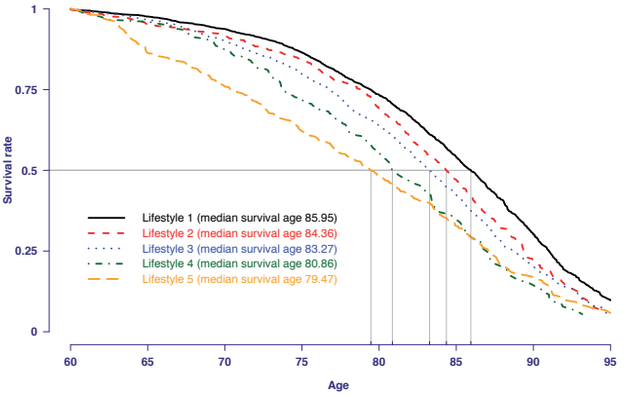
types based on the full UK postcode. This has the advantage that the number of households covered by the same postcode is around 15 on average, and the likelihood of the households being homogeneous with respect to socioeconomic status is therefore greatly increased. Postcodes were introduced to the United Kingdom by the state-owned Royal Mail for the purpose of automating the sorting of mail. UK postcodes are alphanumeric and have covered the entire country since 1974. The full list is available electronically from the Royal Mail as the Postcode Address File (PAF), and UK postcodes are copyrighted.

Postcodes have been widely adopted beyond their original mail-sorting purpose, including consumer profiling for marketing, and premium calculations for general insurance and bulk-annuity pricing. Socioeconomic group is determined by occupation, so we will call our postcode-driven categories lifestyle groups to distinguish them.

There are around 1.8 million postcodes in the UK, covering around 27 million postcode addresses, of which around 1.6 million postcodes are residential. A postcode can cover a whole street, part of a street, or even a single building. Around 200,000 postcodes are for commercial addresses only, and some are non-geographic (such as mailbox addresses). An average of around 15 residential households are covered by a single postcode, providing a high degree of accuracy in determining where a person lives just from their postcode alone. In most cases, a combination of a house number and a postcode is enough to deliver a letter to the correct address.

Postcodes in the UK usually take the form of one of the following patterns: A9 9AA, A99 9AA, A9A 9AA, AA9 9AA, AA99 9AA or AA9A 9AA, where A signifies a letter and 9 a digit. Unfortunately, there is no check digit, so there is no way of knowing if a conformant postcode is actually valid, short of looking up a database of current valid postcodes. The first one or two letters of a postcode form the postcode area, of which there are 124 in the UK. This corresponds to a geographic region and can thus

Figure 5. Kaplan-Meier survival curves for males by lifestyle group, with the median survival age death marked for each



Source: Longevitas Ltd. Males in UK annuity portfolio observed between 2004–2006, lifestyle group determined by optimising homogeneous mortality groups via Mosaic.

be used for modelling regional variations in mortality, although in practice we usually find that there is little or no regional variation after allowing for socioeconomic factors. The first half of the postcode – the EH4 part of EH4 2AB – is known as the postal district, and it is a common mistake to think that this contains all the usable data for postcode profiling. It does not: with an average coverage of 8,800 households, the postal district is even less homogeneous than the postcode sector, and is of little use for modelling mortality differentials as a result. Note that the Royal Mail

Table 1: Some geodemographic profilers in the UK

Name	Provider	No. of type codes	Sample for EH4 2AB: (i) Code	(ii) Description
Acorn	CACI	57	13	Prosperous professionals
CAMIO	Eurodirect	57	5B	Young and older single mortgages
FSS	Experian	45	E13	Fully committed funds
Mosaic	Experian	61	A02	Cultural leadership

typically recodes or reassigns postcodes periodically, so any models driven directly or indirectly by postcode need to be updated regularly.

The key to using postcodes is not the postcode itself, but the type code assigned to that postcode. This type code is one of a smallish number of descriptive categories which summarises what sort of person lives at that sort of address. There are a number of commercial profiling systems available in the UK, which take the form of a database of all valid UK postcodes and the mapping onto the appropriate type code, as shown in Table 1.

Each profiling system has descriptive names and profiles for each category – for example, the postcode EH4 2AB is type code 02 (Cultural leadership) in Mosaic, but type code 13 (Prosperous professionals) in Acorn. These geodemographic profiles were developed primarily for direct marketing purposes, but, as we shall see, they are also particularly effective at predicting mortality differentials.

Similar postcode-driven systems apply in other countries, including the USA (zip code), Canada (postal code) and the Netherlands (postal code). As in the UK, these countries use hierarchical systems, so a given postcode can be used to give both regional and geodemographic information. A similar approach can be applied to other countries, but the full address is usually required for geodemographic modelling, not just the postcode. For example, the German Postleitzahl 89079 tells you the policyholder is in the area of Donaustetten in Baden-Württemberg, but this covers hundreds of households and cannot on its own be used for geodemographic profiling. However, geodemographic profiling based on full address is possible in countries such as France, Germany, Spain, Italy and Japan, and both Experian and Eurodirect market geodemographic profiling systems for these and other territories.

Survival analysis

Before we turn to using postcodes and geodemographic types, we will begin with a simple demonstration of survival analysis for a well-understood risk factor like gender. In *Life & Pensions* May 2007 (Richards, 2007), we showed why survival analysis is a particularly good way to analyse longevity and other risks. A survival curve simply shows the proportion of a population surviving to a given age, starting at one on the left (where everyone is alive) and decreasing monotonically to the right as the proportion of survivors dwindles to zero. Survival analysis has long been used in medical trials, and a survival curve is the probability measure well-known to actuaries as ${}_tp_x$. Survival curves are also the natural thing to study for pensions and annuities, since ${}_tp_x$ is used directly in pricing calculations.

Figure 3 shows the survival curves for population data in the United Kingdom. As expected, there is a pronounced difference between males and females: the probability of reaching any given age is lower for males than for females. The median age at death – that is, the age at which 50% of lives are dead – is three and a half years higher for females.

We have shown the survival curve approach and how it encapsulates the mortality differences between males and females. We now turn to the survival curves for an insured population. A good place to start is the empirical survival curve from your experience data, also known as the Kaplan-Meier estimator. This is a non-parametric approach to survival analysis first introduced by Kaplan and Meier (1958) in one of the most cited scientific papers ever published. It involves no model fitting or parameter estimation and is simply a rearrangement of the mortality experience data. As always in the mortality analysis of life-company data, it is critical to ensure the independence of your observations, so it is necessary to handle people with multiple policies correctly. An effective computerised procedure for de-duplication of multiple policies is described in Richards (2008), and the same procedure has been used for the data here. To illustrate the Kaplan-Meier approach, we first show the empirical survival curves for men and women in a UK annuity portfolio so it can be compared with the population data in Figure 3.

The curve in Figure 4 is a little uneven because this is the empirical survival curve – no smoothing or model-fitting has taken place – and the number of lives is not as large as in Figure 3. The curve in Figure 4 is nevertheless directly comparable with Figure 3, and it shows us exactly what we would have expected: women have a higher probability of reaching a given age than men, a consequence of their lower mortality. Note that members of the insurer population tend to live longer, with a median survival age around three years higher than the general population. It is also interesting to note that the difference between males and females in Figure 4 remains at around three and a half years, almost exactly the same as for the population. Figures 3 and 4 are obviously relevant to the EU directive prohibiting the use of gender for insurance pricing, as discussed in Richards (2004). The gap in the survival curves in Figure 4 equates to around a 10% difference in annuity prices for men and women – at 3% interest, the survival curves in Figure 4 yield an annuity rate of 17.73 for females, which is 10.2% higher than the equivalent 16.09 for males. This difference is obviously highly significant for pricing annuities, where a typical pricing margin might be around 5%. Richards and Jones (2004) showed that gender remained the most significant rating factor for mortality and longevity even after allowing for differences in socioeconomic status and pension size.

There are a number of commercial profiling systems available in the UK, which take the form of a database of all valid UK postcodes and the mapping onto the appropriate type code

Having established survival curves as an effective means of visualising differentials between males and females, we now turn to the use of geodemographic type to indicate mortality differentials. Table 1 lists a number of commercially available products for taking a UK postcode and

mapping it onto a socioeconomic or lifestyle profile. In each case there is a problem for actuaries – the number of types is too large and unwieldy for direct use. For example, the Mosaic classification would require a model with over 60 parameters if each Mosaic type were present in the data. We therefore use an optimisation algorithm to find the best mapping from (say) 61 Mosaic types onto a more manageable number of lifestyle groups (five, say). The optimisation criterion is to find the Mosaic mapping that produces the lowest value for Akaike's Information Criterion, or AIC (Akaike, 1987).

A five-level lifestyle grouping based on Mosaic type codes yields the survival curves in Figure 5. As can be seen by comparing with Figure 4, the gaps in survival can actually be more extreme for lifestyle than by gender. This means that the difference in annuity price can in fact be larger between the lifestyle groups than it is between the genders – at 3% interest, the survival curves in Figure 5 yield an annuity rate of 16.90 for lifestyle group 1 and 13.54 for group 5, a difference of 24.8%. However, this is a logical consequence of sorting people into five groups, where there is more scope for extremes of high and low life expectancy. A more representative comparison might be between lifestyle groups 1 and 3, say, which is nevertheless still highly significant, and results in a financially material difference in annuity price – lifestyle group 3 has an implied annuity rate of 15.73, which means the rate for group 1 is 7.4% higher. As before, these differences are highly material compared with a typical annuity pricing margin of around 5%.

We have seen how postcodes can be highly effective in predicting mortality differentials for pensioners. Since UK insurers already have this information for existing annuitants and must also capture it for new business, it is natural to want to exploit postcodes in annuity pricing. This has already been announced by Legal & General in 2007, and Norwich Union announced in June 2008 that it would follow suit. It is only a matter of time before geodemographic models are a core part of annuity pricing throughout the UK. It is debatable whether this will spread to other countries such as the Netherlands which have similar hierarchical postcodes, and in countries where they do not exist – such as France and Germany – some of the geodemographic systems in Table 1 are available for household-level profiling using the full address.

Practicalities

It is important to navigate some potential pitfalls when using geodemographic profiles. A minor issue is how to handle postcodes that are missing, corrupt or unrecognisable, which we assign to the pseudo-type code 98. Another issue is how to handle valid postcodes that are non-residential addresses, which we assign to the pseudo-type code 99. A particular problem arising from all this is where a block of business has missing postcodes for a specific reason. One example is where pensions are paid to a trustee for onward forwarding to the pensioner and so the insurer holds no addresses for such pensioners. In this situation a specific and distinct class of business is not profiled for a systematic reason, rather than having profiles missing at random. In this case people with missing postcodes could have markedly higher or lower mortality because a missing postcode was simply a marker for a specific sub-category of business.

Another example is a life office where annuitants were given the head office address upon death to suppress mailing. The vast majority of deaths thus all had a non-residential postcode and therefore ended up in category 99, which (unsurprisingly!) proved to be a category of very high mortality. Conversely, there may also be a connection with lower mortality – foreign and overseas addresses will not be profiled, and so end up being assigned the pseudo-Mosaic code 98. If wealthier annuitants are disproportionately likely to live overseas, or if death reporting is less prompt than in the UK, then code 98 will be predictive of low mortality.

One way to detect these sorts of data problem is to calculate the Cramer's V statistic for all categorical variables. Cramer (1999) defines a statistic measuring the strength of association or dependency between two categorical variables. It takes the value 0 for no association, and the value 100 where two variables are perfectly associated and knowledge of one variable completely specifies the other. We include the death status as a categorical variable as well, which helps identify the sort of situation described above.

Table 2 shows the results of all two-way associations between the categorical variables for this life office data set. The Cramer's V statistic is symmetric, so only the values in the upper right of the matrix are shown. The diagonal is not shown, as all the values are 100 (a variable is always perfectly associated with itself). In a clean data set such as this one, the strongest association should usually be between year of birth and death status (people with earlier years of birth should be more likely to be dead). If the association between death status and type were larger, this would be evidence of one of the systematic issues outlined above.

A model that incorporates both geodemographic type and pension size will usually be better than using either variable on its own

Interestingly, we can see that while there is an association between pension size band and geodemographic type (the value 9.7 in Table 2), it is not overwhelming. This is perhaps surprising, as we might expect them both to be proxies for socioeconomic group, and so to be strongly correlated. This figure suggests that while the two are linked, they are not replacements for each other. This means that a model that incorporates both geodemographic type and pension size will usually be better than using either variable on its own.

Table 2: Cramer's V statistic for life-office pensioner data set (all ages)

	Gender	Region code	Size band	Death status	Type
Birth year	21.6	3.1	11.4	54.4	4.0
Gender		4.8	16.1	12.4	5.6
Region code			5.9	6.4	20.6
Size band				17.4	9.7
Death status					10.4

Source: Own calculations using life-office annuitant data. The status variable takes the value 1 for a death, zero otherwise. The type variable is Experian's Postcode Mosaic Type (61 levels, plus two further levels for commercial addresses and unrecognised postcodes). The region code is the UK region extracted from the postcode (124 levels). The relatively high association between type and region code comes from the group of unrecognised postcodes, which are assigned a dummy type code of 98 and a dummy region code of XX.

It is instructive to examine why pension size and geodemographic profiles might both be powerful predictors of pensioner mortality. In each case, the variable in question is merely acting as a proxy for the true underlying drivers of mortality differentials – smoking, diet, drinking and other health behaviours (or absence of them). In the presence of information on smoker status in a model, for example, we would expect a much reduced impact of proxies for socioeconomic group such as pension size or geodemographic type. This reducing role of pension size is echoed by Richards and Jones (2004), who rated pension size as only the fifth most important rating factor, after age, gender, lifestyle and duration since retirement.

A closing point about postcodes

At the time of writing, two UK life insurers have announced they are using postcodes for pricing individual pension annuities – Legal & General in 2007 and Norwich Union in 2008. Much of the press coverage focused on regional differences in life expectancy, with comparisons of different cities or TV regions. However, it is important to distinguish between postcode-derived region and postcode-derived lifestyle. The former approach groups by physical location and includes hundreds of thousands of heterogeneous households. The latter approach, as advocated in this article, combines postcodes in geographically disparate postcodes but of homogeneous lifestyle. A regional interpretation of postcode is a much weaker predictor of pensioner longevity than a lifestyle definition based on a geodemographic profile of the postcode.

> It is instructive to examine why pension size and geodemographic profiles might both be powerful predictors of pensioner mortality

Conclusions

Mortality varies significantly by socioeconomic group, so it is essential to allow for this in life insurance pricing. The traditional actuarial approach of using benefit size on its own has demonstrable weaknesses, but geodemographic profiling is a low-cost and easy way to counter them. Life insurers already have this geodemographic data, so using addresses and postcodes represent an obvious step in strengthening their pricing models. L&P

Stephen J Richards is director at Longevitas Ltd

All data processing, profiling and modelling was done with the Longevitas software (www.longevitas.co.uk) for life insurance risks. Justin Armsworth of Experian provided the Mosaic type codes. Graphs were created in R.

References

Akaike, H. (1987)
Factor analysis and AIC
Psychometrica 52, 317–333

Cramer, H. (1999)
Mathematical Methods of Statistics
Princeton University Press, ISBN13: 978-0-691-00547-8

Kaplan, E.L. and Meier, P. (1958)
Nonparametric estimation from incomplete observations
Journal of The American Statistical Association 53, 457–481

Legal and General plc (2007)
Legal and General links with Hargreaves Lansdown to pioneer postcode-rated annuities
www.legalandgeneralmediacentre.com

Longevitas Development Team (2008)
www.longevitas.co.uk

McLoone, P. (2000)
Carstairs Scores for Scottish Postcode Sectors from the 1991 Census
Public Health Research Unit, University of Glasgow

R Development Core Team (2007)
R: A language and environment for statistical computing
R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0
www.r-project.org

Richards, S.J. (2004)
Gender differentials in mortality
The Actuary magazine, February 2004

Richards, S.J. and Jones, G.L. (2004)
Financial aspects of longevity risk
Staple Inn Actuarial Society, London

Richards, S.J. (2007)
Modelling pensioner longevity
Life & Pensions magazine, May 2007

Richards, S.J. (2008)
Applying survival models to pensioner mortality data
Institute of Actuaries Sessional Meeting Paper, 25 February 2008

Submission guidelines for technical articles

Life & Pensions welcomes the submission of technical articles on topics relevant to our practitioner readership. Core areas include solvency and economic capital modelling, the measurement and management of financial, biometric and operational risks, market-consistent valuation and financing of life and pension balance sheets and cashflows, and investment management. This list is not an exhaustive one.

The most important publication criteria are originality, exclusivity and relevance. In the interests of our readers, we attempt to strike a balance between these. Thus, while we will not publish executive summaries of longer papers (on the grounds of exclusivity), we may accept papers that draw partially but not completely on research submitted elsewhere should our referees recommend it on the grounds of originality and relevance to practitioners.

Given that *Life & Pensions* technical articles are shorter than those in dedicated academic journals, clarity of exposition is another yardstick for publication. Once received by the editor and his team, submissions are logged, and checked against the criteria above. Articles that fail to meet the criteria are rejected at this stage.Articles are then sent to one or more anonymous referees for peer review. Our referees are drawn from the actuarial, risk management, treasury and investment departments of major life and pensions companies, in addition to academia and regulatory bodies. Depending on the feedback from referees, the technical editor makes a decision to reject or accept the submitted article. His decision is final.

We also welcome the submission of brief communications. These are also peer-reviewed contributions to *Life & Pensions* but the process is less formal than for full-length technical articles. Typically, brief communications address an extension or implementation issue arising from a full-length article that while satisfying our originality, exclusivity and relevance requirements, does not deserve full-length treatment.

Submissions should be sent to the editor at *technical@incisivemedia.com*. The preferred format is MS Word, although Adobe PDFs are acceptable. The maximum recommended length for articles is 3,500 words, and for brief communications 1,000 words, with some allowance for charts and/or formulas. We expect all articles and communications to contain references to previous literature. We reserve the right to cut accepted articles to satisfy production considerations. Authors should allow four to eight weeks for the refereeing process.