Subject CS2

CMP Upgrade 2022/23

CMP Upgrade

This CMP Upgrade lists the changes to the Syllabus, Core Reading and the ActEd material since last year that might realistically affect your chance of success in the exam. It is produced so that you can manually amend your 2022 CMP to make it suitable for study for the 2023 exams. It includes replacement pages and additional pages where appropriate.

Alternatively, you can buy a full set of up-to-date Course Notes / CMP at a significantly reduced price if you have previously bought the full-price Course Notes / CMP in this subject. Please see our 2023 *Student Brochure* for more details.

We only accept the current version of assignments for marking, *ie* those published for the sessions leading to the 2023 exams. If you wish to submit your script for marking but only have an old version, then you can order the current assignments free of charge if you have purchased the same assignments in the same subject in a previous year, and have purchased marking for the 2023 session.

This CMP Upgrade contains:

- all significant changes to the Syllabus and Core Reading
- additional changes to the ActEd Course Notes and Assignments that will make them suitable for study for the 2023 exams.

0 Changes to the Syllabus

Only minor formatting and wording changes have been made to the Syllabus objectives.

1 Changes to the Core Reading

This section contains all the ***non-trivial*** changes to the Core Reading.

Chapter 7

Section 1.1, page 5

The expression in the last sentence has been corrected. The final paragraph should read:

Observing lives between (say) integer ages x and $x+1$, and limiting the period of investigation, are also forms of censoring. Censoring might still occur at unpredictable times – by lapsing a life policy, for example – but survivors will certainly be lost to observation at a known time, either on attaining age $x+1$ or when the investigation ends.

Chapter 13

Section 3, page 20

The first line under the Moving average heading has been corrected. It should read:

A *moving average process* of order q, denoted $MA(q)$, is a sequence $\{X_t\}$ defined by the rule:

Chapter 14

Section 1.1, page 4

The function name in the penultimate paragraph of the R box has been corrected. The line now reads:

As the `ts.plot()` function plots a line graph by default, the points can be added with the `points()` function:

Section 2.4, page 25

The expression in the second line of the second paragraph has been corrected. It should read:

The asymptotic variance of $\tilde{\phi}_k$ is $1/n$ for each $k>p$. Again a normal approximation can be used, so that values of the SPACF outside the range $\pm 2/\sqrt{n}$ may suggest that the $AR(p)$ model is inappropriate.

Chapter 15

Section 3.2, page 25

The final equation on this page should be:

$$\frac{d}{d\theta} l(\theta)\bigg|_{\theta=\hat{\theta}} = 0$$

Section 3.2, page 36

The first paragraph on this page is Core Reading and has been reformatted:

The `fitdistr()` function uses a numerical algorithm for the Weibull distribution, which requires starting values. If no values are provided, then the function automatically calculates a starting point.

Chapter 19

Section 3.7, page 27

There are some errors in the equations embedded in the text of the example in section 3.7. This example should read:

This model can be expanded to deal with expenses as the following example demonstrates.

Each year an insurance company issues a number of household contents insurance policies, for each of which the annual premium is £80. The aggregate annual claims from a single policy have a compound Poisson distribution; the Poisson parameter is 0.4 and individual claim amounts have a gamma distribution with parameters α and λ. The expense involved in settling a claim is a random variable uniformly distributed between £50 and £b (> £50). The amount of the expense is independent of the amount of the associated claim. The random variable S represents the total aggregate claims and expenses in one year from this portfolio. It may be assumed that S has approximately a normal distribution.

(i) Suppose that:

$$\alpha = 1;\ \lambda = 0.01;\ b = 100$$

Show that the company must sell at least 884 policies in a year to be at least 99% sure that the premium income will exceed the claims and expenses outgo.

(ii) Now suppose that the values of α, λ and b are not known with certainty but could be anywhere in the following ranges:

$$0.95 \le \alpha \le 1.05;\ 0.009 \le \lambda \le 0.011;\ 90 \le b \le 110$$

By considering what, for the insurance company, would be the worst possible combination of values for α, λ and b, calculate the number of policies the company must sell to be at least 99% sure that the premium income will exceed the claims and expenses outgo.

2 Changes to the ActEd material

This section contains all the ***non-trivial*** changes to the ActEd text.

Chapter 4

Sections 9 and 10

A new section on the expected time spent in a state has been added between Sections 9 and 10. Replacement pages can be found at the end of this Upgrade.

Summary, page 48

The following entry has been added for the new section on the expected time spent in a state:

Expected time spent in a state

Let Y_k be the random variable denoting the amount of time spent in State k over the period from time s to time t. Conditional on the process being in State i at time s, the expected value of Y_k is given by:

$$E[Y_k \mid X_s = i] = \int_0^{t-s} p_{ik}(w) dw$$

Chapter 10

Section 7.1, page 24

The following paragraph has been added underneath the assumptions in this section:

In practice, the statistic is robust to (some) departures from this requirement. A less conservative approach would be to ensure that all expected values are greater than 1 and not more than 20% of the values are less than 5.

Chapter 11

Section 3.1, pages 13-16

Some additional detail on splines has been added on pages 13 and 14. Replacement pages for pages 13 to 16 (for ease of fitting into the 2022 notes) can be found at the end of this Upgrade.

Chapter 12

Sections 2.4 and 2.5, pages 27-34

There are various changes to these pages. Replacement pages can be found at the end of this Upgrade.

Summary, pages 41-44

There are various changes to these pages. Replacement pages can be found at the end of this Upgrade.

Practice question 12.3, page 47

The wording for this question has been updated. The (now four) question parts read:

(a) Set out the revised model that uses a cubic spline function as suggested by your colleague.

(b) Give a possible reason for the inadequate fit of the original model on the observed data and explain how the use of a cubic spline function could improve the fit of the model.

A second colleague has challenged the use of cubic splines for this purpose, arguing that the resulting model tends to be too 'rough'.

(c) Explain what is meant by 'rough' in this context and describe how the method of *p*-splines could be used to help address this difficulty when fitting the model to past data.

(d) Explain how the method of *p*-splines can also be used to forecast future mortality rates.

Practice solution 12.3, pages 53-54

There are various changes to these pages. Replacement pages can be found at the end of this Upgrade.

Chapter 14

Section 1.5, pages 15-16

There are various changes to these pages. Replacement pages can be found at the end of this Upgrade.

Section 4.2, page 38

A question has been added under the R box on this page. Replacement pages can be found at the end of this Upgrade.

Chapter 16

Section 2.3, page 10

The first ActEd paragraph in this section has been replaced with the following:

In Section 2.2, we considered the specific case where the underlying distribution is $X \sim Exp(\lambda)$ and we saw that, for the given values of α_n and β_n :

$$\lim_{n\to\infty}\left[F\left(\beta_n x+\alpha_n\right)\right]^n = \lim_{n\to\infty}\left\{1-\frac{e^{-x}}{n}\right\}^n = e^{-e^{-x}}$$

More generally, for any underlying distribution, if we can identify α_n and β_n such that $\lim_{n\to\infty}\left[F(\beta_n x+\alpha_n)\right]^n$ exists, then it is given by the CDF of a generalised extreme value distribution. In other words, the quantity $\frac{X_M-\alpha_n}{\beta_n}$ tends in distribution to a GEV distribution.

Formulae exist to determine values of α_n and β_n for each underlying distribution that result in convergence (if such values exist). However, this is beyond the Subject CS2 syllabus.

This result is similar to the Central Limit Theorem, which states that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ converges in distribution to the $N(0,1)$ distribution. However, whilst the limit is always the $N(0,1)$ distribution for the standardised sample mean, there are different types of extreme value distribution. The distribution that appears in the limit depends on the underlying distribution of the data.

Section 2.5, pages 17 and 18

A new section, Section 2.6 has been added to these pages. Replacement pages can be found at the end of this Upgrade.

Chapter 17

Practice solution 17.5, page 55

The final expression in the solution to Question 17.5 has a mistake in the power. The power should be $2^{1/\alpha}-1$ instead of $2^{1/\alpha-1}$. The final limit should be:

$$\lim_{u\to 0^+} u^{2^{1/\alpha}-1}$$

Practice solution 17.6, page 56

The solution to part (ii) has been updated to the following:

(ii) ***Comment***

In terms of increasing probabilities, the order is Frank (0.0583), Gumbel (0.0986), then Clayton (0.1089). Each of these is higher than 0.0198, which we would get if we used the product copula. This is because, for the values of the parameters given, the Frank, Gumbel and Clayton copulas all assume that X and Y are positively associated (whereas the product copula assumes they are independent). [1]

Studies also suggest that if one member of a married couple dies, this can precipitate the death of the other member ('broken heart syndrome'). On this basis, we might choose to use a copula function where there is a degree of positive interdependence throughout, *eg* the Frank copula (with a positive parameter). [1]

Although we used the same parameter $\alpha = 5$ in each of the three copula functions, the effect of this parameter on the calculation will vary depending on the copula, and so the results are not directly comparable. We would need further information on the strength of the relationship and on any tail dependence to decide between the three copulas. [1]

We can choose Gumbel or Clayton (with a positive parameter) if there appears to be tail dependence, if say we believe accidental deaths are more relevant at younger ages or the 'broken heart syndrome' applies at older ages.

Chapter 19

Section 3.3, page 11

The labelling of equations from Equation 19.4 onwards is incorrect. Equation 19.4 should be labelled 19.2 and so on.

Chapter 21

Section 3.2, page 32

There is a typo in the solution to part (ii)(b). It should read:

Similarly, the estimated probability that a claim from Region 1 for a Large amount is fraudulent is $\frac{3}{3+176} = 0.0168$, *ie* 1.68%.

Section 3.3, page 35

There is a mistake in the penultimate paragraph of this page. This paragraph is discussing the values of the quantity $\sum_{k=1}^{K} p_{jk}(1-p_{jk})$ and not the Gini index. The paragraph should read:

For a classification problem where the data points are divided into m distinct categories, this quantity must take a value between 0 and $1-\frac{1}{m}$. As $m \to \infty$, the upper limit of this quantity tends to 1.

Section 3.3, page 41

Around halfway down the page, part (i)(a) should be part (i)(b).

Section 3.3, pages 45, 46, 47

There is a typo in the titles of the graphs on these pages. The first line of the title should read:

Predicted vs. observed median house

Section 3.3, page 46

There is an error in the section reference around halfway down the page. It should read:

When we introduced random forests in Section 3.3, we discussed considering subsets of the input variables at each split point.

Summary, page 60

The expression for the penalised log-likelihood in the Penalised generalised linear models section is incorrect. It should be:

Penalised generalised linear models

Penalised regression is an adaptation of the method of maximum likelihood where a penalty is applied to constrain the estimated values of the parameters to improve their reliability for making predictions. The method involves maximising the penalised likelihood:

$$l(\beta_0,\beta_1,\ldots,\beta_d \mid \mathbf{x},\mathbf{y}) - \lambda g(\beta_{0,}\beta_1,\ldots,\beta_d)$$

3 Changes to the X Assignments

Overall

There have been minor changes throughout the assignments, including changes to mark allocations.

Additional solutions to mathematical parts of questions have been added in order to illustrate how they could be answered in the exam, using Word.

More significant changes are listed below.

Assignment X1

Solution X1.3, page 3

The solution for Chain 1 does not reflect the latest Core Reading on periodicity. The solution should read:

Chain 1 is not periodic or aperiodic. It is not possible to return to State 1 at all and State 2 is aperiodic.

Solution X1.11, page 26

The second paragraph of part (v) has been updated to the following:

This is unlikely to be the case in practice as the department may be particularly busy at weekends or during the winter months, resulting in longer waiting times for patients. The transition rates may also depend on the period of time that a patient has been in a particular state. So, a time-inhomogeneous model is likely to be more appropriate.

Assignment X2

Question X2.5, page 2

The part reference in part (ii) is incorrect. It should read:

Write down an integral expression for $p_{12}(x,x+t)$ in terms of transition rates and the probabilities in part (i).

Solution X2.6, page 8

There is a typo in the expression for the sum of a geometric series at the top of the page. It should read:

$$a+ar+ar^2+\ldots+ar^{n-1}=\frac{a(1-r^n)}{1-r}$$

Question X2.7, page 3

The first paragraph of the question has been updated to the following:

A two-year study of 18 jockeys was undertaken. The jockeys were observed from the time of their first horse race until the time they first fell off a horse during a race or until they were censored. Jockeys who had not yet fallen off a horse at the end of the study were treated as censored.

Part (c) has been updated to the following:

(c) the number of jockeys censored at the end of the study, assuming that no other censoring took place between the time of the final fall and the end of the two years.

Assignment X3

Solution X3.6, page 8

The solution to part (ii) has been updated to the following (the paragraph underneath part (ii) has also been removed):

From the conclusions in part (i), the graduated rates appear to be consistent with the observed rates and there is no significant evidence of overall bias. [½]

So, based on these tests alone, there is nothing to suggest that using the graduated rates to calculate the benefits payable is not reasonable. [½]

However, there could be other issues with the graduation that would not be captured by these tests … [½]

… such as a small number of outliers, clumping of bias or it not being smooth. [½]

These issues could be formally tested using other graduation tests. [½]

Inspection of the standardised deviations shows that the graduated rates are generally less than the observed rates (even though we found no statistically significant overall bias). [½]

Using lower mortality rates will tend to overstate the value of pension fund liabilities. So, using these graduated rates may lead to larger (usually employer) contributions than are necessary. This may not be a problem for the scheme unless the employer struggles to pay these larger amounts. [1]

The observed rates will reflect the current mortality rates and not the future mortality rates that will be experienced by the scheme's pensioners. Mortality rates may improve over time. If the valuation does not anticipate this improvement, then the scheme's pension liabilities may be undervalued. This problem could be mitigated by projecting the observed rates. [1]
[Maximum 2]

Solution 3.7, page 10

The final two paragraphs of part (ii) have been replaced with the following:

The larger the magnitude of $\hat{b}_x$, the larger the impact the time trend has on that particular age and the larger the change in mortality. For example, we saw a 10-year reduction in mortality of 3.8% at age 65 compared to 12.2% at age 75 for values of $\hat{b}_x$ of 0.28 and 1.3, respectively. [1]

Strictly speaking, this is only true when specifically comparing either two positive or two negative values of $\hat{b}_x$. For example, when $\hat{b}_x = -0.95$ the change is $e^{-0.1\times(-0.95)} = e^{0.095} = 1.0997$ or a 9.97% increase. However, we only see a 9.5% reduction when $\hat{b}_x = 1$, even though it is larger in magnitude.

[Total 4]

Markers: Please award ½ mark for each description and ½ mark for suitable evidence in each case.

Solution 3.11, page 18

The alternative hypothesis has been added to part (i):

(i) ***Statistical tests***

The null hypothesis for all the tests is that the true underlying rates of the population are consistent with the rates given in the standard table. The alternative hypothesis is that they are not. [1]

Solution 3.11, page 19

The final sentence in the solution to part (i)(b) and the subsequent commentary paragraph have been updated to the following:

Since the p-value is greater than 5%, there is insufficient evidence to reject the null hypothesis at the 5% significance level. We conclude that there is no significant evidence of overall bias between the mortality rates of the population and those in the standard table. [1]

As both np and nq are larger than 5, we could instead use the normal approximation to the binomial distribution. Markers: please award full marks to students who use this approach.

The observed standard normal value, including the continuity correction, is:

$$\frac{5.5 - 16\times 0.5}{\sqrt{16\times 0.5\times 0.5}} = -1.25$$

The result is not significant at the 5% level.

Solution 3.11, page 20

The final paragraph of part (i)(c) has been updated to the following:

The observed number of positive groups is greater than 1, so there is insufficient evidence to reject the null hypothesis at the 5% level of significance. We conclude that there is no significant evidence of excess clumping or grouping of deviations of the same sign. [1]

Assignment X4

Solution X4.10, page 19

There is a typo in the first line of the calculation at the top of the page. It should be:

$$\begin{aligned} P\left(X_M \le 495\right) &= P\left(\frac{X_M - 500}{\frac{1}{50}500} \le \frac{495-500}{\frac{1}{50}500}\right) \\ &\approx \exp\left(\frac{495-500}{\frac{1}{50}500}\right) \\ &= \exp(-0.5) = 0.60653 \end{aligned}$$

Solution X4.10, page 19

The comment for part (iii)(c) incorrectly refers to the distribution of the standardised sample mean instead of the standardised sample maximum. It should read:

The probabilities are similar, suggesting that the GEV distribution provides a reasonable approximation to the standardised sample maximum distribution for $n = 50$.

Solution X4.10, page 20

The first line of the solution to part (iv)(c) references the wrong question part. It should read:

Here we have that $W = X - 400 \mid X > 400$. Using the CDF from part (iv)(a), the required probability is:

4 Changes to the Y Assignments

Overall

There have been minor corrections to the Y assignments, which are outlined below.

Assignment Y1

Solution Y1.3, page 18

There is a typo in the second paragraph of part (v). It has been updated to the following:

Specifically, as $e^{\hat{\beta}} = 0.3319$, then according to the model, the hazard for patients undergoing the new treatment is 66.81% lower than those that aren't.

Assignment Y2

Question Y2.4, page 6

Part (v)(a) should ask for a matrix with the same number of rows as the test data set, not the same number of rows as the entire `swiss` data set. It has been updated to the following:

Repeat the steps in parts (iv)(a) and (iv)(b) to generate 1,000 decision trees on bootstrapped samples of the training data, calculating (and storing) the predicted value of `Fertility` for each province of the test data for each tree. You should store your results in a matrix called `preds` that has the same number of rows as the test data and 1,000 columns, one for each generated decision tree.

5 Other tuition services

In addition to the CMP you might find the following services helpful with your study.

5.1 Study material

We also offer the following study material in Subject CS2:

- Flashcards
- Revision Notes
- ASET (ActEd Solutions with Exam Technique) and Mini-ASET
- Mock Exam and AMP (Additional Mock Pack).

For further details on ActEd's study materials, please refer to the 2023 *Student Brochure*, which is available from the ActEd website at **ActEd.co.uk**.

5.2 Tutorials

We typically offer the following (face-to-face and/or online) tutorials in Subject CS2:

- Regular Tutorials (five full days / ten half days)
- Block Tutorials (five days)
- a Preparation Day for the practical exam.
- an Online Classroom.

For further details on ActEd's tutorials, please refer to our latest *Tuition Bulletin*, which is available from the ActEd website at **ActEd.co.uk**.

5.3 Marking

You can have your attempts at any of our assignments or mock exams marked by ActEd. When marking your scripts, we aim to provide specific advice to improve your chances of success in the exam and to return your scripts as quickly as possible.

For further details on ActEd's marking services, please refer to the 2023 *Student Brochure*, which is available from the ActEd website at **ActEd.co.uk**.

5.4 Feedback on the study material

ActEd is always pleased to receive feedback from students about any aspect of our study programmes. Please let us know if you have any specific comments (*eg* about certain sections of the notes or particular questions) or general suggestions about how we can improve the study material. We will incorporate as many of your suggestions as we can when we update the course material each year.

If you have any comments on this course, please send them by email to **CS2@bpp.com**.

Calculate the probability that a life in the sick state dies without ever recovering to the healthy state.

Solution

Here we can use the jump chain since the times are irrelevant. The life must either go straight to the dead state at the next jump, or to state T. If the life goes to state T, then it definitely dies without recovering (as it cannot then re-enter state H or state S). The probability is therefore:

$$\frac{0.05+0.15}{1.00+0.05+0.15}=\frac{0.20}{1.20}=\frac{1}{6}$$

Alternatively, we could calculate the required probability as:

$$1-P(\text{life enters state H when it leaves state S})$$

ie:

$$1-\frac{1.00}{1.20}=\frac{1}{6}$$

9.a Expected time spent in state *k* when starting in state *i*

Let Y_k be the random variable denoting the amount of time spent in State k over the period from time s to time t. Conditional on the process being in State i at time s, the expected value of Y_k is given by:

$$E[Y_k \mid X_s = i] = \int_0^{t-s} p_{ik}(w)dw$$

To see this why this is the case, consider the following set of indicator variables:

$$I_w = \begin{cases} 1 \text{ if } X_{s+w} = k \\ 0 \text{ if } X_{s+w} \neq k \end{cases}$$

Taken together, these indicator variables form a random function:

$$f(w) = I_w = \begin{cases} 1 \text{ if } X_{s+w} = k \\ 0 \text{ if } X_{s+w} \neq k \end{cases}$$

If we were to plot a realisation of $f(w)$ for $0 < w < t - s$, then we'd get periods of time where the function takes the value 0, when the process is not in State k, and periods of time where the function takes the value 1, when the process is in State k. The area under this function gives the total time spent in State k between times s and t:

$$Y_k = \int_0^{t-s} f(w)\, dw = \int_0^{t-s} I_w\, dw$$

So, conditioning on the starting state:

$$E[Y_k \mid X_s = i] = E\left[\int_0^{t-s} I_w dw \mid X_s = i\right]$$

$$= \int_0^{t-s} E[I_w \mid X_s = i]dw$$

Provided that the function we're integrating is a 'well-behaved' function, we can interchange the expectation and the integral as we've done above. Now:

$$E[I_w \mid X_s = i] = p_{ik}(w) \times 1 + (1 - p_{ik}(w)) \times 0$$

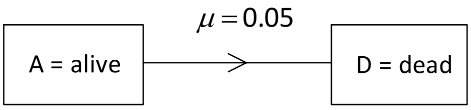
$$= p_{ik}(w)$$

So:

$$E[Y_k \mid X_s = i] = \int_0^{t-s} p_{ik}(w)dw$$

Question

Consider the following two-state Markov jump process that models the mortality of a particular population, where time is measured in years:



(i) Calculate the expected amount of time spent in the alive state between the ages of 50 and 60 for an individual currently aged 50 exact.

(ii) Calculate the expected lifetime of an individual from birth.

Solution

(i) ***Expected time spent alive***

Let X_t be the value of the process (A or D) when the individual is age t. Let Y_A be the random variable denoting the amount of time spent in the alive state over the period from age 50 to 60. Given we start in the alive state, the expected value of Y_A is given by:

$$E[Y_A \mid X_{50} = A] = \int_0^{10} p_{AA}(w)dw$$

Given that once we leave state A there is no return, we have, using the result from Section 7:

$$p_{AA}(w) = p_{\overline{AA}}(w) = e^{-0.05w}$$

So:

$$\begin{aligned} E[Y_A \mid X_{50} = A] &= \int_0^{10} e^{-0.05w}dw \\ &= \left[-\frac{1}{0.05} e^{-0.05w} \right]_0^{10} \\ &= \frac{1}{0.05}(1 - e^{-0.5}) \\ &= 7.869 \end{aligned}$$

The expected time spent in the alive state between the ages of 50 and 60 is 7.869 years.

(ii) ***Expected lifetime from birth***

We now require the overall expected amount of time spent in the Alive state, *ie* integrating to infinity:

$$\begin{aligned}\int_0^\infty p_{AA}(w)dw &= \int_0^\infty e^{-0.05w}dw \\ &= \left[-\frac{1}{0.05}e^{-0.05w}\right]_0^\infty \\ &= \frac{1}{0.05} \\ &= 20\end{aligned}$$

So, the expected future lifetime from birth is 20 years.

Alternatively, in part (ii) we could have used the result from Section 7 about the distribution of holding times. The distribution of the holding time in the alive state is exponential with parameter 0.05. So, the expected amount of time spent in the alive state from birth is given by the reciprocal of the parameter, *ie* 20 years.

In fact, according to this model, the expected future lifetime for an individual of any age is 20 years. This is due to the memoryless property of the exponential distribution.

So, for linearity, we must have:

$$\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = 0 \tag{1}$$

and: $$\beta_1 + \beta_2 + \cdots + \beta_n = 0 \tag{2}$$

These conditions simplify the summation in the formula for a natural cubic spline by reducing the number of required parameters.

The definition of $\phi_j(x)$ leads to the following form for the natural cubic spline over the whole age range:

$$\boldsymbol{\mu_x = \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \Phi_j(x)}$$

where:

$$\boldsymbol{\Phi_j(x) = \phi_j(x) - \left[\frac{x_n - x_j}{x_n - x_{n-1}}\right]\phi_{n-1}(x) + \left[\frac{x_{n-1} - x_j}{x_n - x_{n-1}}\right]\phi_n(x)}$$

The functions $\{1, x, \Phi_1(x), \Phi_2(x), \ldots, \Phi_{n-2}(x)\}$ form what is known as a basis for the set of natural cubic splines with knots at $x_1 < x_2 < \cdots < x_n$. This means that any natural cubic spline with these knots can be represented as a linear combination of these functions, *ie* we can write:

$$N(x) = \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \Phi_j(x)$$

for any natural cubic spline, $N(x)$, for some values of the parameters $\alpha_0, \alpha_1, \beta_1, \beta_2, \ldots, \beta_{n-2}$. So, we can fit a natural cubic spline to data by estimating these parameters.

To see that this is equivalent to the formula for μ_x given earlier, which states that $\mu_x = \alpha_0 + \alpha_1 x + \sum_{j=1}^{n} \beta_j \phi_j(x)$, we will make use of equations (1) and (2).

Multiplying (2) by x_n gives:

$$\beta_1 x_n + \beta_2 x_n + \cdots + \beta_n x_n = 0 \tag{3}$$

Then subtracting (1) from (3):

$$\beta_1(x_n - x_1) + \beta_2(x_n - x_2) + \cdots + \beta_{n-1}(x_n - x_{n-1}) = 0$$

$$\Rightarrow \beta_1(x_n - x_1) + \beta_2(x_n - x_2) + \cdots + \beta_{n-2}(x_n - x_{n-2}) = -\beta_{n-1}(x_n - x_{n-1})$$

$$\Rightarrow \beta_{n-1} = -\left[\beta_1\left(\frac{x_n - x_1}{x_n - x_{n-1}}\right) + \cdots + \beta_{n-2}\left(\frac{x_n - x_{n-2}}{x_n - x_{n-1}}\right)\right] = -\sum_{j=1}^{n-2} \beta_j \left(\frac{x_n - x_j}{x_n - x_{n-1}}\right) \tag{4}$$

In addition, rearranging (2) gives:

$$\beta_n = -(\beta_1 + \beta_2 + \cdots + \beta_{n-1})$$

and using (4):

$$\beta_n = -(\beta_1 + \beta_2 + \cdots + \beta_{n-2}) + \beta_1\left(\frac{x_n - x_1}{x_n - x_{n-1}}\right) + \cdots + \beta_{n-2}\left(\frac{x_n - x_{n-2}}{x_n - x_{n-1}}\right)$$

$$= \beta_1\left(\frac{x_n - x_1}{x_n - x_{n-1}} - 1\right) + \cdots + \beta_{n-2}\left(\frac{x_n - x_{n-2}}{x_n - x_{n-1}} - 1\right)$$

$$= \beta_1\left(\frac{x_{n-1} - x_1}{x_n - x_{n-1}}\right) + \cdots + \beta_{n-2}\left(\frac{x_{n-1} - x_{n-2}}{x_n - x_{n-1}}\right) = \sum_{j=1}^{n-2} \beta_j \left(\frac{x_{n-1} - x_j}{x_n - x_{n-1}}\right)$$

Alternatively, this expression for β_n can be obtained by multiplying (2) by x_{n-1} and then subtracting (1).

So:

$$\mu_x = \alpha_0 + \alpha_1 x + \sum_{j=1}^{n} \beta_j \phi_j(x)$$

$$= \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \phi_j(x) + \beta_{n-1}\phi_{n-1}(x) + \beta_n \phi_n(x)$$

$$= \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \phi_j(x) - \sum_{j=1}^{n-2} \beta_j \left(\frac{x_n - x_j}{x_n - x_{n-1}}\right)\phi_{n-1}(x) + \sum_{j=1}^{n-2} \beta_j \left(\frac{x_{n-1} - x_j}{x_n - x_{n-1}}\right)\phi_n(x)$$

$$= \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \left[\phi_j(x) - \left(\frac{x_n - x_j}{x_n - x_{n-1}}\right)\phi_{n-1}(x) + \left(\frac{x_{n-1} - x_j}{x_n - x_{n-1}}\right)\phi_n(x)\right]$$

$$= \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \Phi_j(x)$$

For the purposes of graduation, we restrict our attention to the set of natural cubic splines. More generally, the functions $\{1, x, x^2, x^3, \phi_1(x), \phi_2(x), \ldots, \phi_n(x)\}$ form what is known as a basis for the set of cubic splines (not just natural cubic splines) with knots at $x_1 < x_2 < \cdots < x_n$. This means that any cubic spline with these knots can be represented as a linear combination of these functions. So, for any cubic spline, $S(x)$, we can write it as:

$$S(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \sum_{j=1}^{n} \beta_j \phi_j(x)$$

for some values of the parameters $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \ldots, \beta_n$.

However, for natural cubic splines, we restrict the function to be linear before the first knot and linear after the last knot. This means we don't need α_2 and α_3 (*ie* they are 0) and we can simplify the sum to represent it in terms of the $\Phi_j(x)$ instead of the $\phi_j(x)$, as shown above.

3.2 The graduation process

The stages in spline graduation are, therefore:

Step 1 – make decisions about knots

Choose the number and value of the knots.

For each knot, the position of x (*ie* the age) is specified, but the value of μ_x is not. It is not necessary for the knots to be equally spaced.

Step 2 – preliminary calculations

Calculate the $\Phi_j(x)$.

Step 3 – estimate the parameter values

Fit the equation $\mathring{\mu}_x = \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \Phi_j(x)$ using weighted least squares, where the weights are proportional to the inverse of the estimated variance of $\tilde{\mu}_x$.

That is, we determine the values of α_0, α_1, β_1, β_2, ..., β_{n-2} that minimise the expression:

$$S = \sum_x w_x (\hat{\mu}_x - \mathring{\mu}_x)^2 = \sum_x w_x \left[\hat{\mu}_x - \left(\alpha_0 + \alpha_1 x + \beta_1 \Phi_1(x) + \beta_2 \Phi_2(x) + \cdots + \beta_{n-2} \Phi_{n-2}(x) \right) \right]^2$$

The value of w_x should be proportional to $E_x^c \big/ \hat{\mu}_x$.

Step 4 – calculate the graduated rates

Calculate the graduated rates using the estimated values of α_0, α_1, β_1, β_2, ..., β_{n-2} from Step 3.

Step 5 – test

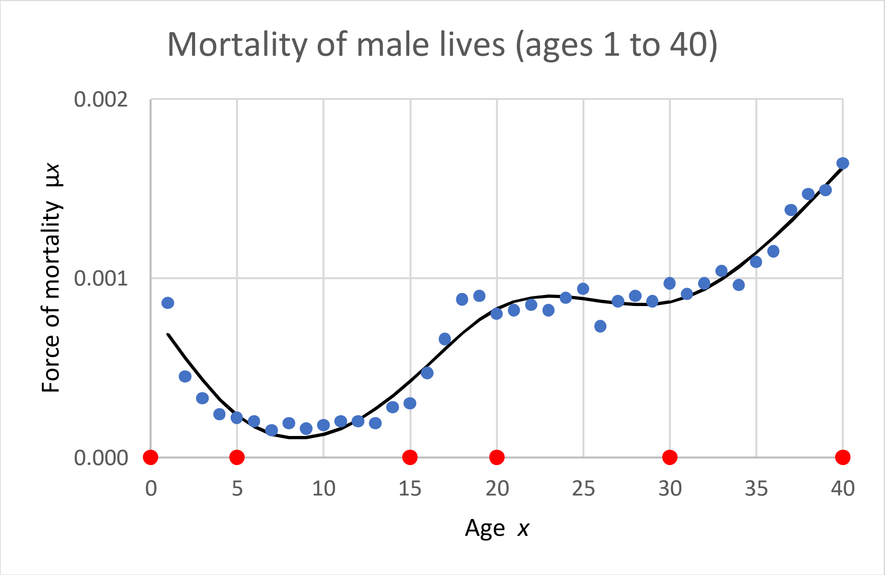
The greater the number of knots, the more closely will the graduated rates adhere to the crude rates, but the less smooth the graduation will be.

The resulting graduation would be subjected to tests of goodness-of-fit to the data (see Chapter 10) which may assist in finding the optimal number of knots.

The remarks in Section 1.4 apply also to experiences graduated by this method.

3.3 Examples of graduations using spline functions

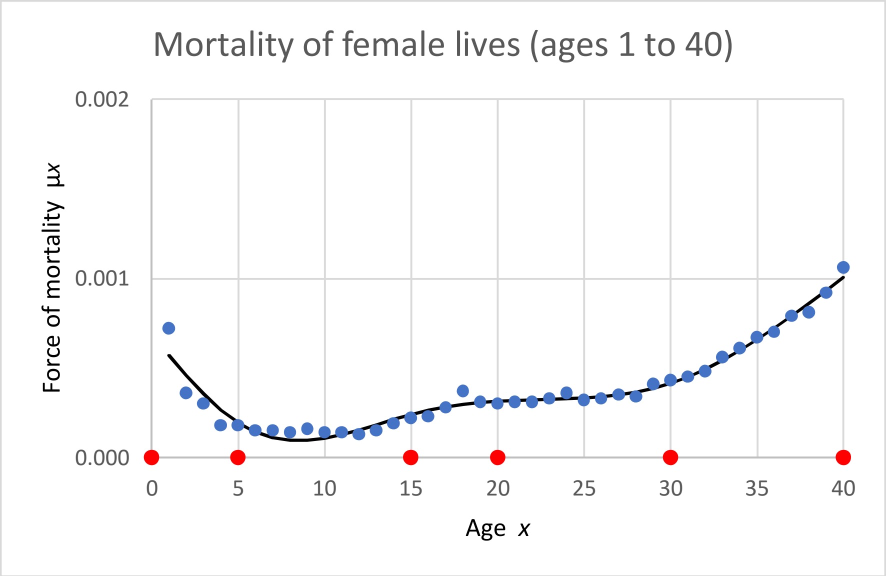
The graphs below show two graduations (one for males and one for females) for a large population of lives similar to those used in English Life Tables No 15 (ELT15). The circles show the crude estimates of the force of mortality at each age (calculated by dividing the number of deaths aged x nearest birthday by the corresponding central exposed to risk). The solid line shows the graduated values, which have been calculated by fitting a cubic spline function with 6 knots, positioned at ages 0, 5, 15, 20, 30 and 40 (marked by circles on the x-axis).



The mortality for the male lives over the age range shown includes several contrasting features:

- relatively high infant mortality at the start (see age 1), which then decreases
- a flat period of very low mortality between ages 5 and 15
- an ‘accident hump’ in the late teenage years (see ages 18 and 19)
- another flat period between ages 20 and 30
- increasing mortality rates from around age 30.

As can be seen, the spline function is able to follow these twists and turns, producing a set of graduated rates that progress smoothly but also adhere closely to the crude rates.



The mortality of the female lives has the same features, although these are less pronounced. Again, the spline function produces a set of graduated rates that progress smoothly but also adhere closely to the crude rates.

Question

(i) Write down a formula for calculating the values of μ_x in the spline graduations illustrated in the graphs above in terms of the fitted parameters, the age x and the functions $\Phi_j(x)$.

(ii) Let $\phi_j(x) = \max\left\{(x - x_j)^3, 0\right\}$. Write down the formula used to calculate $\Phi_1(x)$ in terms of the functions $\phi_j(x)$.

Solution

(i) These graduations each use $n=6$ knots. So the equation for μ_x will have the form:

$$\mu_x = \alpha_0 + \alpha_1 x + \beta_1 \Phi_1(x) + \beta_2 \Phi_2(x) + \beta_3 \Phi_3(x) + \beta_4 \Phi_4(x)$$

(ii) The formula for calculating $\Phi_1(x)$ is:

$$\Phi_1(x) = \phi_1(x) - \left(\frac{x_6 - x_1}{x_6 - x_5}\right)\phi_5(x) + \left(\frac{x_5 - x_1}{x_6 - x_5}\right)\phi_6(x)$$

Here, the first knot is at $x_1 = 0$ and the last two are at $x_5 = 30$ and $x_6 = 40$, so this is:

$$\Phi_1(x) = \phi_1(x) - 4\phi_5(x) + 3\phi_6(x)$$

This page has been left blank so that you can easily put in the replacement pages.

Adding cohort effects to the Lee-Carter model

An age-period-cohort extension of the Lee-Carter model may be written:

$$\ln m_{x,t} = a_x + b_x^1 k_t + b_x^2 h_{t-x} + \varepsilon_{x,t}$$

where h_{t-x} is the overall level of mortality for persons born in year $t-x$. See A. E. Renshaw and S. Haberman (2006) 'A cohort-based extension of the Lee-Carter model for mortality reduction factors', *Insurance, Mathematics and Economics* 38, pp. 556-70.

So in this model we now have two 'b_x' parameters for each age x:

- b_x^1, which is the extent to which the time trend affects mortality rates at age x, and
- b_x^2, which is the extent to which the cohort affects mortality rates at age x.

The superscripts on the symbols are *not* powers, but indicate the two different parameter values at the given age. b_x^1 corresponds to b_x in the standard Lee-Carter model. If we use c to represent the cohort year (as we did in Section 2.2 above), then we could alternatively write the age-period-cohort Lee-Carter model as:

$$\ln m_{x,t,c} = a_x + b_x^1 k_t + b_x^2 h_c + \varepsilon_{x,t}$$

which shows more clearly how the three factors (age x, period t and cohort c) all influence the projected mortality rate.

In this case the h_{t-x} can be estimated from past data and the forecasting achieved using time series methods similar to those described for the k_t parameter in Section 2.3 above.

2.5 Forecasting using *p*-splines

In Chapter 11, the idea of graduation using splines was introduced. In this section, we extend this idea to forecasting.

Splines

We first recap how splines are used for modelling a set of observed mortality rates that are assumed to vary only according to age x, *ie* using a single factor model.

Recall from Chapter 11 that a spline is a polynomial of a specified degree defined on a piecewise basis. The pieces join together at knots, where certain continuity conditions are fulfilled to ensure smoothness. In Chapter 11, the splines were fitted to age-dependent mortality rates, so the knots were specified in terms of ages. Typically, the polynomials used in splines in mortality forecasting are of degree 3 (*ie* cubic).

In Chapter 11 we saw that a (natural) cubic spline function, with n knots at values $x_1, x_2, \dots x_n$, can be written as:

$$f(x) = \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} \beta_j \Phi_j(x)$$

where:

$$\Phi_j(x) = \phi_j(x) - \left[\frac{x_n - x_j}{x_n - x_{n-1}}\right]\phi_{n-1}(x) + \left[\frac{x_{n-1} - x_j}{x_n - x_{n-1}}\right]\phi_n(x)$$

and:

$$\phi_j(x) = \begin{cases} 0 & x < x_j \\ (x - x_j)^3 & x \geq x_j \end{cases}$$

Question

State the continuity conditions that are incorporated in this function, which ensure the smoothness of the joins between the successive cubic functions at each knot.

Solution

The continuity conditions are, for each knot (*ie* at values $x_1, x_2, \dots x_n$):

- the value of the cubic leading into the knot and leading out of the knot are equal at the knot
- the first derivative of the of the cubic leading into the knot and leading out of the knot are equal at the knot
- the second derivative of the cubic leading into the knot and leading out of the knot are equal at the knot.

We also impose the condition that before the first knot, and after the last knot, the spline function is linear (this is what makes it a *natural* cubic spline).

So, we can fit a natural cubic spline to observed mortality rates across an age range by estimating the parameters α_0, α_1 and the β_j.

To construct the model, we choose the number of knots (and hence the number of splines to use), and the degree of the polynomials in each spline. We can then use the splines in a regression model, such as the Gompertz model.

To illustrate, the Gompertz model can be written as:

$$\ln[E(D_x)] = \ln E_x^c + \alpha + \beta x \qquad (1)$$

where $E(D_x)$ is the expected deaths at age x, E_x^c is the central exposed to risk at age x, and α and β are parameters to be estimated.

Rearranging (1):

$$\ln[E(D_x)] - \ln E_x^c = \alpha + \beta x$$

which is equivalent to:

$$\ln\left[\frac{E(D_x)}{E_x^c}\right] = \alpha + \beta x \qquad (2)$$

where $\dfrac{E(D_x)}{E_x^c}$ is the true force of mortality μ for lives labelled as aged x (since $E(D_x) = \mu E_x^c$).

If the age definition for deaths and exposed to risk is 'age nearest birthday', then this will be the force of mortality at *exact* age x, *ie* μ_x.

Question

Show how the Gompertz model just defined relates to the Gompertz law (as shown on page 32 of the *Tables*).

Solution

With deaths and exposed to risk aged x nearest birthday, from (2) the Gompertz model can be written:

$$\ln\left[\frac{E(D_x)}{E_x^c}\right] = \ln \mu_x = \alpha + \beta x \qquad (3)$$

The Gompertz law for the force of mortality at exact age x is usually written as:

$$\mu_x = Bc^x$$

This gives:

$$\ln \mu_x = \ln B + x \ln c$$

So (3) represents the Gompertz law with $B = e^\alpha$ and $c = e^\beta$.

However, the Gompertz law is a deterministic formula, whereas the Gompertz *model* is a stochastic model of the number of deaths occurring in a specified age group.

So instead of writing the model as in (3) we could alternatively define it in stochastic form by replacing $E(D_x)$ with D_x:

$$\ln\left[\frac{D_x}{E_x^c}\right]=\alpha+\beta x+e_x$$

where e_x is a random error term with mean zero (and for which a probability distribution would need to be assumed).

If in (1) we replace the term $\alpha+\beta x$ by a smooth function defined using splines, we have:

$$\ln[E(D_x)]=\ln E_x^c+\sum_{j=1}^{s}\theta_j B_j(x) \qquad (4)$$

where $B_j(x)$ are the set of basis **splines, and θ_j are the parameters to be estimated. The number of splines is s (see Macdonald *et al* (2018)).**

So all we are doing here is replacing the (relatively inflexible) function $\alpha+\beta x$ with the (much more flexible) spline function (*eg* a cubic spline). We should be able to get a much better fit to the observed data than if we just tried to fit the Gompertz law itself.

Here the $B_j(x)$ are a set of basis splines defined over a particular set of knots. Recall from Chapter 11 that this means that any spline of the same degree that is defined over the same set of knots (and has the required continuity conditions) can be represented as a linear sum of these functions. If we are working with a cubic spline basis, this means that we can write any such cubic spline in the form $\sum_{j=1}^{s}\theta_j B_j(x)$. So, a cubic spline can be fitted to data by estimating the θ_j parameters.

This is similar to how the functions $\{1, x, \Phi_1(x), \Phi_2(x),\ldots,\Phi_{n-2}(x)\}$ form a basis for natural cubic splines with a particular set of knots, as discussed in Chapter 11.

A difference between (4) and the spline graduations we described in Chapter 11 is that, in Chapter 11 we fitted the spline directly to observed values of $\hat{\mu}_x$, whereas here we are fitting the spline to observed values of $\ln\hat{\mu}_x$, *ie* using log-transformed data.

To see that this is the case, consider:

$$\frac{E(D_x)}{E_x^c}=\mu_x \;\Rightarrow\; \ln\left[\frac{E(D_x)}{E_x^c}\right]=\ln\mu_x \;\Rightarrow\; \ln E(D_x)=\ln E_x^c+\ln\mu_x$$

Comparing this with Equation (4) we can see that the spline function is being used to model $\ln\mu_x$.

Question

Give one reason for using the log-transformed data in this way.

Solution

Mortality for the middle and older ages generally varies approximately exponentially with increasing age, so the logarithm of the mortality rate would be expected to follow an approximately linear progression. By transforming the data in this way, we can expect to be able to use a simple polynomial to fit the data.

p-splines

Spline models will adhere more closely to past data if the number of knots is large and the degree of the polynomial in the spline function is high. For forecasting, we ideally want a model which takes account of important trends in the past data but is not influenced by short-term 'one-off' variations. This is because it is likely that the short-term variations in the past will not be repeated in the future, so that taking account of them may distort the model in a way which is unhelpful for forecasting. On the other hand, we do want the model to capture trends in the past data which are likely to be continued into the future.

Models which adhere too closely to the past data tend to be 'rough' in the sense that the coefficients for adjacent years do not follow a smooth sequence.

By this we mean that the sequence of estimated parameter values $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_s$ forms an uneven progression. In practice, it is found that roughness in these parameters also leads to a corresponding roughness in the mortality rates that are predicted by the fitted model. So if we can reduce the roughness in the fitted parameters, we should consequently produce a model with a smoother progression of mortality rates from age to age.

The method of *p*-splines attempts to find the optimal model by introducing a penalty for models which have excessive 'roughness'. The method may be implemented as follows:

- **Specify the knot spacing and degree of the polynomials in each spline.**
- **Define a *roughness penalty*, $P(\theta)$, which increases with the variability of adjacent coefficients. This, in effect, measures the amount of roughness in the fitted model.**

  $P(\theta)$ is a function of the fitted parameter values $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_s$ such that, the more irregular the progression of $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_s$ is, the higher $P(\theta)$ will be.

- **Define a smoothing parameter, λ, such that if $\lambda = 0$, there is no penalty for increased roughness, but as λ increases, roughness is penalised more and more.**

  When $\lambda = 0$ we recover the ordinary maximum likelihood problem. As $\lambda \to \infty$ the fitted penalty reduces to 0 to compensate, *ie* the progression of the fitted parameters becomes more regular.

- **Estimate the parameters of the model, including the number of splines, by maximising the penalised log-likelihood**

  $$l_p(\theta) = l(\theta) - \lambda P(\theta)$$

  where $l(\theta)$ is the log-likelihood from model (4).

So, when we wish to estimate the parameters $\theta_1, \theta_2, \ldots, \theta_s$, we first define a likelihood function (in terms of those parameters) that is proportionate to the probability of the observed mortality rates occurring. The 'log-likelihood' is the natural log of this function.

For example, say we use the Poisson model for data with the age definition 'age nearest birthday' and assume that:

$$D_x \sim Poi(\mu_x E_x^c)$$

Then the likelihood for this model with observed deaths denoted by d_x is given by:

$$L \propto \prod_x \frac{\exp(-\mu_x E_x^c) \times (\mu_x E_x^c)^{d_x}}{d_x\,!}$$

Taking logs:

$$\ln(L) = l(\theta) = \sum_x \left[-\mu_x E_x^c + d_x \ln(\mu_x) \right] + \text{constant}$$

If we then use Model (4):

$$\ln[E(D_x)] = \ln E_x^c + \sum_{j=1}^{s} \theta_j B_j(x)$$

ie:

$$\ln \mu_x = \ln\left(\frac{E(D_x)}{E_x^c} \right) = \sum_{j=1}^{s} \theta_j B_j(x)$$

Then we have:

$$l(\theta) = \sum_x \left[-\exp\left[\sum_{j=1}^{s} \theta_j B_j(x) \right] E_x^c + d_x \sum_{j=1}^{s} \theta_j B_j(x) \right] + \text{constant}$$

We then need to choose a penalty function to construct the full penalised likelihood:

$$l_p(\theta) = l(\theta) - \lambda P(\theta)$$

The penalised log-likelihood is effectively trying to balance smoothness and adherence to the data.

'Rougher' values of $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$ would cause the mortality rates to adhere more closely to the data, so the normal (unpenalised) maximum likelihood estimates would tend to result in parameters (and hence fitted mortality rates) that do not progress as smoothly as desired. The penalty factor means that the estimation process automatically compensates for this feature. We can exercise some control over the balance between smoothness and adherence both through the choice of penalty function and by changing the value of λ.

Question

Explain why the following might be a suitable choice of penalty function:

$$P(\theta)=(\theta_1-2\theta_2+\theta_3)^2+(\theta_2-2\theta_3+\theta_4)^2+\cdots+(\theta_{s-2}-2\theta_{s-1}+\theta_s)^2$$

Solution

The function can be alternatively written as:

$$P(\theta)=([\theta_3-\theta_2]-[\theta_2-\theta_1])^2+([\theta_4-\theta_3]-[\theta_3-\theta_2])^2+\cdots+([\theta_s-\theta_{s-1}]-[\theta_{s-1}-\theta_{s-2}])^2$$

This is the same as:

$$\begin{aligned}P(\theta)&=(\Delta\theta_2-\Delta\theta_1)^2+(\Delta\theta_3-\Delta\theta_2)^2+\cdots+(\Delta\theta_{s-1}-\Delta\theta_{s-2})^2\\&=\left(\Delta^2\theta_1\right)^2+\left(\Delta^2\theta_2\right)^2+\cdots+\left(\Delta^2\theta_{s-2}\right)^2\end{aligned}$$

where Δ^r indicates the value of the rth order difference. So, minimising $P(\theta)$ will attempt to select values of θ_j that minimise the sum of the squares of the 2nd differences, in a similar way to the smoothness test in graduation, where we want the 3rd differences to be small.

For example, let's define:

$$f(a,b,c)=(c-b)-(b-a)$$

and compare the values of this function for the sequence $(4,5,7)$ *and the sequence* $(4,7,5)$. *The function yields:*

$$f(4,5,7)=(7-5)-(5-4)=2-1=1$$

$$f(4,7,5)=(5-7)-(7-4)=-2-3=-5$$

So the 'rougher' progression yields the larger absolute value of the second difference.

Squaring the function ensures that the higher absolute values are the most penalised (by making all the values positive), and also places a proportionately greater penalty on the larger absolute differences. This should ultimately result in a smoother final outcome.

The penalty function in this question is known as a second-order penalty. This is because it relates to the 2nd differences.

Question

Model (4) is to be used for fitting a cubic spline to mortality data observed between ages 20 and 60 with knots, x_i, at ages 20, 30, 40, 50 and 60. A set of 7 basis splines is to be used to fit the cubic spline. The model equation is given by:

$$\ln(\mu_x)=\sum_{j=1}^{7}\theta_j B_j(x)$$

In this model, each of the θ_j can be associated with a particular age, y_j. The values of y_j for this exercise are $\{10,20,30,40,50,60,70\}$. You do not need to know how they are derived.

The penalty function to be used is the following second order penalty function:

$$P(\theta)=\left(\theta_1-2\theta_2+\theta_3\right)^2+\left(\theta_2-2\theta_3+\theta_4\right)^2+\cdots+\left(\theta_{s-2}-2\theta_{s-1}+\theta_s\right)^2$$

(i) Show that if $P(\theta)=0$ then the points (y_j,θ_j) lie on a straight line.

For this particular modelling exercise, you may assume that:

$$\sum_{j=1}^{7}B_j(x)=1$$

$$\sum_{j=1}^{7}\left[jB_j(x)\right]=1+i+\frac{x-x_i}{x_{i+1}-x_i} \qquad \text{for} \qquad x_i<x\le x_{i+1}$$

(ii) Show that if the points (y_j,θ_j) lie on a straight line, then $\sum_{j=1}^{7}\theta_j B_j(x)$ is a linear function between the ages of 20 and 60, stating the equation of this function.

(iii) Comment on the implications to your answer to part (ii) if this model were to be fitted using the given penalty function and using a very large value of the smoothing parameter, λ.

Solution

(i) ***Showing the points lie on a straight line***

If $P(\theta)=0$, then:

$$\left(\theta_{j-1}-2\theta_j+\theta_{j+1}\right)^2=0$$

$$\Leftrightarrow \theta_{j-1}-2\theta_j+\theta_{j+1}=0$$

Rearranging gives:

$$\theta_{j+1} - \theta_j = \theta_j - \theta_{j-1}$$

This means that the θ_j are equidistant. As both the θ_j and y_j are equidistant, the points (y_j, θ_j) must lie on a straight line given by the slope:

$$\frac{\theta_j - \theta_{j-1}}{y_j - y_{j-1}}$$

(ii) ***Showing the spline is a linear function***

We know from part (i) that the θ_j are equidistant. Let a denote the constant first difference of the θ_j. We have that:

$$\begin{aligned}\sum_{j=1}^{7} \theta_j B_j(x) &= \theta_1 B_1(x) + \theta_2 B_2(x) + \ldots + \theta_7 B_7(x) \\ &= \theta_1 B_1(x) + (\theta_1 + a) B_2(x) + \ldots + (\theta_1 + 6a) B_7(x) \\ &= \theta_1 \sum_{j=1}^{7} B_j(x) + \sum_{j=1}^{7} \left[(j-1) a B_j(x) \right]\end{aligned}$$

The second sum can be written as:

$$\sum_{j=1}^{7} \left[(j-1) a B_j(x) \right] = a \sum_{j=1}^{7} \left[j B_j(x) \right] - a \sum_{j=1}^{7} B_j(x)$$

From the results given in the question, we have, for $x_i < x \le x_{i+1}$:

$$\begin{aligned}\sum_{j=1}^{7} \theta_j B_j(x) &= \theta_1 + a\left(1 + i + \frac{x - x_i}{x_{i+1} - x_i}\right) - a \\ &= \theta_1 + ai + a\frac{x - x_i}{x_{i+1} - x_i}\end{aligned}$$

ie:

$$\sum_{j=1}^{7} \theta_j B_j(x) = \begin{cases} \theta_1 + a + a\dfrac{x-20}{30-20} & \text{for } 20 < x \le 30 \\ \theta_1 + 2a + a\dfrac{x-30}{40-30} & \text{for } 30 < x \le 40 \\ \theta_1 + 3a + a\dfrac{x-40}{50-40} & \text{for } 40 < x \le 50 \\ \theta_1 + 4a + a\dfrac{x-50}{60-50} & \text{for } 50 < x \le 60 \end{cases}$$

This can also be written as:

$$\sum_{j=1}^{7}\theta_j B_j(x)=\frac{a}{10}x+\theta_1-a \qquad \text{for } 20<x\leq 60$$

which is of the form $y=mx+c$.

(iii) ***Comment***

For a very large value of the smoothing parameter, λ, we know that the penalty function will be small (*ie* close to 0) when maximising the penalised likelihood.

From parts (i) and (ii), this means that $\sum_{j=1}^{7}\theta_j B_j(x)$ should closely resemble a linear function, *ie*:

$$\ln(\mu_x)=\sum_{j=1}^{7}\theta_j B_j(x)$$

is close to linear.

Therefore, the model will closely resemble a Gompertz model.

Forecasting

We now turn to forecasting future mortality rates by age x and time period t , *ie* using a two-factor model. The basic process is to use a (spline) function to model values of $\ln m_{x,t}$ by time period (or year) t , using a different spline function for each age (or group of ages) identified by x . So now we are fitting the function by time period, rather than by age.

So we have:

$$\ln[E(D_{x,t})]=\ln E^c_{x,t}+\sum_{j=1}^{s}\theta_j B_j(t)$$

or:

$$\ln\left[m_{x,t}\right]=\sum_{j=1}^{s}\theta_j B_j(t) \qquad (5)$$

Recall that if x is an integer and we are grouping data by age nearest birthday, then the mortality functions μ and m are essentially interchangeable.

Forecasting using *p*-splines is effected at the same time as the fitting of the model to past data. The past data used will consist of deaths and exposures at ages x for a range of past years t .

Forecasting may be carried out for each age separately, or for many ages simultaneously. In the case of a single age x , we wish to construct a model of the time series of mortality rates $m_{x,t}$ for age x over a period of years, so the knots are specified in terms of years.

Having decided upon the forecasting period (the number of years into the future we wish to forecast mortality), we add to the data set dummy deaths and exposures for each year in the forecast period. These are given a weight of 0 when estimating the model, whereas the existing data are given a weight of 1. This means that the dummy data have no impact on $l(\theta)$. We then choose the regression coefficients in the model (5) so that the penalty $P(\theta)$ is unchanged.

When fitting and forecasting at the same time, the spline given by $\sum_{j=1}^{s} \theta_j B_j(t)$ spans both the period for which we have observed data and the desired future projection period. To fit this model, we need to provide dummy data for the projection periods. Here dummy data means made up death counts and exposed to risk (the actual values used don't matter).

We then still want to maximise the penalised log-likelihood:

$$l_p(\theta) = l(\theta) - \lambda P(\theta)$$

However, we want this to be maximised based on the observed data only and not the dummy data made up for the future periods. To ensure that the dummy data do not affect the penalised log-likelihood, we can:

- give the real data a weight of 1 and the dummy data a weight of 0 when constructing the likelihood. Essentially this means that the dummy data are ignored for the purposes of calculating the first term, $l(\theta)$
- ensure the contribution to the penalty function for the future θ_j coefficients is 0 (this is what the Core Reading is referring to when it states that the penalty function is unchanged when adding in a projection period).

Consider the second-order penality we looked at previously:

$$\left(\theta_1 - 2\theta_2 + \theta_3\right)^2 + \left(\theta_2 - 2\theta_3 + \theta_4\right)^2 + \ldots + \left(\theta_{s-2} - 2\theta_{s-1} + \theta_s\right)^2$$

To ensure that the contribution to the penalty function for the coefficients in the projection period is 0, we need:

$$\left(\theta_{j-1} - 2\theta_j + \theta_{j+1}\right)^2 = 0$$

As we saw in an earlier question, this means that the first difference of the relevant θ_j must be constant. If we assume that the associated time periods of these coefficients, t_j, are also equidistant, then this means that the points (t_j, θ_j) lie on a straight line.

Further, if we use the structure of cubic basis functions described in the same earlier question, this means that the fitted cubic spline will be linear over the projection period. So, when using this model, the spline fitted to the past data is linearly extrapolated for the projection period.

Once the parameters in the period of observed data are estimated, the equation of the straight-line extrapolation is completely defined by the last few values, *ie* the first point at which we require $\left(\theta_{j-1} - 2\theta_j + \theta_{j+1}\right)^2 = 0$. So, the last few years of past data heavily influence this projection.

Advantages and disadvantages of the *p*-spline approach

The *p*-spline approach has the advantages that it is a natural extension of methods of graduation and smoothing, and it is relatively straightforward to implement in R.

It has the following disadvantages:

- **When applied to ages separately, mortality at different ages is forecast independently. So there is a danger that there will be roughness between adjacent ages. This can be overcome by fitting the model and forecasting in two dimensions (age and time) simultaneously.**
- **There is no explanatory element to the projection (in the way that time-series methods use a structure for mortality and an identifiable time series for projection).**

  So, when we fit the model we will obtain a set of numbers for the spline coefficients, but these don't have any natural interpretation, and so it is not easy to compare different models in terms of the differences in the parameter values obtained.

- ***p*-splines tend to be over-responsive to an extra year of data (though this can be ameliorated by increasing the knot spacing).**

  This means that if we fit the model to n years of data, and fit it again to $n+1$ years of data (*eg* because we've just gathered another complete year of observed experience), the model changes more dramatically than we would normally expect (*eg* compared to the case where we are fitting a standard mathematical formula like the Gompertz model).

Extensions and combinations with the Lee-Carter model

Several variations on the *p*-spline approach have been proposed.

R.J. Hyndman and M.S. Ullah (2007) 'Robust forecasting of mortality and fertility rates: a functional data approach', *Computational Statistics and Data Analysis* 51, pp. 4,942-4,956, proposed a model which combines the *p*-spline method with the Lee-Carter model. This can be written as follows:

$$\ln m_{x,t} = a_x + \sum_{j=1}^{J} k_{t,j}\, b_j(x) + \text{error term}$$

Chapter 12 Summary

Mortality projection

Projections of mortality can be made using two-factor models (age x and time period t, or age x and cohort c), or three-factor models (age, time period and cohort). In three-factor models the factors are linked by $x = t - c$.

It is important to know the advantages and disadvantages of using the various two and three-factor models.

The three projection approaches are based on expectation, extrapolation, and explanation.

Methods based on expectation

These use simple deterministic models (*eg* reduction factors), based on expectations of target future mortality rates based on expert opinion and/or on recent historical trends.

Methods based on extrapolation

Lee-Carter model

Two-factor stochastic model (age and period):

$$\ln m_{x,t} = a_x + b_x k_t + \varepsilon_{x,t}$$

where:

a_x is the general shape of mortality at age x

k_t is the effect of time t on mortality

b_x is the extent to which mortality is affected by the time trend at age x

$b_x k_t$ is the effect of time t on mortality at age x

$\varepsilon_{x,t}$ is the error term (independently normally distributed random variables with zero mean and common variance to be estimated).

Time series methods are used to project k_t.

For data collected over n years, the parameter a_x is estimated as follows:

$$\hat{a}_x = \frac{1}{n}\sum_{t=1}^{n} \ln\left(\hat{m}_{x,t}\right)$$

SVD can then be applied to the values $\ln\left(\hat{m}_{x,t}\right) - \hat{a}_x$ in order to estimate b_x and k_t whilst using the constraints $\sum_{t=1}^{n} k_t = 0$ and $\sum_{x} b_x = 1$.

Age-period-cohort version of the Lee-Carter model:

$$\ln m_{x,t} = a_x + b_x^1 k_t + b_x^2 h_{t-x} + \varepsilon_{x,t}$$

where:

h_{t-x} is the effect of cohort year $t-x$ on mortality

b_x^2 is the extent to which mortality is influenced by the cohort effect at age x.

Splines

Spline functions can be used for modelling mortality rates by age using:

$$\ln[E(D_x)] = \ln E_x^c + \sum_{j=1}^{s} \theta_j B_j(x) \;\Rightarrow\; \ln\left(m_x\right) = \sum_{j=1}^{s} \theta_j B_j(x)$$

To use splines for modelling mortality rates by time period, rather than age, the $B_j(x)$ would be replaced by $B_j(t)$ with knots at times $t_1, t_2, \ldots t_s$.

p-splines

When estimating the parameters $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$ using maximum likelihood techniques, we maximise the penalised log-likelihood:

$$l_p(\theta) = l(\theta) - \lambda P(\theta)$$

where $P(\theta)$ is a roughness penalty that increases with the degree of irregularity in the progression of $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$. This is designed to produce a smoother progression of fitted rates with age and/or duration.

An extension of the Lee-Carter model with p-splines:

$$\ln m_{x,t} = a_x + \sum_{j=1}^{J} k_{t,j}\, b_j(x) + \text{error term}$$

where:

the a_x represent the average pattern of mortality across years, but smoothed using *p*-splines.

$b_j(x)$ are a set of basis functions

$k_{t,j}$ are time series coefficients

Methods based on explanation

Projections are made separately by cause of death and combined.

Possible methods include:

- cause-deleted life table approach
- multiple state (Markov) modelling

Difficulties of the approach include:

- forecasting future changes in the risk factors / disease states
- allowing for the lag between changes in the risk factors and their effect on mortality
- difficulties in identifying and categorising the cause of death

Sources of error in mortality forecasting

The main sources of error are:

- model mis-specification
- parameter uncertainty
- incorrect judgement or prior knowledge
- random variation, including seasonal effects
- data errors.

The practice questions start on the next page so that you can keep the chapter summaries together for revision purposes.

(ii)(a) ***Revised projection model using cubic spline function***

The mortality projection model would now be:

$$\ln\left[E\left(D_{x,t}\right)\right]=\ln E_{x,t}^{c}+\sum_{j=1}^{J}\theta_j B_j(t)$$

(b) ***Reasons for inadequate fit and how it could be improved by a cubic spline function***

The trend in logged mortality over time is unlikely to follow a quadratic function over the entire period considered.

Even if the fitted model captures the general change in mortality over time, there are likely to be periods of time where the function fits poorly, *ie* significant periods where the model over or under predicts mortality.

This is due to the relative inflexibility of fitting a single polynomial model (quadratic or otherwise). We are taking a 'one size fits all' approach by trying to fit one function across the entire time period for a particular age.

We could try higher degree polynomials, which are more flexible, to better capture local changes in shape. However, this still has the problem of fitting one function to the entire data set for that age.

On the other hand, spline functions are very flexible models in terms of the shapes being fitted over different periods of time.

As a spline function is piecewise constructed from different polynomials corresponding to different periods of time, it can better capture local changes in shape and so we should expect to see a higher degree of adherence to data.

Cubic splines, in particular, are often used for mortality models and so seem a sensible choice.

(c) ***Use of p-splines***

The problem with splines is that they can be *too* flexible and may cause the model to include historical trend variations that are either short-term or past-specific, and which are not expected to recur in future.

To include these features in the model may then be inappropriate or unhelpful when we attempt to use the model for forecasting purposes.

One symptom of this over-adherence, or roughness, in the model, is that the sequence of estimated parameters $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_J$ may form an uneven progression. Smoothing this progression can help reduce the roughness in the predicted values from the model.

The method of *p*-splines attempts to find an optimal model by introducing a penalty for models which have excessive 'roughness'.

The method may be implemented as follows:

- Specify the knot spacing and degree of the polynomials in each spline.
- Define a *roughness penalty*, $P(\theta)$, which increases with the variability of adjacent coefficients. This, in effect, measures the amount of roughness in the fitted model.
- Define a smoothing parameter, λ, such that if $\lambda = 0$, there is no penalty for increased roughness, but as λ increases, roughness is increasingly penalised.
- Estimate the parameters of the model, including the number of splines, by maximising the penalised log-likelihood:

  $$l_p(\theta) = l(\theta) - \lambda P(\theta)$$

  where $l(\theta)$ would be the usual log-likelihood for the model.

The penalised log-likelihood is effectively trying to balance smoothness and adherence to the data.

(d) ***Forecasting with p-splines***

When using *p*-splines, forecasting is done at the same time as fitting the model to the observed data. This means that the spline given by $\sum_{j=1}^{J} \theta_j B_j(t)$ now spans both the observed and forecast periods.

The steps are the same as those outlined in part (ii)(c) with the exception that dummy data are now incorporated for the forecast period (*ie* made up death data and exposed to risk data).

To ensure that the dummy data do not affect the penalised log-likelihood, the real data are given a weight of 1 and the dummy data are given a weight of 0 when constructing the likelihood.

We also need to ensure that the contribution to the penalty function for the future period is 0.

These two conditions ensure that we still output the penalised maximum likelihood estimate based on the observed data only. However, we also obtain the fitted coefficients for the spline over the projection period, so we can forecast the mortality rates.

The forecasted rates depend on the nature of the spline function and penalty function. For a particular set of cubic basis splines and choice of penalty function, the spline function fitted in the observed period is linearly extrapolated into the projection period.

(iii) ***Disadvantages of using p-splines***

When applied to ages separately, mortality at different ages is forecast independently so there is a danger that there will be roughness between adjacent ages.

There is no explanatory element to the projection (in the way that time series methods use a structure for mortality and an identifiable time series for projection).

p-splines tend to be over-responsive to adding an extra year of data.

12.4 (i) ***Calculating the 1-year probability of dying*** $\boldsymbol{q_{70}}$

A healthy person aged 70 can die over one year by following any one of the following three paths:

(1) transition directly from H to D within one year (HD)

(2) transition from H to Y followed by transition from Y to D in one year (HYD)

(3) transition from H to Z followed by transition from Z to D in one year (HZD)

We need the sum of the probabilities of following each path.

The one-year probabilities are denoted by P_{HD}, P_{HYD} and P_{HZD} respectively.

Healthy directly to dead

We need:

$$P_{HD} = \int_{t=0}^{1} p_{HH}(t)\,\mu_{HD}\,dt$$

which is the probability of someone staying healthy until time t, then dying from healthy at that point, integrated over all possible times t within the one year $(0 < t < 1)$.

This page has been left blank so that you can easily put in the replacement pages.

Example 2

The monthly inflation figures are obtained by seasonal differencing of the Retail Prices Index. If x_t is the value of the RPI in month t, the annual inflation figure reported is:

$$\frac{x_t - x_{t-12}}{x_{t-12}} \times 100\%$$

1.5 Method of moving averages

The method of moving averages makes use of a simple linear filter to eliminate the effects of periodic variation.

A linear filter is a transformation of a time series x (the input series) to create an output series y that satisfies:

$$y_t = \sum_{k=-\infty}^{\infty} a_k x_{t-k}$$

The collection of weights $\{a_k : k \in Z\}$ forms a complete description of the filter. The objective of the filtering is to modify the input series to meet particular objectives, or to display specific features of the data. For example, an important problem in analysis of economic time series is detection, isolation, and removal of deterministic trends.

In practice a filter $\{a_k : k \in Z\}$ normally contains only a relatively small number of non-zero components.

A very simple example of a linear filter is the difference operator $\nabla = 1 - B$. Using this filter produces:

$$y_t = (1-B)x_t = x_t - x_{t-1}$$

So, in this case, we have $a_0 = 1$, $a_1 = -1$ and $a_k = 0$ for all other values of k.

As a second example, suppose that the input series is a white noise process e, and the filter takes the form:

$$a_0 = 1, a_1 = \beta_1, \ldots, a_q = \beta_q \text{ and } a_k = 0 \text{ for all other values of } k$$

Then the output series is $MA(q)$, since we have:

$$y_t = \sum_{k=0}^{q} \beta_k e_{t-k}$$

Conversely, applying a filter of the form:

$$a_0 = 1, a_1 = -\alpha_1, \dots, a_p = -\alpha_p \text{ and } a_k = 0 \text{ for all other values of } k$$

to an input series x that is $AR(p)$ recovers the original white noise series:

$$y_t = x_t - \sum_{k=1}^{p} \alpha_k x_{t-k} = e_t$$

If x is a time series with seasonal effects with even period $d = 2h$, then we define a smoothed process y by:

$$\boldsymbol{y_t = \frac{1}{2h}\left(\frac{1}{2}x_{t-h} + x_{t-h+1} + \cdots + x_{t-1} + x_t + \cdots + x_{t+h-1} + \frac{1}{2}x_{t+h}\right)}$$

This ensures that each period makes an equal contribution to y_t.

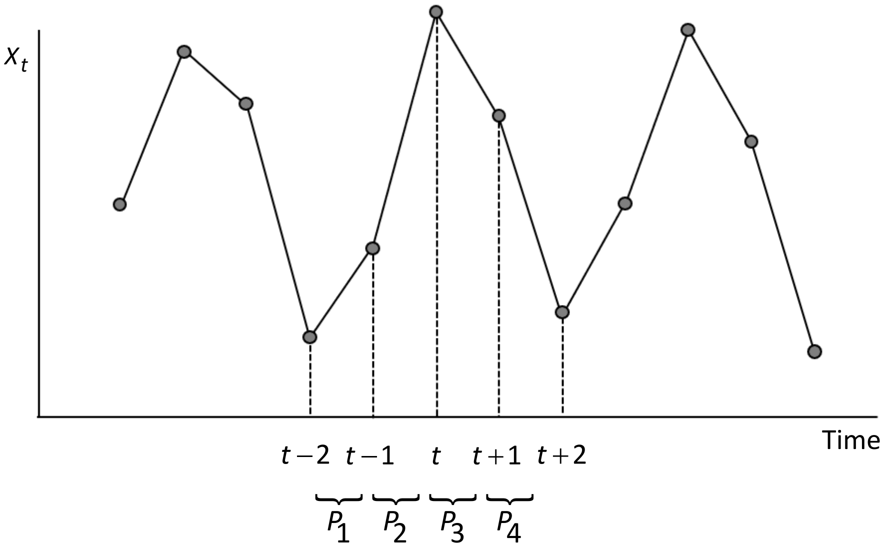
For example, with quarterly data a yearly period will have $d = 4 = 2h$, so $h = 2$ and we have:

$$y_t = \frac{1}{4}\left(\frac{1}{2}x_{t-2} + x_{t-1} + x_t + x_{t+1} + \frac{1}{2}x_{t+2}\right)$$

In this case the filter has weights:

$$a_k = \begin{cases} \frac{1}{8} & \text{for } k = -2, 2 \\ \frac{1}{4} & \text{for } k = -1, 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

To see why this gives a symmetric average round time t, consider the following sample path of a time series that has a seasonal effect with period $d = 4$:



In order to construct a symmetric average around time t, we can consider the four intervals (*ie* two either side of time t) from $t-2$ to $t+2$, denoted by P_1 to P_4 in the diagram above.

We first take the average of the observed time series values at the start and end of each interval. An average of these values then gives a symmetric average around time t:

$$y_t = \frac{1}{4}\left(\frac{x_{t-2}+x_{t-1}}{2}+\frac{x_{t-1}+x_t}{2}+\frac{x_t+x_{t+1}}{2}+\frac{x_{t+1}+x_{t+2}}{2}\right)$$

Simplifying gives:

$$y_t = \frac{1}{4}\left(\frac{1}{2}x_{t-2}+x_{t-1}+x_t+x_{t+1}+\frac{1}{2}x_{t+2}\right)$$

as before.

Alternatively, consider the average of the four values $t-2$ to $t+1$, which gives:

$$z_t = \frac{1}{4}\left(x_{t-2}+x_{t-1}+x_t+x_{t+1}\right)$$

Although this takes the average of four values, the length of the seasonal period, this average is not symmetric around time t. We could also consider the average of the four values from $t-1$ to $t+2$, which gives:

$$r_t = \frac{1}{4}\left(x_{t-1}+x_t+x_{t+1}+x_{t+2}\right)$$

Again, this is not symmetric around time t. However, if we take the average of these averages, we do get a symmetric average around time t:

$$\begin{aligned}\frac{z_t+r_t}{2} &= \frac{\frac{1}{4}\left(x_{t-2}+x_{t-1}+x_t+x_{t+1}\right)+\frac{1}{4}\left(x_{t-1}+x_t+x_{t+1}+x_{t+2}\right)}{2}\\ &= \frac{1}{4}\left(\frac{1}{2}x_{t-2}+x_{t-1}+x_t+x_{t+1}+\frac{1}{2}x_{t+2}\right)\end{aligned}$$

as before.

This type of series, constructed by taking a symmetric average around time t, is also called a *centred* moving average. Such a centred moving average introduces the practical problem that the average can only be calculated in retrospect, *ie* there will be a natural delay.

The same can be done with odd periods $d = 2h+1$, but the end terms x_{t-h} and x_{t+h} do not need to be halved.

For example, with data every 4 months, a yearly period will have $d=3=2h+1$, so $h=1$ and we have:

$$y_t=\frac{1}{3}\left(x_{t-1}+x_t+x_{t+1}\right)$$

In this case the filter has weights:

$$a_k=\begin{cases}\frac{1}{3} & \text{for } k=-1,0,1\\ 0 & \text{otherwise}\end{cases}$$

As with most filtering techniques, care must be taken so that the smoothing of the data does not obscure the very effects which the procedure is intended to uncover.

This implies that:

$$X_n = Z_n + 2X_{n-1} - X_{n-2}$$

Hence:

$$X_{n+2} = 2X_{n+1} - X_n + Z_{n+2}$$

So the two-step ahead forecast is:

$$\hat{x}_n(2) = 2\hat{x}_n(1) - x_n + \hat{z}_n(2)$$

Since Z is $ARMA(1,1)$, it has a defining equation of the form:

$$Z_n = \mu + \alpha(Z_{n-1} - \mu) + e_n + \beta e_{n-1}$$

So:

$$Z_{n+2} = \mu + \alpha(Z_{n+1} - \mu) + e_{n+2} + \beta e_{n+1}$$

and:

$$\hat{z}_n(2) = \hat{\mu} + \hat{\alpha}(\hat{z}_n(1) - \hat{\mu})$$

Hence the two-step ahead forecast can be expressed as:

$$\hat{x}_n(2) = 2\hat{x}_n(1) - x_n + \hat{\mu} + \hat{\alpha}(\hat{z}_n(1) - \hat{\mu})$$

An $ARIMA(p,d,q)$ process with $d > 0$ is not stationary and therefore has no stationary variance. It should come as no surprise, then, that the prediction variance for the k-step ahead forecast increases to infinity as k increases. This is easily seen in the case of the random walk process.

Predicting three steps ahead in R using the $ARIMA(1,0,1)$ model fitted to the data generated in Section 2.1:

```
    predict(fit, n.ahead = 3)

$pred
Time Series:
Start = 301
End = 303
Frequency = 1
[1] 1.1164950 0.7184494 0.4749197

$se
Time Series:
Start = 301
End = 303
Frequency = 1
[1] 0.9495467 1.4808587 1.6359396
```

The `$pred` component of the output contains the predicted values and the `$se` component contains estimated standard errors.

The following code outputs predictions and estimated standard errors for 15 to 20 steps ahead:

```
    predict(fit, n.ahead = 20)$pred[15:20]

[1] 0.09215216 0.09174233 0.09149159 0.09133819 0.09124433 0.09118691

    predict(fit, n.ahead = 20)$se[15:20]

[1] 1.722052 1.722053 1.722053 1.722053 1.722053 1.722053
```

As indicated in Section 4.1, the predicted values and standard errors are converging.

The predicted values are converging to the estimated mean of the process, which is 0.0911 to 4 decimal places.

The standard errors are converging to 1.722053, which is the square root of γ_0 of the fitted model.

Question

The equation of the model fitted to the data generated in Section 2.1 is:

$$X_t = 0.0911 + 0.6118(X_{t-1} - 0.0911) + 0.5849e_{t-1} + e_t$$

where the variance of the white noise terms is estimated to be 0.9016.

If the last time for which an observed value is available is n, the step-ahead forecast for time $n+s$ has the structure:

$$\hat{x}_n(s) = 0.0911 + 0.6118(\hat{x}_n(s-1) - 0.0911) + 0.5849\hat{e}_{n+s-1} + \hat{e}_{n+s}$$

(i) Show that $\hat{x}_n(s) \to 0.0911$ as $s \to \infty$ by considering the structure of the step-ahead forecast for $s > 1$ or otherwise.

The $MA(\infty)$ representation of X_t is given by:

$$X_t = 0.0911 + 1.1967\sum_{j=1}^{\infty}\left[0.6118^{j-1}e_{t-j}\right] + e_t$$

Let X'_{n+s} represent the value of the series at time $n+s$ given the values of the process and white noise terms up to time n (*ie* these values can be treated as constants).

(ii) Derive an expression for the variance of X'_{n+s} .

(iii) Show that the standard deviation of X'_{n+s} tends to 1.722 as $s \to \infty$.

Solution

(i) ***Limit of step-ahead forecast***

For $s \geq 1$, the step-ahead forecasts are given by:

$$\hat{x}_n(s) = 0.0911 + 0.6118(\hat{x}_n(s-1) - 0.0911) + 0.5849\hat{e}_{n+s-1} + \hat{e}_{n+s}$$

However, the expectation of future white noise is 0. This gives:

$$\hat{e}_{n+s} = 0$$

$$\hat{e}_{n+s-1} = 0 \text{ for } s > 1$$

So, for $s > 1$, the forecast simplifies to:

$$\hat{x}_n(s) = 0.0911 + 0.6118(\hat{x}_n(s-1) - 0.0911) \qquad \text{for } s > 1$$

This can also be written as:

$$\hat{x}_n(s) - 0.0911 = 0.6118(\hat{x}_n(s-1) - 0.0911) \qquad \text{for } s > 1$$

Repeated backwards substitution of $\hat{x}_n(s-1) - 0.0911$ on the RHS gives:

$$\hat{x}_n(s) - 0.0911 = 0.6118^{s-1}(\hat{x}_n(1) - 0.0911) \qquad \text{for } s > 1$$

The RHS tends to 0 as $s \to \infty$, *ie* $\hat{x}_n(s) \to 0.0911$ as $s \to \infty$.

(ii) ***Variance of future terms***

From the question, the $MA(\infty)$ representation of X_t is given by:

$$X_t = 0.0911 + 1.1967\sum_{j=1}^{\infty}\left[0.6118^{j-1}e_{t-j}\right] + e_t$$

To determine the variance of X'_{n+s}, we first consider the variance of the first few future values of the process. Using the $MA(\infty)$ representation, the variance of X'_{n+1} is given by:

$$\text{var}(X'_{n+1}) = \text{var}(1.1967\sum_{j=1}^{\infty}\left[0.6118^{j-1}e_{n+1-j}\right]) + \text{var}(e_{n+1})$$

$$= 1.1967^2\sum_{j=1}^{\infty}\left[0.6118^{2j-2}\,\text{var}(e_{n+1-j})\right] + \text{var}(e_{n+1})$$

As we are assuming the white noise terms up to time n are known, the sum is 0 (as the variance of a constant is 0). So:

$$\text{var}(X'_{n+1}) = \text{var}(e_{n+1}) = 0.9016$$

The square root of this is 0.9495, which is the same, to 4DP, as the standard error given in the R output for the 1-step-ahead forecast. We are only using the parameter estimates to 4DP in these calculations so the figures may differ to the R output slightly due to rounding.

The variance of X'_{n+2} is given by:

$$\begin{aligned}\text{var}(X'_{n+2}) &= \text{var}(1.1967\sum_{j=1}^{\infty} 0.6118^{j-1} e_{n+2-j}) + \text{var}(e_{n+2}) \\ &= 1.1967^2 \sum_{j=1}^{\infty}\left[0.6118^{2j-2}\,\text{var}(e_{n+2-j})\right] + \text{var}(e_{n+2})\end{aligned}$$

Assuming we know the values of the white noise terms up to time n, only the first term of the sum is non-zero:

$$\begin{aligned}\text{var}(X'_{n+2}) &= 1.1967^2 \times 0.6118^0 \times \text{var}(e_{n+1}) + \text{var}(e_{n+2}) \\ &= 0.9016(1.1967^2 + 1) \\ &= 2.1928\end{aligned}$$

The square root of this is 1.4808, which is the standard error given in the R output for the 2-step-ahead forecast to 3 decimal places.

In general, for X'_{n+s}, only the first $s-1$ terms of the sum are non-zero:

$$\begin{aligned}\text{var}(X'_{n+s}) &= \text{var}(1.1967\sum_{j=1}^{\infty} 0.6118^{j-1} e_{n+s-j}) + \text{var}(e_{n+s}) \\ &= 1.1967^2 \sum_{j=1}^{\infty}\left[0.6118^{2j-2}\,\text{var}(e_{n+s-j})\right] + \text{var}(e_{n+s}) \\ &= 1.1967^2 \times 0.9016\left(1 + 0.6118^2 + 0.6118^4 + \ldots + 0.6118^{2s-4}\right) + 0.9016 \\ &= 1.2912\left(1 + 0.6118^2 + 0.6118^4 + \ldots + 0.6118^{2s-4}\right) + 0.9016\end{aligned}$$

(iii) ***Limit of the variance***

As $s \to \infty$, the brackets tend to the infinite sum of a geometric progression with first term 1 and common ratio 0.6118^2. Using the formula for the sum to infinity of a geometric progression, we have:

$$1 + 0.6118^2 + 0.6118^4 + \ldots = \frac{1}{1 - 0.6118^2}$$

So:

$$\begin{aligned}\operatorname{var}(X'_{n+s}) \underset{s\to\infty}{\to} &\ 1.2912\frac{1}{1-0.6118^2}+0.9016 \\ &= 2.9652\end{aligned}$$

The standard deviation is the square root of the variance, which is $\sqrt{2.9652} = 1.722$ as required.

The square root calculated here is actually 1.721965, which differs slightly from the limit of standard errors in the R output. Again, this is due to using the values of the parameter estimates to 4DP in these calculations.

These results also intuitively make sense. We would expect this stationary series to fluctuate around its mean in the future. So, the step-ahead point forecasts tend to the fitted mean value. The variance of the series dictates the size of the fluctuations around the mean and so it's logical that our long-term prediction intervals use the variance in their construction.

Prediction intervals constructed in this way tend to be too narrow (eg a 95% prediction interval actually has a lower than 95% chance of containing the corresponding time series value). This is because we have taken the fitted model for granted and only considered the variation in the future errors. The true time series values will be generated by the unknown true model, which may be of a different order and have different parameter values.

4.3 Exponential smoothing

The Box-Jenkins methodology is demanding, requiring a skilled operator to produce reliable results. There are many instances in which a company needs no more than a simple forecast of some future value without having to employ a trained statistician to provide it. A much simpler forecasting technique, introduced by Holt in 1958, uses a weighted combination of past values to predict future observations.

One-step ahead forecast using exponential smoothing

$$\hat{x}_n(1) = \alpha(x_n + (1-\alpha)x_{n-1} + (1-\alpha)^2 x_{n-2} + \cdots)$$

The weights used here are $\alpha, \alpha(1-\alpha), \alpha(1-\alpha)^2, \ldots$.

Here α is a single parameter, either chosen by the user or estimated by least squares from past data. Typically, a value in the range 0.2 to 0.3 is used.

A value of α between 0 and 1 will give a weighted average of historic values with less emphasis on values that are further back in time.

The geometrically decreasing weights give rise to the name *exponential smoothing*.

Since the weights sum to 1, the exponential smoothing filter is a weighted average of historic values, with the weights decreasing geometrically as we go further back in time.

Question

Show that the weights sum to 1 when $0 < \alpha < 1$.

This page has been left blank so that you can easily put in the replacement pages.

	GEV distributions (for the maximum value) corresponding to common loss distributions		
Type **Shape parameter**	**WEIBULL** $\gamma < 0$	**GUMBEL** $\gamma = 0$	**FRÉCHET** $\gamma > 0$
Underlying distribution	**Beta** **Uniform** **Triangular**	**Chi-square** **Exponential** **Gamma** **Lognormal** **Normal** **Weibull**	**Burr** ***F*** **Log-gamma*** **Pareto** ***t***
Range of values permitted	$x < \alpha - \dfrac{\beta}{\gamma}$	$-\infty < x < \infty$	$x > \alpha - \dfrac{\beta}{\gamma}$

* Note that $X \sim loggamma$ if $\ln X \sim Gamma$.

Unhelpfully, the extreme value distribution corresponding to the Weibull distribution from the *Tables* is actually of the Gumbel type (rather than the Weibull type).

Mathematicians have identified criteria that can be used to determine which family a particular distribution belongs to. As a rough guide:

- **underlying distributions that have finite upper limits (*eg* the uniform distribution) are of the Weibull type (which also has a finite upper limit)**
- **'light tail' distributions that have finite moments of all orders (*eg* exponential, normal, lognormal) are typically of the Gumbel type**
- **'heavy tail' distributions whose higher moments can be infinite are of the Fréchet type.**

2.6 Fitting a GEV distribution

So far, we have focused on the distribution of the standardised block maxima and obtained results about its limiting distribution, which depends on the underlying distribution of the data.

However, for an observed set of claims data, we generally won't know the underlying distribution (although we could try and fit distributions to it) meaning we don't know the appropriate limiting GEV. We also can't calculate the standardised block maxima, which depend on the unknown values α_n and β_n, as they also depend on the unknown underlying distribution.

Further, we are generally more interested in modelling the block maxima themselves, rather than the standardised block maxima, so that we can, for example, calculate the probability of the maximum claim in the next block exceeding a certain value.

One approach we can take is directly fitting a GEV distribution to the sample of raw block maxima values, without standardising them. We justify this below.

We know that the quantity $\frac{X_M - \alpha_n}{\beta_n}$ tends in distribution to a GEV distribution (assuming we are working with an underlying distribution for which the limit exists). This means that, for large enough n:

$$\frac{X_M - \alpha_n}{\beta_n} \div GEV(\alpha, \beta, \gamma)$$

for some values of the parameters α, β and γ.

We can recover the block maximum from the standardised block maximum as follows:

$$X_M = \frac{X_M - \alpha_n}{\beta_n} \times \beta_n + \alpha_n$$

Now, it turns out that if we have a random variable that follows a GEV distribution and we multiply it by a constant or add a constant, then it can be shown (see below) that the resulting random variable follows another GEV distribution but with different parameters. This means that, for large enough n:

$$X_M \div GEV(\alpha', \beta', \gamma')$$

for some values of the parameters α', β' and γ'.

So, assuming we have a large enough block size, the block maxima themselves approximately follow a GEV distribution. This means we can try fitting a GEV distribution directly to the block maxima by, for example, using maximum likelihood estimation.

Question

Starting with the result that, for large enough n:

$$\frac{X_M - \alpha_n}{\beta_n} \div GEV(\alpha, \beta, \gamma)$$

show that $X_M \div GEV(\alpha', \beta', \gamma')$, stating the values of α', β' and γ', by considering the CDFs of $\frac{X_M - \alpha_n}{\beta_n}$ and X_M or otherwise.

Solution

Our starting point is the result in the question:

$$\frac{X_M-\alpha_n}{\beta_n} \div GEV(\alpha,\beta,\gamma)$$

We need to consider the two cases $\gamma=0$ and $\gamma\neq 0$.

If $\gamma\neq 0$, then:

$$P\left(\frac{X_M-\alpha_n}{\beta_n}\le x\right)\approx \exp\left(-\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\frac{1}{\gamma}}\right)$$

So:

$$\begin{aligned}P\left(X_M\le x\right)&=P\left(\frac{X_M-\alpha_n}{\beta_n}\le\frac{x-\alpha_n}{\beta_n}\right)\\&\approx \exp\left(-\left(1+\frac{\gamma\left(\frac{x-\alpha_n}{\beta_n}-\alpha\right)}{\beta}\right)^{-\frac{1}{\gamma}}\right)\\&=\exp\left(-\left(1+\frac{\gamma\left(x-\alpha_n-\beta_n\alpha\right)}{\beta_n\beta}\right)^{-\frac{1}{\gamma}}\right)\\&=\exp\left(-\left(1+\frac{\gamma'\left(x-\alpha'\right)}{\beta'}\right)^{-\frac{1}{\gamma'}}\right)\end{aligned}$$

This is the CDF of the $GEV(\alpha',\beta',\gamma')$ distribution, meaning that $X_M \div GEV(\alpha',\beta',\gamma')$ where:

$$\alpha'=\alpha_n+\beta_n\alpha$$

$$\beta'=\beta_n\beta$$

$$\gamma'=\gamma$$

If $\gamma=0$, then:

$$P\left(\frac{X_M-\alpha_n}{\beta_n}\le x\right)\approx \exp\left(-\exp\left(-\frac{(x-\alpha)}{\beta}\right)\right)$$

So:

$$P(X_M \le x) = P\left(\frac{X_M - \alpha_n}{\beta_n} \le \frac{x - \alpha_n}{\beta_n}\right)$$

$$\approx \exp\left(-\exp\left(-\frac{\left(\frac{x-\alpha_n}{\beta_n} - \alpha\right)}{\beta}\right)\right)$$

$$= \exp\left(-\exp\left(-\frac{(x - \alpha_n - \beta_n \alpha)}{\beta_n \beta}\right)\right)$$

$$= \exp\left(-\exp\left(-\frac{(x - \alpha')}{\beta'}\right)\right)$$

This is the CDF of the $GEV(\alpha', \beta', \gamma')$ distribution, meaning that $X_M \div GEV(\alpha', \beta', \gamma')$ where:

$$\alpha' = \alpha_n + \beta_n \alpha$$

$$\beta' = \beta_n \beta$$

$$\gamma' = 0$$

Note that in both cases the α and β parameters undergo the same transformations and the γ parameter remains unchanged.

For example, suppose we have monthly claim data stored in a data frame `data` with the first column `month`, representing the number of months since the start of the investigation, and the second column `claim`.

To calculate the block maxima for these claims using block sizes of 12 months, we would use the following R code:

```
data$block <- (data$month-1) %/% 12 + 1
blockmax <- aggregate(claim ~ block, data, max)
```

We can plot a histogram of the block maxima using the `hist()` function and an empirical density function using `density()` in the `plot()` function (if there is enough data). We can then superimpose a GEV distribution to see if it is a good approximation.

```
GEV <- function(x,alpha,beta,gamma){
1/beta*(1+gamma*(x-alpha)/beta)^-
(1+1/gamma)*exp(-((1+gamma*(x-alpha)/beta)^(-1/gamma))) }
lines(<sequence of x values>,GEV(<sequence of x
values>,<alpha>,<beta>,<gamma>))
```

The `qqplot()` function is used to compare the sample data to simulated values from a fitted GEV model.

We can estimate the maximum likelihood values as we did in Chapter 15 by defining a function that calculates the negative of the log-likelihood and using the function `nlm()` on this function as before.

Question

In the question in Section 2.1, the block maximum, X_M took the values:

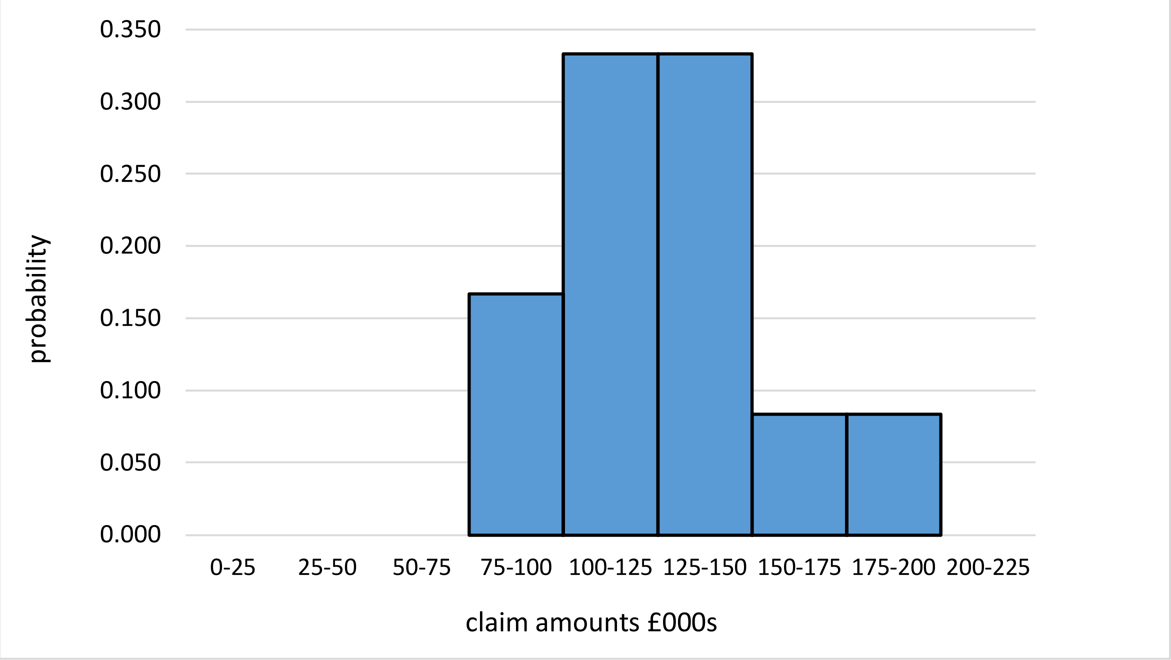
$$\{102, 152, 147, 128, 145, 113, 84, 140, 185, 118, 94, 104\}$$

when the block size was 5.

Plot these points on a frequency diagram and suggest a type of GEV distribution that might be appropriate.

Solution

A frequency diagram representing the above block maximum data is given below:



The data set is too small to be able to suggest a type of GEV distribution with any confidence. However, the upper tail decay appears to be rapid, which might lead us to consider a Gumbel-type GEV distribution.

This page has been left blank so that you can easily put in the replacement pages.