Anna Ojrzyńska

Uniwersytet Ekonomiczny w Katowicach

RODZINA MODELI LEE-CARTERA

Wprowadzenie

Publikacja modelu umieralności Lee-Cartera w 1992 r. zapoczątkowała ponowne zainteresowanie prognozowaniem umieralności. Od tamtego czasu opracowano kilka innych metod, jednakże metodologia zaproponowana przez Lee oraz Cartera jest nadal jedną z najlepiej dostępnych i powszechnie stosowanych. Autorzy publikacji z 1992 r. zaproponowali model opisujący zmiany w umieralności z uwzględnieniem zmieniającego się czasu. W metodzie tej logarytm współczynnika zgonów jest równy sumie dwóch składników, z których jeden nie zależy od czasu, a drugi jest iloczynem parametru pokazującego ogólny poziom umieralności oraz parametru, który wskazuje, jak szybko lub wolno zmienia się umieralność w danym wieku w zależności od zmian ogólnej umieralności w czasie. Parametry tego modelu są estymowane na podstawie danych historycznych.

Modyfikację do tego modelu zaproponowali m.in.: Lee i Miller [2001], Booth, Maindonald i Smith [2002], Hyndman i Ullah [2005] oraz De Jong [2006].

Celem artykułu jest przedstawienie wybranych modyfikacji klasycznego modelu Lee-Cartera oraz zastosowanie ich do szacowania współczynnika zgonów w Polsce. Na podstawie danych o współczynnikach zgonów dla jednorocznych grup wiekowych kobiet i mężczyzn w latach 1990-2005 dla Polski zostaną oszacowane parametry modelu umieralności Lee-Cartera oraz wybranych modyfikacji tego modelu. Następnie zostanie oceniona dobroć dopasowania modeli, a tym samym zostanie zweryfikowana możliwość użycia tych modeli do prognozowania umieralności w Polsce.

1. Metodologia badawcza

W latach 90. Lee i Carter podjęli próbę zastosowania teorii procesu błądzenia przypadkowego z dryfem do modelowania oraz prognozowania współczynników zgonów $m_{x,t}$ w grupach wieku x i dla kolejnych lat kalendarzowych t. Model ten ma postać [Rossa, 2009]:

$$\ln m_{x,t} = a_x + b_x k_t + \varepsilon_{x,t}, \quad (1)$$

gdzie:

a_x − średni poziom logarytmu współczynnika zgonów w poszczególnych grupach wieku, uśredniony względem czasu kalendarzowego,

b_x − wskazują, jak szybko logarytmy cząstkowych współczynników zgonów, tj. $\ln m_{x,t}$, zmieniają w odpowiednich grupach wieku x,

k_t − opisuje ogólną tendencję zmian umieralności w ciągu badanego okresu, traktowany jest efekt wpływu czasu kalendarzowego t na zmianę w poziomie cząstkowych współczynników zgonów,

$\varepsilon_{x,t}$ − niezależne składniki losowe, o rozkładach normalnych o wartości oczekiwanej równej 0 i stałej wariancji σ^2.

Dla zapewnienia jednoznaczności rozwiązania Lee i Carter przyjęli dodatkowe warunki ograniczające [Papież, 2008], tj. że suma parametrów b_x dla wszystkich grup wieku jest równa 1, natomiast suma parametrów k_t jest równa 0.

W propozycji Lee-Cartera szereg czasowy k_t jest traktowany jako wycinek procesu stochastycznego błądzenia przypadkowego z dryfem, opisanym formułą [Rossa, 2009]:

$$k_t = c + k_{t-1} + e_t, \quad (2)$$

gdzie:

c − reprezentuje pewną stałą (dryf),

e_t − składnik losowy o rozkładzie normalnym z wartością oczekiwaną równą 0 i pewną skończona wariancją.

Model zaproponowany przez Lee i Millera jest modyfikacją modelu Lee-Cartera i różni się od jego pierwotnej postaci pod trzema względami [Lee, Miller, 2001]:

1. Zakres czasowy danych, na podstawie których szacowano parametry modelu uległ skróceniu – były to lata 1950-1989.
2. Dokonano korekty parametru k_t, polegającej na dopasowaniu oszacowań tego parametru do oczekiwanej długości życia noworodka w roku t (e_0 jest jednym z parametrów tablic trwania życia).
3. Do wyznaczenia prognoz wartości współczynników zgonu jako wartości teoretycznych dla ostatniego okresu szacowania przyjmuje się wartości rzeczywiste (empiryczne).

W ocenie autorów tej modyfikacji głównym źródłem wysokich błędów prognoz, wyznaczonych przez Lee i Cartera było duże niedopasowanie pomiędzy wartościami teoretycznymi dla ostatniego okresu szacowania modelu (tj. dla

1989 r.) z wartościami rzeczywistymi (empirycznymi) w tym roku. W modelu Lee-Milera rozwiązaniem tego problemu jest zastosowanie dodatkowego warunku, iż parametr k_t przyjmuje wartość 0 w ostatnim roku szacowania modelu.

Analizując współczynniki zgonów w latach 1900-1995, autorzy zauważyli również, że wzorzec umieralności nie jest stały w czasie. W związku z tym szacowania modelu umieralności dokonano na podstawie danych z lat 1950-1989.

Model zaproponowany przez Booth-Maindonalda-Smitha również różni się od klasycznego modelu Lee-Cartera w trzech kwestiach [Booth, Maindoinald, Smith, 2002]:

1. Okres, na podstawie którego są szacowane parametry modelu jest dobierany na podstawie statystycznych kryteriów dobroci dopasowania, przy założeniu liniowości parametru k_t.
2. Korekty parametru k_t dokonuje się na podstawie rozkładu liczby zgonów wg wieku w poszczególnych latach, z wykorzystaniem własności rozkładu Poissona.
3. Nie dokonuje się żadnych zmian w wartościach teoretycznych dla ostatniego okresu szacowania.

Metoda Hyndmana i Ullaha wykorzystuje natomiast analizę funkcjonalną do modelowania logarytmów współczynników zgonów. Jest ona rozszerzeniem modelu Lee-Cartera w następujących kwestiach [Booth, Hyndman, Tickle, de Jong, 2006]:

1. Zakłada się że umieralność to gładka funkcja wieku, która jest obserwowana z błędami; gładkie współczynniki zgonów są szacowane za pomocą nieparametrycznych metod wygładzania.
2. Wykorzystuje więcej niż jeden zestaw składników (k_t, b_x).
3. Do prognozowania parametrów modelu są wykorzystywane bardziej ogólne metody szeregów czasowych niż błądzenie losowe z dryfem. Do wykładniczego wygładzania zastosowano modele przestrzeni stanów.
4. Zastosowanie odpornych metod estymacji pozwoli na dokładniejsze oszacowania współczynników zgonu dla nietypowych lat m.in. w okresach konfliktów zbrojnych, epidemii.
5. Brak korekty parametru k_t.

Podejście Hyndmana i Ullaha może być wyrażone za pomocą równania o postaci:

$$\ln m_{x,t} = a(x) + \sum_{j=1}^{J} k_{t,j} b_j(x) + e_t(x) + \sigma_t(x)\varepsilon_{x,t}, \qquad (3)$$

gdzie:
$a(x)$ − średni wzorzec umieralności w wieku x,
$b_j(x)$ − jest podstawową funkcją,
k_t − parametr szeregu czasowego,
$\sigma_t(x)\varepsilon_{x,t}$ − błąd obserwacji, wynikający z różnicy między obserwowanymi wskaźnikami a krzywymi sklejanymi,
$e_t(x)$ − błąd szacowania, wynikający z różnicy między krzywymi sklejanymi i ich oszacowanymi odpowiednikami.

Ostatnią z przedstawionych w tym referacie modyfikacji modelu Lee-Cartera jest podejście De Jonga i Tickle'a, które wykorzystuje metodologię przestrzeni stanów do modelowania logarytmów współczynników zgonu. Modele przestrzeni stanów obejmują szeroki zakres elastycznych wielowymiarowych modeli szeregów czasowych, których model Lee-Cartera jest szczególnym przypadkiem. Ogólna metodologia dopuszcza mnóstwo specjalizacji i uogólnień oraz zawiera ocenę oszacowania nieznanych parametrów, wnioskowanie i prognozowanie łącznie z obliczeniem błędów prognoz. Model Lee-Cartera może być zapisany w postaci równania:

$$y_t = a + bk_t + \varepsilon_t, \tag{4}$$

gdzie:
y_t − wektor logarytmów współczynników zgonu w każdym wieku w roku t,
a *i* b − wektory odpowiednich parametrów modelu Lee-Cartera dla każdego wieku,
k_t − jest wskaźnikiem poziomu umieralności w roku t, tak jak w modelu Lee-Cartera,
ε_t − wektor błędów w każdym wieku w roku t.

W 2006 r. De Jong i Tickle opracowali bardziej ogólną specyfikację [Booth, Hyndman, Tickle, de Jong, 2006]:

$$y_t = Xa + Xbk_t + \varepsilon_t, \tag{5}$$

gdzie:
X jest macierzą o liczbie wierszy większej niż liczba kolumn.

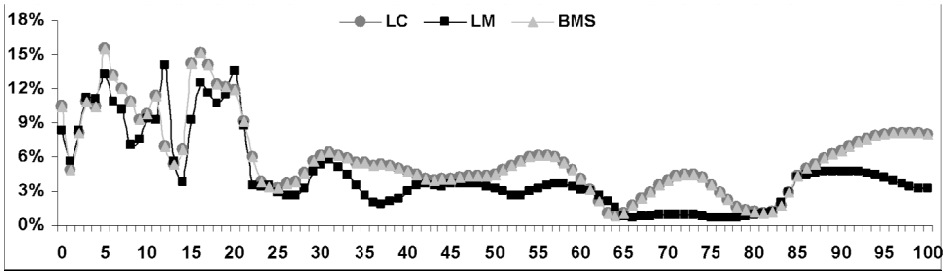
Gdy $X = I$, wówczas model redukuje się do postaci równania (4). Równanie (5) rozwiązuje problem modelu LC zapisanego równaniem (4), gdzie istnieje parametr a i b dla każdego wieku. W modelu (5) macierz X ma mniej kolumn niż wierszy, co oznacza, że wektorów parametrów a i b jest mniej niż istniejących grup wiekowych. Efekty szeregu czasowego k_t nie są niezależne we wszystkich grupach wiekowych, ale są ograniczone przez strukturę macierzy X, nakładając gładkość w każdej grupie wiekowej.

W obecnej analizie macierz X jest oparta na krzywych B-sklejanych, w których obowiązuje forma kwadratowa logarytmu współczynnika zgonu między węzłami w różnych grupach wieku. Oszacowania modelu metodą największej są obliczone przy użyciu filtrowania Kalmana i wygładzania.

2. Analiza empiryczna

Badanie umieralności przeprowadzono na podstawie danych dotyczących współczynników natężenia zgonów oraz stanu ludności według wieku w Polsce w latach 1990-2010. Estymację parametrów przeprowadzono zarówno dla kobiet, jak i dla mężczyzn na podstawie lat 1990-2005. Okres 2006-2010 posłużył natomiast do wyznaczenia prognoz oraz weryfikacji modelu.

Do oceny prognoz wykorzystano średni bezwzględny błąd procentowy MAPE[1] oraz średni błąd procentowy MPE[2]. Obliczone wartości błędu MAPE pozwolą porównać dokładność prognoz otrzymywanych z wykorzystaniem modyfikacji modelu Lee-Cartera. Wartości błędów MPE pozwolą natomiast ocenić czy wyznaczone prognozy przeszacowują lub niedoszacowują rzeczywiste wartości współczynników zgonu. Rysunki 1 i 3 prezentują wartości średnich bezwzględnych błędów procentowych prognoz dla jednorocznych grup wiekowych mężczyzn i kobiet.



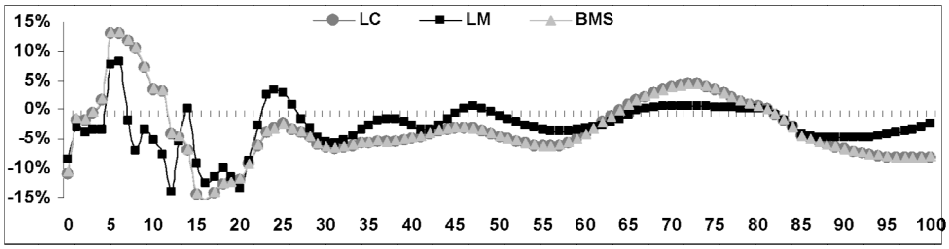
Rys. 1. Średnie bezwzględne błędy procentowe prognoz współczynnika zgonu według wieku mężczyzn

Prognozy umieralności mężczyzn cechuje większa dokładność aniżeli w przypadku prognoz wyznaczonych dla kobiet. Błędy dla prognoz umieralności mężczyzn poniżej 20 roku życia nie przekraczają 14%. W starszych grupach

[1] $$MAPE = \frac{1}{m}\sum_{\tau}^{m}\left|\frac{y_{\tau}^{P} - y_{\tau}}{y_{\tau}}\right|$$

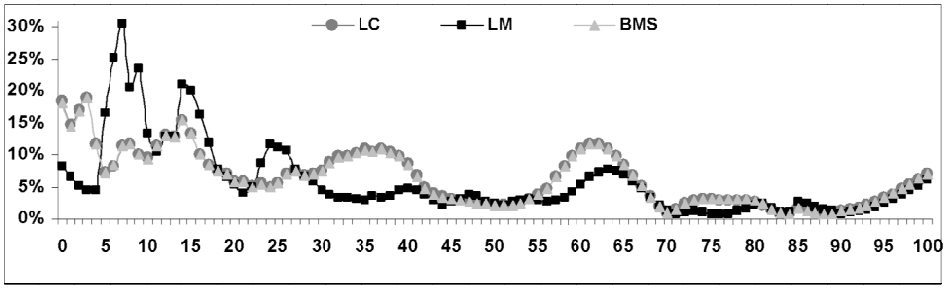
[2] $$MPE = \frac{1}{m}\sum_{\tau}^{m}\frac{y_{\tau}^{P} - y_{\tau}}{y_{\tau}}$$

wiekowych, tzn. między 20-85 rokiem życia, oscylują natomiast wokół poziomu 4%. Z wyjątkiem mężczyzn w wieku 12 i 20 lat prognozy wyznaczone na podstawie modelu Lee-Milera były obarczone niższymi błędami MAPE. Prognozy obliczone natomiast na podstawie modelu Lee-Cartera oraz modelu Booth-Maindonalda-Smitha były bardzo zbliżone. Analizując średnie błędy prognoz dla mężczyzn (rys. 2), można zauważyć, że częściej prognozy te niedoszacowują rzeczywistych wartości współczynnika zgonu według wieku.



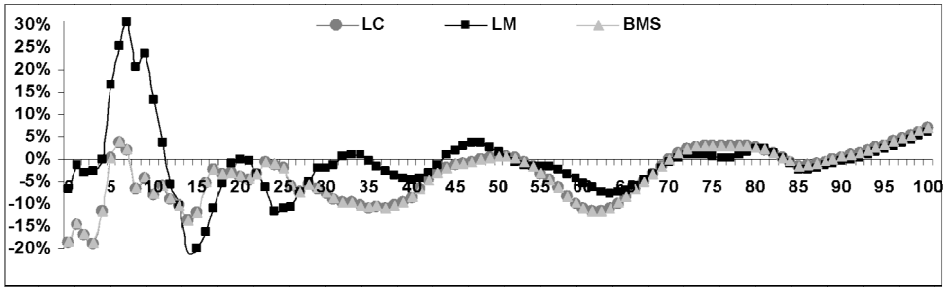
Rys. 2. Średnie błędy procentowe prognoz współczynnika zgonu według wieku mężczyzn

W przypadku kobiet prognozy umieralności są obarczone dość dużymi MAPE, szczególnie jest to widoczne w najmłodszych grupach wiekowych. Dla kobiet powyżej 30 roku życia model Lee-Milera charakteryzuje się najlepszym dopasowaniem. W młodszych grupach wieku nie można jednoznacznie określić, który z przedstawionych modeli zapewnia najdokładniejsze prognozy, przykładowo dla kobiet poniżej 4 roku życia jest nim model Lee-Milera, natomiast w wieku 5-12, 15-18, 23-26 modele Lee-Cartera oraz Booth-Maindonalda-Smitha.



Rys. 3. Średnie bezwzględne błędy procentowe prognoz współczynnika zgonu według wieku kobiet

Analizując średnie błędy prognoz dla kobiet, można zauważyć, że częściej prognozy te niedoszacowują rzeczywistych wartości współczynnika zgonu według wieku. Tylko w przypadku prognoz wyznaczonych na podstawie modelu Lee-Milera dla kobiet w wieku 6-12 lat oraz kobiet w najstarszym wieku (dla wszystkich trzech modeli) można zaobserwować, iż prognozy te przeszacowują rzeczywiste współczynniki zgonu.



Rys. 4. Średnie błędy procentowe prognoz współczynnika zgonu według wieku kobiet

Dodatkowo obliczono średni bezwzględny błąd procentowy łącznie we wszystkich grupach wiekowych $MAPE_t^m$ oraz MPE_t^m, a wyniki przedstawiono w tab. 1.

Tabela 1

Błędy prognoz ex post umieralności kobiet i mężczyzn

	MAPE			MPE		
	LC	LM	BMS	LC	LM	BMS
mężczyźni	5,97%	4,37%	5,90%	-3,26%	-2,83%	-3,10%
kobiety	6,34%	5,60%	6,31%	-3,75%	-0,81%	-3,64%

Obliczone średnie błędy prognoz potwierdzają, iż najlepszym dopasowaniem do danych charakteryzowała się modyfikacja modelu umieralności zaprezentowana przez Lee i Milera. Wykorzystując właśnie ten model do prognozowania umieralności w latach 2006-2010, można spodziewać się błędów na poziomie 4,3% dla mężczyzn i 5,6% dla kobiet.

Podsumowanie

Przeprowadzona w tej pracy próba prognozowania współczynnika zgonów została oparta na trzech modelach należących do rodziny modeli Lee-Cartera. Pierwszym z modeli była klasyczna i pierwotna metodologia zaproponowana przez Lee i Cartera w 1992 r. Drugim z modeli była modyfikacja zaproponowana przez Lee i Milera. Ostatnim przedstawionym modelem był model Booth-Maindonalda-Smitha.

Wyniki badań wskazały, że zaproponowana modyfikacja klasycznego modelu Lee-Cartera przedstawiona przez Lee i Milera poprawia jakość dopasowania modelu do danych empirycznych dla kobiet i mężczyzn powyżej 30 roku życia. Przeprowadzona analiza pozwala twierdzić, że model Lee-Cartera można

wykorzystywać do szacowania umieralności w Polsce. Należy jednak rozważać też inne modyfikacje klasycznego modelu, które mogą dać bardziej wiarygodne wyniki prognoz umieralności.

Literatura

Booth H., Maindonald J., Smith L. (2002): *Applying Lee-Carter under Conditions of Variable Mortality Decline*. „Population Studies", No. 56.

Booth H., Hyndman R.J., Tickle L., de Jong P. (2006): *Lee-Carter Mortality Forecasting: A Multi-country Comparison of Variants and Extensions*. „Demographic Research", No. 15.

Hyndman R.J., Ullah M.S. (2005): *Robust Forecasting of Morality and Fertility Rates: A Functional Data Approach.* Working paper 2105, Department of Econometrics and Business Statistics, Moriah University.

De Jong, Tickle L. (2006): *Extending Lee-Carter Method for Forecasting Morality.* „Demography" 2001, 38(4).

Lee R.D., Carter L. (1992): *Modelling and Forecasting US Mortality*. „Journal of the American Statistical Association", Vol. 87(419).

Lee R.D., Miller T. (2001): *Evaluating the Performance of the Lee-Carter Method for Forecasting Mortality*. „Demography", 38(4).

Papież M. (2008): *Możliwość wykorzystania modelu Lee-Cartera do szacowania wartości w dynamicznych tablicach trwania życia*. Zeszyty Naukowe SAD PAN, nr 18.

Rossa A. (2009): *Dynamiczne tablice trwania życia oparte na metodologii Lee-Cartera i ich zastosowanie do obliczania wysokości świadczeń emerytalnych*. Acta Universitatis Lodziensis. Folia Oeconomica, 231.

FAMILY OF LEE-CARTER MODELS

Summary

This paper presents a proposal for the application of selected models of the group of models using the Lee and Carter methodology for forecasting mortality rates. These include the original Lee-Carter, the Lee-Miller (2001) and Booth-Maindonald-Smith (2002) variants, and the more flexible Hyndman-Ullah (2005) and de Jong (2006) extensions. Based on estimates of mortality rates derived from the selected models was verified the ability to use these models to estimate mortality rates in Poland.