Journal of Statistical and Econometric Methods, vol.3, no.2, 2014, 15-30
ISSN: 2241-0384 (print), 2241-0376 (online)
Scienpress Ltd, 2014

Testing 'Clemmensen's hook' in the death rate from breast cancer

Gustav N. Kristensen[1]

Abstract

Breast cancer is the most common form of cancer among women. Since 1945 the number of cases has almost doubled. However, since 1990 the survival rate has improved significantly.
The literature discusses models of the age distribution of breast cancer mortality developed on the basis of a two-disease theory of breast cancer incidence called Clemmensen's hook.
On the basis of Danish data on the death rate from breast cancer the existence of Clemmensen's hook is discussed critically.
A model encompassing five discussed studies on the subject is developed. When the cohort effects are included in the model, the two-disease theory for breast cancer disappears.
The paper concludes that Clemmensen's hook does not exist as the overlapping of two curves corresponding to pre- and post-menopausal tumors for breast cancer, respectively.

Mathematics Subject Classification: Statistics; numerical analysis; special functions

Keywords: Cohort effects; Clemmensen's hook; breast cancer; age-period-cohort model

[1] University of Southern Denmark. E-mail: guk@sam.sdu.dk

Article Info: *Received* : April 1, 2014. *Revised* : May 2, 2014.
Published online : May 15, 2014.

1 Introduction

Breast cancer is the most common form of cancer among women. Since 1945 the number of cases has doubled in Denmark. However, since 1990 the survival rate has improved significantly in the western world. Today, in summary, about 85% of the breast cancer patients are still free from breast cancer five years after treatment.

The purpose of this article is to create an econometric model to make a prognosis for the development in the death rate from breast cancer. In that context, the literature has argued for the existence of a Clemmensen’s hook in the modeling of death rates from breast cancer over lifetime. The model developed here questions its existence.

Manton and Stallard [10] interpret Clemmensen’s hook as a result of breast cancer being two separate diseases. The curve indicating the death rate for breast cancer is thus interpreted as the overlapping of two curves corresponding to pre- and post-menopausal tumors, respectively.

Clayton and Schifflers [3, 4] talk about Clemmensen’s hook as follows: “The age curve [of death rates] shows the phenomenon of Clemmensen’s hook; rates increase to a maximum at 50-54 then fall back slightly before continuing their upward trend from the age of 65 onwards”.

Likewise, Cayuela et al. [2] mention that Clemmensen’s hook “has been observed in different countries with reference to both incidence and mortality and is interpreted as the overlapping of two curves corresponding to pre- and post-menopausal tumors, respectively.

However, in more recent studies we see that Clemmensen’s hook disappears.

Bouchardy et al. [1] write that “the typical age incidence curve of breast cancer described a progressive increase of risk with age, with a slope down around the menopause age, called Clemmensen’s hook. This typical curve by age is no longer observed in Geneva …”.

Similarly, Fuglede et al. [5] conclude that “important changes over the past decade in the age-specific incidence pattern of breast cancer in particular around the time of menopause [commonly denoted Clemmensen’s hook] were indicated.”

For a better description of the development of breast cancer over the lifecycle of women this article will present an econometric age-period-cohort model to test the existence of a Clemmensen’s hook in the death rate from breast cancer based on Danish data. The method is based on Kristensen [9] including secular cohort effects. For a discussion of the age-period-cohort models see also Clayton and Schifflers [3, 4], Holford [7], Osmond and Gardner [11], and Rostgaard et al. [12].

2 Data

The Danish data on the *death rate* from breast cancer are obtained from: The Danish Health and Medicines Authority (Statens Serum Institut): "Cancer in breast", B-020. In principle, the present article is based on the total dataset for deaths from breast cancer in Denmark 1977-2012. The explanatory variables are:

T period (or year), 1960 = 1

Age age at death

Dbc actual death rate from breast cancer. Dbc is shown in Figure 1.

CohBorn cohort indicated by the year of birth of the youngest person(s) in an age group. The cohort is followed by a dummy equal to one following an age group diagonal over period and age.

B_{Age} vector for age specific cohort coefficients that express protective and detrimental effects according to year of birth. This vector can with good reason be seen as the "susceptibility parameter s as systematically varying with birth cohorts" as mentioned by Manton and Stallard [10].

The article applies 5-year age groups. From the age group 30-34 to the age group 85+ there are no years in which the death rate is zero in the Danish data. Simplified, 30 indicates the age group 30-34, and similarly 35 indicates the age group 35-39, etc. In the empirical estimations the youngest included cohort 30-34 (where the youngest member was 30 years old) started in 1977. The age-specific death rate from breast cancer is shown in Figure 1.

Looking back, the death rate was increasing from 1945 onwards. For women in the age group 55-59 the highest death rate was reached in 1992. That is, the death rate in that age group was (in the data set) increasing in the period 1977-1992 and declining from 1992. For women in the age group 65-69 the highest death rate was reached in 2000 and started declining after 2000. No real decline is seen in the age groups above 80.

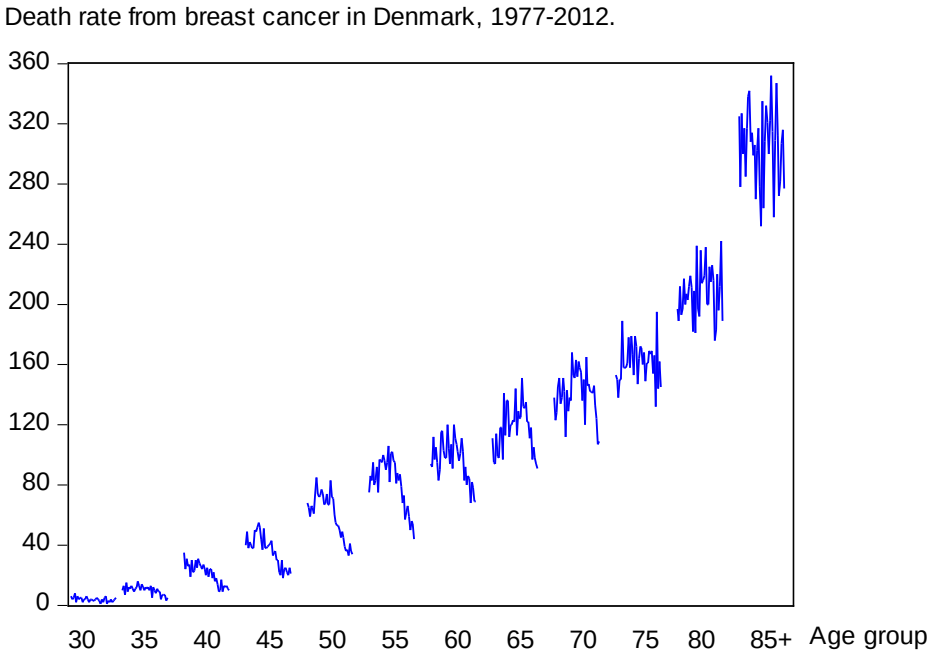

Figure 1: Death rate from breast cancer in Denmark 1977-2012 distributed into 5-year age groups from 30-34 to 85+.

3 Model for death rates

The above- mentioned articles talk about the death rate from breast cancer being exponentially increasing with age. Simplified and in the present notation that is:

$$Dbc = \alpha e^{\alpha_1 Age} \tag{1}$$

The “new mortality trend” (Gavrilova and Gavrilov, [6]), which lets the death rate curve at all ages decline with almost the same percentage, was added to equation (1) in a semi- logarithmic version:

$$\text{Log}(Dbc) = \alpha_0 + \alpha_1 \text{ Age} - \alpha_2 T \tag{2}$$

This formula was the basis for developing equation (3).
The equation for the death rate from breast cancer including the cohort effects is:

$$\begin{aligned}\text{Log(Dbc)} = {} & \alpha_1/\text{Age} + \alpha_2 \text{Age} + \alpha_3 \text{Age}^2 \\ & + \alpha_4/\text{T} + \alpha_5 \text{Age/T}^2 + \alpha_6 \text{Age}^2/\text{T}^3 \\ & + \beta_1 \text{Coh1892} + \beta_2 \text{Coh1893} + \ldots\ldots + \beta_{91} \text{Coh1982} \end{aligned} \quad (3)$$

There is no constant element (origo regression), and therefore we can use dummies for the entire period 1892-1982. The equation was estimated by WLS using Age as weight.

$$\begin{aligned}\text{Log(Dbc)*Age} = {} & \alpha_1 + \alpha_2 \text{Age}^2 + \alpha_3 \text{Age}^3 \\ & + \alpha_4 \text{Age/T} + \alpha_5 \text{Age}^2/\text{T}^2 + \alpha_6 \text{Age}^3/\text{T}^3 \\ & + \beta_1 \text{Coh1892*Age} + \beta_2 \text{Coh1893*Age} + \ldots. \\ & .+ \beta_{91} \text{Coh1982*Age} \end{aligned} \quad (4)$$

A new time series, formed by the beta coefficients in (4) from β_1 to β_{91}, in total 91 observations, is shown in Figure 2. B $\in \{\beta_1, \beta_2, \beta_3 \ldots\ldots\ldots \beta_{91}\}$.

The cohort coefficients relating to earlier and recent periods are based on fewer age-specific rates and are hence less reliable than in the central period. Thus, we see the central period 1932-1952 as having the lowest variance.

The beta coefficients (the cohort effects) can be seen as related to age groups as shown below in Table 1. Applied for forecasting, the *middle* of Table 1 shows in the estimated.

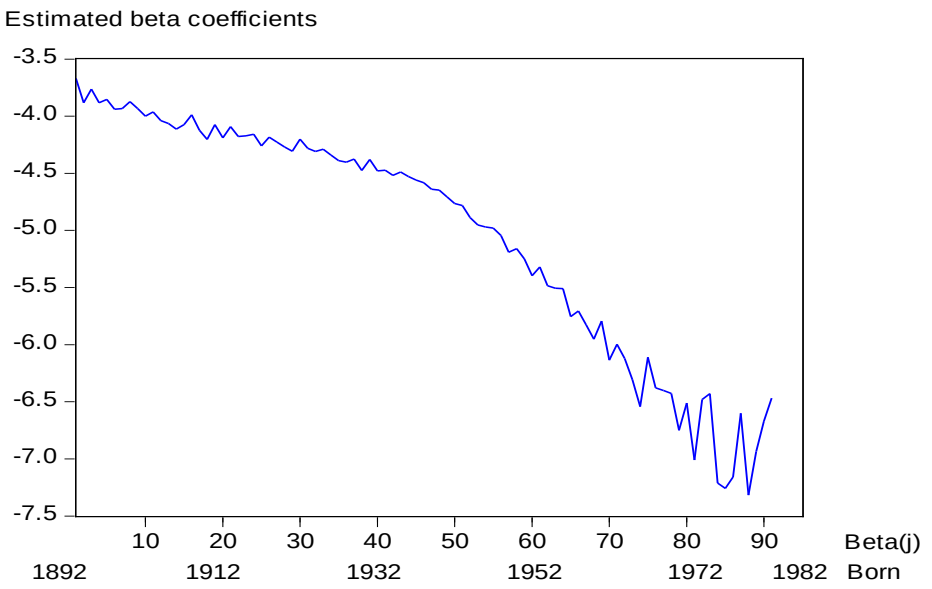

Figure 2: The estimated beta coefficients from equation (4) seen as a time series.

Table 1. The β coefficients seen as time series variables

Age	**30**	**35**	**40**	**:::**	**65**	**70**	**75**	**80**	**85 +**
Year\B *Age*	**B** 30	**B** 35	**B** 40	:::	**B** 65	**B** 70	**B** 75	**B** 80	**B** 85
Ex post fc									
1965	β 44	β 39	β 34	:::	β_9	β_4	**na**	**na**	**na**
:::									
1971	β 50	β 45	β 40	:::	β 15	β 10	β_5	**na**	**na**
1972	β 51	β 46	β 41	:::	β 16	β 11	β_6	β_1	**na**
1973	β 52	β 47	β 42	:::	β 17	β 12	β_7	β_2	**na**
1974	β 53	β 48	β 43	:::	β 18	β 13	β_8	β_3	**na**
1975	β 54	β 49	β 44	:::	β 19	β 14	β_9	β_4	**na**
1976	β 54	β 50	β 45	:::	β 20	β 15	β 10	β_5	**na**
Estimated									
1977	β 56	β 51	β 46	:::	β 21	β 16	β 11	β_6	β_1
1978	β 57	β 52	β 47	:::	β 22	β 17	β 12	β_7	β_2
1979	β 58	β 53	β 48	:::	β 23	β 18	β 13	β_8	β_3
1980	β 59	β 54	β 49	:::	β 24	β 19	β 14	β_9	β_4
:::	:::	:::	:::	:::	:::	:::	:::	:::	:::
2011	β	β	β	::	β	β	β	β	β

	β_{90}	β_{85}	β_{80}	:	β_{55}	β_{50}	β_{45}	β_{40}	β_{35}
2012	β_{91}	β_{86}	β_{81}	:: :	β_{56}	β_{51}	β_{46}	β_{41}	β_{36}
Ex ante fc									
2013	**na**	β_{87}	β_{82}	:: :	β_{57}	β_{52}	β_{47}	β_{42}	β_{37}
2014	**na**	β_{88}	β_{83}	:: :	β_{58}	β_{53}	β_{48}	β_{43}	β_{38}
2015	**na**	β_{89}	β_{84}	:: :	β_{59}	β_{54}	β_{49}	β_{44}	β_{39}
2016	**na**	β_{90}	β_{85}	:: :	β_{60}	β_{55}	β_{50}	β_{45}	β_{40}
2017	**na**	β_{91}	β_{86}	:: :	β_{61}	β_{56}	β_{51}	β_{46}	β_{41}
2018	**na**	**na**	β_{87}	:: :	β_{62}	β_{57}	β_{52}	β_{47}	β_{42}
::::									
2022	**na**	**na**	β_{91}	:: :	β_{66}	β_{61}	β_{56}	β_{51}	β_{46}

beta coefficients. *Above* "Ex post fc" shows the estimated coefficients as applied in ex post forecasts. *Below* "Ex ante fc" shows the estimated coefficients applied in ex ante forecasts.

We cannot say anything for $\beta_x > \beta_{91}$ or $\beta_y < \beta_1$. However, for the individual age groups we can, within limits, make ex ante and ex post forecasts. The origo (English: origin) for the cohort coefficients is age group 85+ in 1977. For simplicity, we treat it as an age group 85-89.

From origo you can go back in time: 1977- 85 = 1892 (the oldest born 1888).

From 2012 the age group 30-34 reaches back to: 2012- 85 = 1927 (the oldest born 1922).

From origo you can go forward in time: 1977+85 = 2062 (the oldest born 2066).

Figure 3 shows the *estimated data* distributed on age groups, the middle group in Table 1. When ex post and ex ante beta values are included for the

individual age groups, all curves are identical for the individual age groups and equal to the beta coefficient curve in Figure 2.

4 Ex post and ex ante forecast

Forecasts become increasingly unsure when moving away from actual data. The ex-post forecast is therefore limited to 1972, and the ex-ante forecast is limited to 2022. Likewise, it is not convenient to apply very high (e.g. 2022) or negative values of T because T^2 and T^3 will then get different signs. Consequently, T=1 for 1960 was applied.

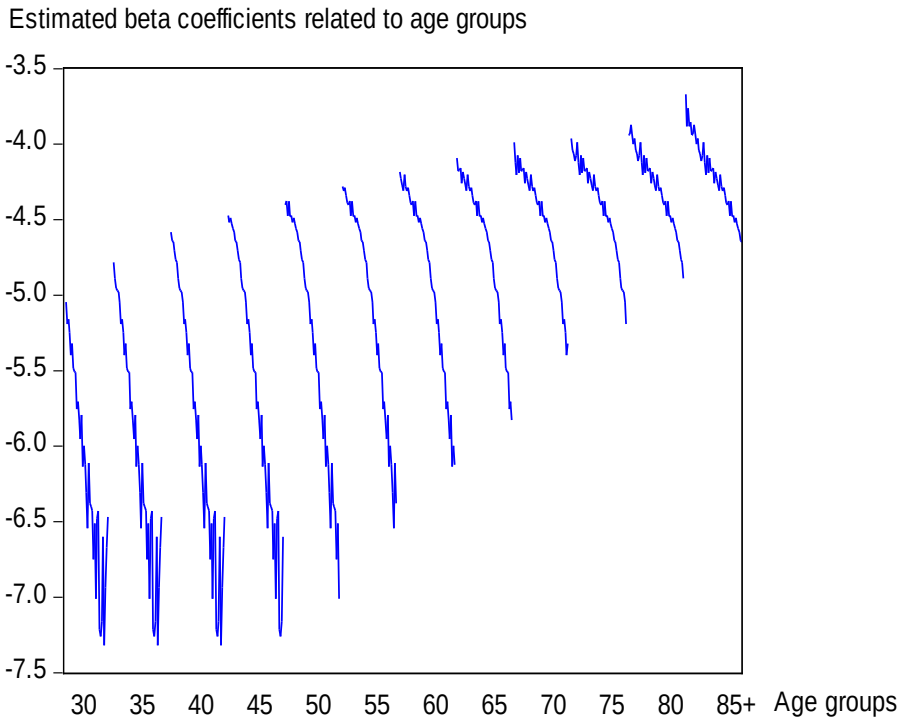

Figure 3: The estimated beta coefficients related to age groups

An irreversible decline in the death rate from breast cancer is estimated to start in 2014 for the age group 80-84, and in 2015 for the age group 85+.

4.1 The disappearance of Clemmensen's hook

In order to show more clearly the disappearance of Clemmensen's hook the curves in Figure 4 are smoothed out by only including every fifth year from1972 and forward to 2022. The outcome is shown in Figure 5.

Following the *'s in 1972 and the **o'**s in 2022 we see the disappearance of Clemmensen's hook. It is seen that the model result from equation (4) is in accordance with all the above- mentioned models in literature from 1980 to 2006. However, there are not two diseases. Clemmensen's hook is a result of the (former) lifestyle until 1932, where the beta-coefficient curve has its kink.

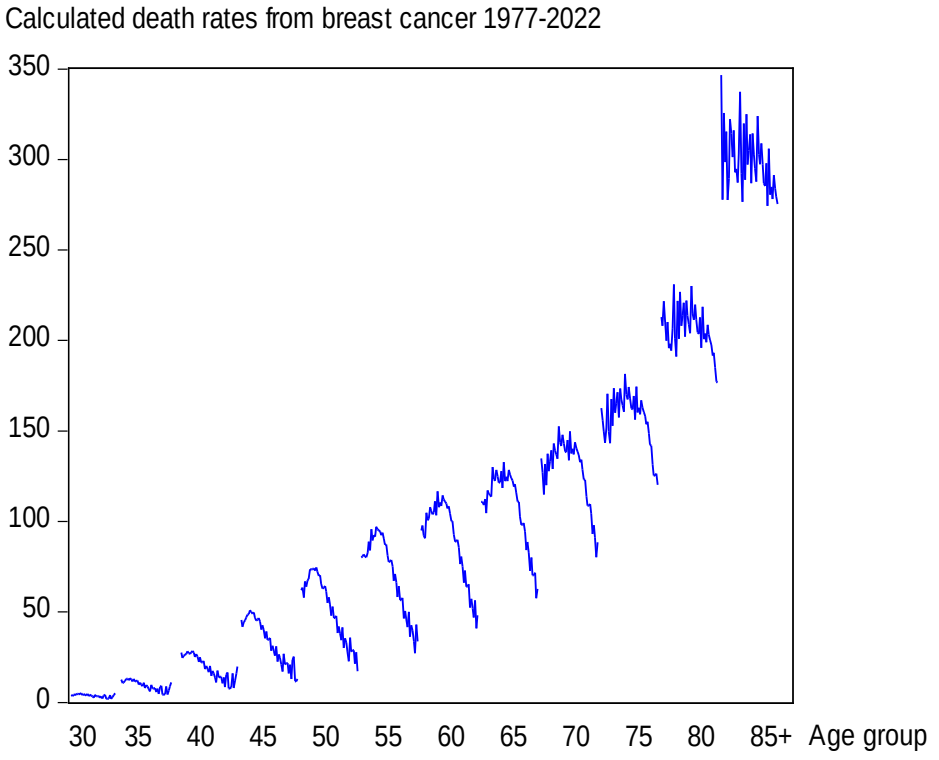

Figure 4: The calculated death rate 1977-2022

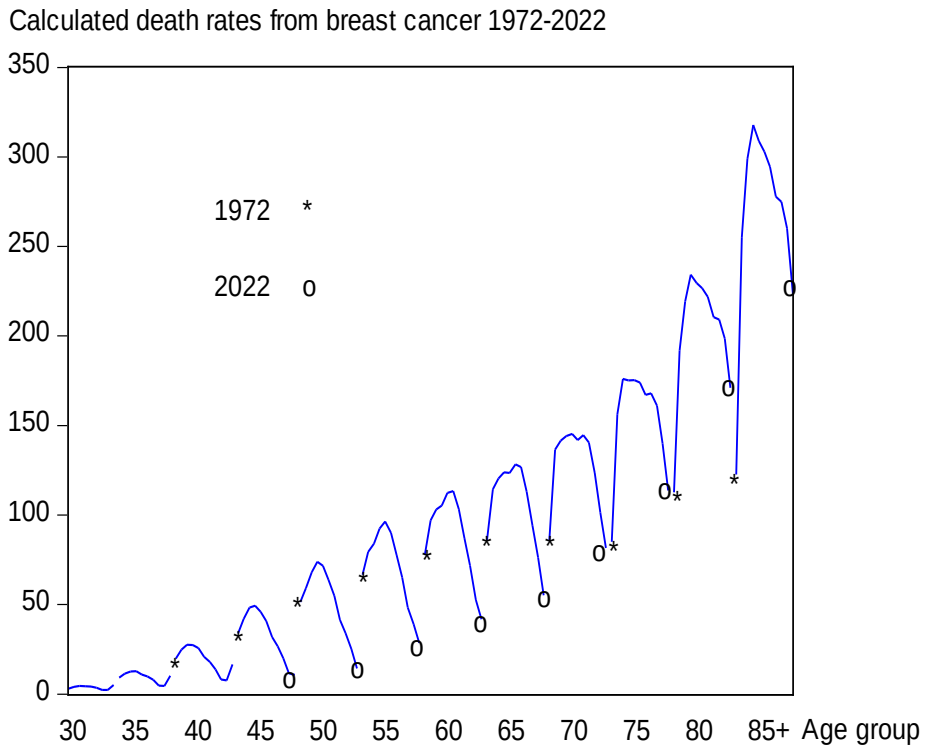

Figure 5: Stylistic calculation of the end of “Clemmensen’s hook” 1972-2022

5 Conclusion

The cases of breast cancer increased rapidly after the Second World War. However, from 1990 the progress in treatment was accelerating, and the death rate started to decline, first for the younger age groups, and will from 2015 include all age groups.

Clemmensen’s hook (according to this study) mirrors a shift in lifestyle among different generations or cohorts modeled by the cohort coefficients. The changing lifestyle forms a beta coefficient curve with a kink in 1932 that creates the impression of a “Clemmensen’s hook”.

Thus, Clemmensen’s hook does not indicate the existence of two diseases in breast cancer.

References

[1] C. Bouchardy, A. Morabia, H.M. Verkooijen, et al. Remarkably change in age-specific breast cancer incidence in the Swiss canton of Geneva and its possible relation with the use of hormone replacement therapy. *BMC Cancer*; **6**(78), (2006), 1-7.

[2] A. Cayuela, S. Rodriguez-Dominguez, M. Ruiz-Burrego, and M. Gili. Age-period-cohort analysis of breast cancer mortality rates in Andalucia (Spain). *Annals of Oncology*, **15**, (2004), 686-688.

[3] D. Clayton, E. Schifflers. Models for temporal variation in cancer rates. I. Age-period and age-cohort models. *Statistics in Medicine,* **6**, (1987), 449-467.

[4] D. Clayton, E. Schifflers. Models for temporal variation in cancer rates. II. The age-period-cohort model. *Statistics in Medicine,* **6**, (1987), 469-481.

[5] N. Fuglede, O. Langballe, A.L. Svendsen, et al. *Development in incidence of breast cancer in non-screened Danish women, 1973-2002* – a population-based study.

[6] N.S. Gavrilova and L.A. Gavrilov. Aging and Longevity: Mortality laws and mortality forecasts for aging populations. *Demografie,* **53**(2): (2011), 109-128.

[7] T.R. Holford. Understanding the effects of age, period, and cohort on incidence and mortality. *Annu. Rev. Publ. Health*, **12**, (1991), 425-457.

[8] G.N. Kristensen. The Vintage Waves in the Death Rate Data for Apoplexy. (Paper presented at the Symposium in Applied Statistics; pp 209-219, (2013a) University of Aarhus).

[9] G.N. Kristensen. Cohort Coefficients. Describing the secular development in protective and detrimental cohort effects associated with apoplexy. *Journal of Statistical and Econometric Methods.* **2**(4), (2013b), 119-127.

[10] K.G. Manton, and E. Stallard. A Two-Disease Model of Female Breast Cancer: Mortality in 1969 Among White Females in the United States. *Journal of the National Cancer Institute*, **64**(1), (1980), 9-16.

[11] C. Osmond, and M.J. Gardner. Age, Period and Cohort Models applied to Cancer Mortality Rates. *Statistics in Medicine*, **1**, (1982), 245-259.

[12] K. Rostgaard, M. Vath, H. Holst, M. Madsen, E. Lynge. Age-period-cohort modeling of breast cancer incidence in Nordic countries. *Statistics in Medicine,* **20**, (2001), 47-61.

[13] Statens Serum Institut. Kræftoverlevelse i Danmark fra 1997 til 2011. (2014).

Appendix

Residual analysis

The error structure

This appendix gives an overview of the residual in a model which include a cohort effect, with special focus on the residual autocorrelation. This demonstrated pattern is seen not only in death rate from breast cancer but also in death rate from: malnutrition, lung cancer, COLD, stroke, and heart attack.

Let us start with the simplest case: First order autocorrelation

$$e_t = \rho\, e_{t-1} + v_t \qquad (1a)$$

the assumptions on v_t are initially: serially uncorrelated with a normal distribution with constant variance: $v_t \approx N(0,\ \sigma^2)$

However, when v_t is not normal independent but has negative autocorrelation we e.g. have

$$v_t = -r\, v_{t-1} + \xi_t\ , \qquad (2a)$$

where $\xi_t \approx N(0,\ \sigma_1^2)$.

Therefore the estimation must be made in two rounds where in first round is

$$e_t = a\, e_{t-1} + \text{res} \qquad (3a)$$

The residual is used as an estimator (substitute) for v_t, and applied as an extra explaining variable to e_t and make a second round estimation

$$e_t = a\, e_{t-1} - b\, v_t + \text{error} \qquad (4a)$$

a and b, the estimates for ρ and r can both become bigger than one although ρ and r both are smaller than one.

Empirical evidence

Above, initially, a WLS estimation of equation (4) was made without including the dummies for the cohort effects, that is:

$$\text{Log(Dbc)*Age} = \alpha_1 + \alpha_2 \text{Age}^2 + \alpha_3 \text{Age}^3 + \alpha_4 \text{Age/T} + \alpha_5 \text{Age}^2/\text{T}^2 + \alpha_6 \text{Age}^3/\text{T}^3 \qquad (5a)$$

The classical test values became

$R^2 = .986$ DW = .46 Obs. = 432

The DW statistic seems to indicate highly significant positive autocorrelation. However, the DW statistic only test for first order autocorrelation, and is not suitably but highly misleading in this case as shown below.

Instead of removing the autocorrelation in (1a) by the cohort dummies we include the lagged residual, actually we use WGLS estimation:

$$\begin{aligned}\text{Log(Dbc)*Age} = \quad & \alpha_1 + \alpha_2 \text{Age}^2 + \alpha_3 \text{Age}^3 \\ & + \alpha_4 \text{Age/T} + \alpha_5 \text{Age}^2/\text{T}^2 + \alpha_6 \text{Age}^3/\text{T}^3 \\ & + \rho_1 \text{e(-1)} \end{aligned} \qquad (6a)$$

The estimated residual coefficient became:

$$\rho_1 = .766$$
t (24.61)

The classical test values became

$R^2 = .994$ DW = 2.80 Obs. = 420

We see that ρ_1 is highly significant; however, the DW statistic (for the remaining residual) now indicates significant negative autocorrelation, indicating that the residual in (1a) was not first order autocorrelated.

We now repeat the procedure – and repeat it again until the remaining residual is white noise.

$$\begin{aligned}\text{Log(Dbc)*Age} = \quad & \alpha_1 + \alpha_2 \text{Age}^2 + \alpha_3 \text{Age}^3 \\ & + \alpha_4 \text{Age/T} + \alpha_5 \text{Age}^2/\text{T}^2 + \alpha_6 \text{Age}^3/\text{T}^3 \\ & + \rho_1 \text{e(-1)} + \rho_2 \text{ee(-1)} \end{aligned} \qquad (7a)$$

Estimated residuals coefficients became:

+ 1.022*e(-1) - .644*ee(-1)
t (29.478) (-11.56)

and the classical test values became

$R^2 = .996$ DW = 2.32 Obs. = 408

Now there is no significant autocorrelation in the residuals. Monte Carlo experiments show that DW in cases like this is upward biased.

The autocorrelation pattern found in (1a) must in principle be equal to the autocorrelation in the variable $B \in \{ \beta_1, \beta_2, \beta_3 \ldots\ldots\ldots \beta_{91} \}$ shown above in Figure 2 and in Table 1 for the individual age groups.

The autocorrelation in the B-variable

Estimating the entire function the formula

$$\text{Log(Dbc)*Age} = \alpha_1 + \alpha_2 \text{Age}^2 + \alpha_3 \text{Age}^3 + \alpha_4 \text{Age}/T + \alpha_5 \text{Age}^2/T^2 + \alpha_6 \text{Age}^3/T^3 + \beta_1 \text{Coh1892} + \beta_2 \text{Coh1893} + \cdots + \beta_{91} \text{Coh1982} \quad (8a)$$

gives the classical test values:

$R^2 = .997$ DW = 2.01 Obs. = 424

Here the DW statistic shows no autocorrelation at all. As the residual is white noise the DW test is "feasible".
As mentioned above, the inclusion of the cohort dummies removed the autocorrelation in (5a). Consequently, the variable B must include (or have similarities to) "e" in (4a). Treating "B" as "e" we get:

$$B = \gamma_1 B(-1) \quad (9a)$$

$$B = 1.0052^* B(-1)$$

t (252.26)

$R^2 = .965$ DW = 2.80 Obs. = 90

Including the lagged residual gives:

$$B = 1.0058^* B(-1) - .4266^* eb(-1)$$

t (276.02) (-4.37)

$R^2 = .971$ DW = 2.18 Obs. = 89

The confusing outcome here is that the coefficients to e(-1) and B(-1) are bigger than one although both theoretically are smaller than one. However, that this is not a problem can be demonstrated by a Monte Carlo experiment.

Creating second order negative autocorrelation

For simulation purposes (1a) can be rewritten as

$$\varepsilon_t = v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \rho^3 v_{t-3} + \rho^4 v_{t-4} + \rho^5 v_{t-5} + \rho^6 e_{t-6} \ldots \quad (10a)$$

However, when v_t is not normal independent but has negative autocorrelation we have

$$v_t = -r v_{t-1} + \xi_t$$

$$v_{t-1} = -r v_{t-2} + \xi_{t-1}$$

$$
\begin{aligned}
v_{t-2} &= -r\, v_{t-3} + \xi_{t-2} \\
v_{t-3} &= -r\, v_{t-4} + \xi_{t-3} \\
v_{t-4} &= -r\, v_{t-5} + \xi_{t-4} \\
v_{t-5} &= -r\, v_{t-6} + \xi_{t-5} \\
v_{t-6} &= -r\, v_{t-7} + \xi_{t-6} \\
v_{t-7} &= -r\, v_{t-8} + \xi_{t-7} \\
v_{t-8} &= -r\, v_{t-9} + \xi_{t-8} \\
&:::::::::::,
\end{aligned}
\qquad (11a)
$$

which inserted in (10a) gives

$$
\begin{aligned}
\varepsilon_t &= \xi_t \\
&+ (\rho - r)\, \xi_{t-1} \\
&+ (\rho^2 + r^2 - \rho r)\, \xi_{t-2} \\
&+ (\rho^3 + \rho r^2 - r^3 - \rho^2 r)\, \xi_{t-3} \\
&+ (\rho^4 + \rho^2 r^2 - \rho^1 r^3 - \rho^3 r)\, \xi_{t-4} \\
&+ (\rho^5 + \rho^3 r^2 - \rho^2 r^3 - \rho^4 r)\, \xi_{t-5} \\
&+ (\rho^6 + \rho^4 r^2 - \rho^3 r^3 - \rho^5 r)\, \xi_{t-6} \\
&+ (\rho^7 + \rho^5 r^2 - \rho^4 r^3 - \rho^6 r)\, \xi_{t-7} \\
&+ (\rho^8 + \rho^6 r^2 - \rho^5 r^3 - \rho^7 r)\, \xi_{t-8} \\
&::::::::::
\end{aligned}
\qquad (12a)
$$

Therefore the estimation must be made in two rounds as shown in (3a) and (4a).

ρ and r are e.g. given the values: $\rho = .90$; $r = -.55$ (inserted as .55), and an e_t is calculated as a proxy for ε_t, here based on 30 lag.

A first order autocorrelation coefficient is calculated as:

$$
\begin{aligned}
e_t &= .640\, e_{t-1} \\
&\quad (83.27)
\end{aligned}
$$

$R^2 = .410$ DW = 2.61 Obs. = 9.969

The residual is called v_t. Second step is calculated as:

$$
\begin{aligned}
e_t &= 1.079\, e_{t-1} - .743\, v_t \\
&\quad (102.20) \quad (-54.09)
\end{aligned}
$$

$R^2 = .544$ DW = 2.01 Obs. = 9.968

We see that a Monte Carlo experiment based on 10.000 random numbers can produce the same result as the empirical data for breast cancer.