Report II
MT7052
Inference and prediction for life and health processeses

Malin Andersen - malinandersen1988@gmail.com
Oscar Potts - oscar.n.potts@gmail.com
Majken Gunnarsson - majken.gunnarsson@gmail.com

December 11, 2024

Part I

1. *Given the described stochastic model, write out the equations for the relevant transition probabilities $P_{i,j}(s,t)$using Kolmogorov's forward equations.*

   We have the Kolmogorov forward equation (KFE)[4] as:

$$\frac{\partial}{\partial t}P_{i,j}(s,t) = \sum_{g\neq j} p_{ig}(s,t)\mu_{gj}(t) - p_{ij}\sum_{g\neq j}\mu_{jg}(t) \tag{1}$$

   In the given problem description we have a 3-state model with a *simple transition graph*, meaning it does not have any looping pattern. This makes the Kolmogorov forward equation relatively simpler with compared to otherwise and we have:

$$\begin{aligned}
\frac{\partial}{\partial t}P_{0,0}(s,t) &= -P_{0,0}(s,t)(\mu_{0,1}(t)+\mu_{0,2}(t))\\
\frac{\partial}{\partial t}P_{0,1}(s,t) &= P_{0,0}(s,t)\mu_{0,1}(t) - P_{0,1}(s,t)\mu_{1,2}\\
\frac{\partial}{\partial t}P_{0,2}(s,t) &= P_{0,0}(s,t)\mu_{0,2}(t) + P_{0,1}(s,t)\mu_{1,2}(t)\\
\frac{\partial}{\partial t}P_{1,1}(s,t) &= -P_{1,1}(s,t)\mu_{1,2}(t)\\
\frac{\partial}{\partial t}P_{1,2}(s,t) &= P_{1,1}(s,t)\mu_{1,2}(t)
\end{aligned} \tag{2}$$

   Where $P_{i,j}(s,t)$ are our transition probabilities between time s and t with $mu_{i,j}$ are our hazard functions. Observe that we have an initial condition of being in state 1 and that the remaining transition probabilities not possible are constant 0. Given the Weibull distribution can also derive a close form expression given the Weibull distribution as

$$\mu_{i,j}(t) = \frac{b_{i,j}}{a_{i,j}}t^(b_{i,j}-1). \tag{3}$$

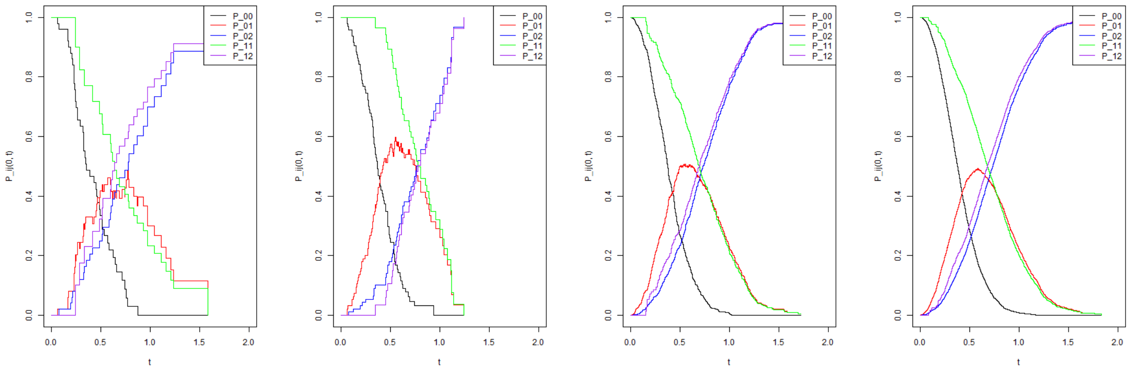
   You could use this parametric form and solve the PDE analytically relatively simply. But the purpose in this assignment is not to solve the system, it is to simulate the system and of course include censoring which is not included here.

2. *Use the supplied sample code to analyse the effect of sample size on the Aalen-Johansen estimated transition probabilities $P_{i,j}(0,t)$ based on the sample sizes $n = 50, 100, 1000, 5000$. Clearly explain what you observe and provide intuition for the estimated transition probabilities as a function of t.*

In figure 1 we see the results from running the supplied code fitting the Aaleen-Johansen[1] estimates for above KFE across the different sample sizes. Overall we do observe that the probability of being alive, healthy or sick $(P_{0,0}, P_{1,1})$, decreases and is near as t increases with the probability of being healthy $P_{0,0}$ being near 0 beyond $t = 1$. As expected we see the probability of death, being $P_{0,1}, P_{1,2}$ increases as t increases in opposite of the probability of being alive. Finally we note the probability of going from healthy to sick $(P_{0,1})$ has a bending pattern, beginning to increase and then decrease with the probability of being sick.

In regards to the sample size we see in 1(a) with a small sample size the transition probabilities $P_{0,t}$ are very unstable and vary a lot. The variation is caused by fewer individuals to represent the transition accurately. It could also be mention that a smaller sample create wider confident intervals, so there's a higher chance of over- or underestimating the probabilities due to natural variation.

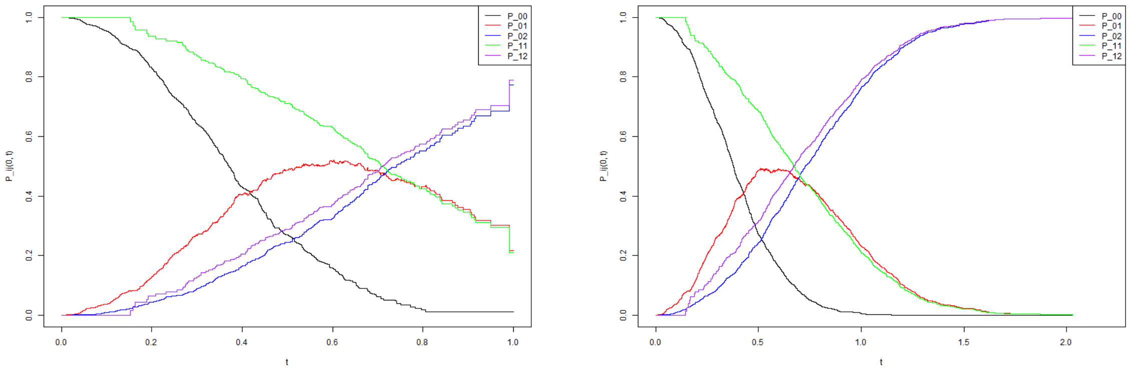
As the sample size increases as seen in images a) to d) the estimates becomes more stable and smoothing out. The probabilities begin to represent the underlying process better as seen in the images a) to d). With more data points the impact of outliers decrease resulting in a narrower confidence interval and more reliable estimates.



(a) $n = 50$ (b) $n = 100$ (c) $n = 500$ (d) $n = 1000$

Figure 1: Estimated Transition probabilities according to Aaleen-Johansen on different amount of datapoints generated.

3. *For* $n = 1000$*, analyse what happens when you use* $Tc \sim Unif(0, 1)$ *and* $Tc \sim Unif(0, 10)$*. What happens with the estimates close to* $t = 1$ *and* $t = 2$*?*



(a) $T_{cens} \sim Unif(0,1)$

(b) $T_{cens} \sim Unif(0,10)$

Figure 2: Estimated Transition probabilities according to Aaleen-Johansen on 1000 datapoints with different distribution of censoring.

Varying the censoring we see in figure 2 (a) and (b) that censoring impacts the Aalen-Johansen[1] estimator because it reduces the amount of data we can use to calculate the transition probabilities. It should be mentioned that we see the method being robust to this, having consistent estimates as in question 2, but that the output-space does not allow us to see all data meaning all ages.

Specifically we see this when censoring occurs early (e.g., $T_{\text{cens}} \sim \text{Unif}(0,1)$ figure 2 (a)), it removes a lot of observations, which makes us unable to say anything for $t > 1$. Again the estimates still align well with what we see in question 2, given the same interval, but we see more variation, or more clear jumps, as $t \to 1$. This is because this censoring makes high observations more scarce, leading to less information of events near $t = 1$

On the other hand, if the censoring is much wider as in the case of $T_{\text{cens}} \sim \text{Unif}(0,10)$, figure 2 (b), we observe data up to time $t = 2.03$, meaning that much of our censoring does not affect the estimates. This results in more uniform observations, with less variability or visible steps, as seen in 2 (a). As the Weibull survival function is exponential, seeing an observation of $T_j > 2$ is rare, and a observation $T_j = 10$ should be impossible. As such we get observation on most of the output space and only rarely does censoring occur. This scenario represents a typical situation where censoring does not imposes an upper limit on our observations.

Part II

Mortality Forecasting

Seen in figure 3 we have a Lexis diagram[5] displaying the observed mortality rates of women in Sweden aged 0 to 90 during the period 1900 − 2020. What is observed is how the general mortality rates has decreased over the years as the plot turns brighter towards modern years. This is likely a consequence of how innovations in healthcare and public health has improved the life expectancy across the years.

Specifically we see brighter survival spots for young women, which may be that pregnancy deaths has gone down through the years. We also see how the death rate of births has decreased but is still a spike in mortality from the dark line at age 0. Finally we do observe a dark horizontal line near 1917, likely being a consequence of famine during the first world war.

What is to expect from these observations is how life expectancy for women has increased the last 100 years, and likely during the last 30 years as well. We should thereby expect that later models show positive effects on life expectancy as we go forward in time.

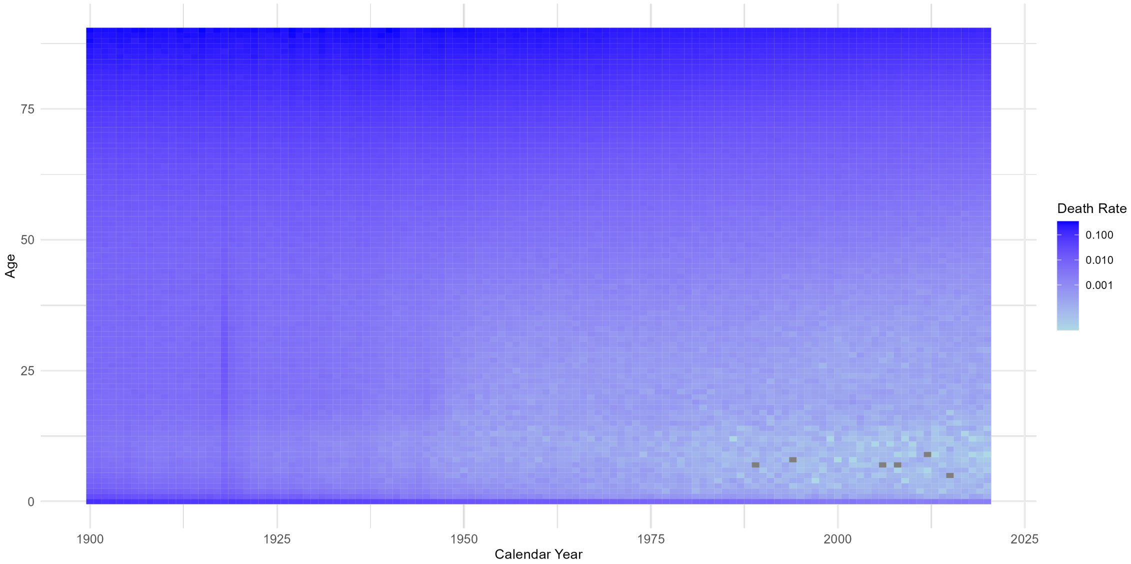


Figure 3: Lexis diagram over survival rates for Swedish women aged 0 to 90 during the period 1900 − 2020.

Poisson Lee-Carter

The Poisson Lee-Carter model[3] is a model which dissects mortality in 3 different factors assuming we have a relationship

$$log(mu) = \alpha_x + \beta_x \kappa_t + \varepsilon_{x,t} \tag{4}$$

between mortality, age and time where $\varepsilon_{x,t}$ is the random noise observed. We will go through the interpretation of the terms shortly, but will first notice and acknowledge the exponential assumptions of the Lee-Carter model.

Fitting a Poisson Lee-Carter model for Swedish women aged $20 - 90$ years old on the years $1990 - 2010$ we see the resulting fit in figure 4. This was done in accordance to Brouhns et al[2] ,in which one begin by estimating the MLEs of the α_x, β_x and κ_t, and second fit a time series model to the estimated κ_t. This was done through the StMoMo in R.

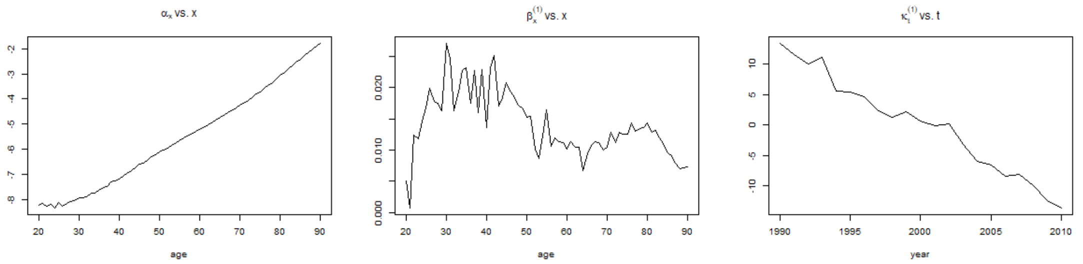


Figure 4: Parameters to a Poisson Lee-Carter model fitted on Swedish female mortality data between 1990-2010 with ages between 20-90 years.

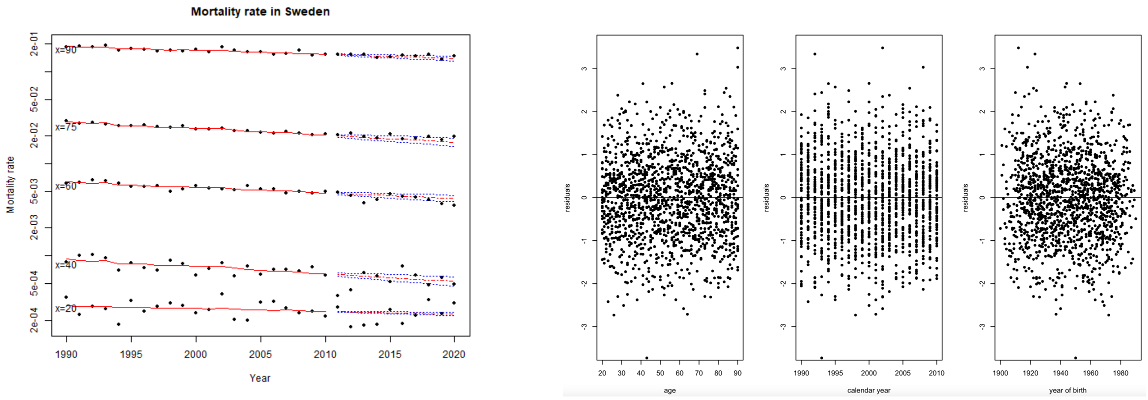
Our finding align closely with the observations made in the first task. The coefficient α_x which describes the baseline age hazard for each age increases almost linear in accordance to time. It can also be interpreted as the expected hazard of a Swedish Women aged x drawn randomly anywhere between 1990 and 2010.

The κ_t measures the time dependence to mortality and as can be seen we have a negative near linear trend between 1990 and 2020. This means that overall the mortality rate has decreased as we have proceeded through the centuries.

This κ_t should be taken in relation to our β_x which tells us which ages are more affected by this linear time trend. We observe that, for ages 25-50, there is a higher relationship to time, and less so for ages above 50, and near no relation for people aged 20-21. This suggests that over the past 30 years mortality rates for ages 25-50 were nearly cut i half compared to those for individuals over 50. Although mortality rates for those over 50 have also declined, the trend is not as strong as that seen in younger age groups.

Note the exponential nature of the Poisson Lee-Carter model, meaning the linear relationship we see is in fact exponential. This implies that middle aged women seeing double the time dependence compared to elder is double in the log-linear sense, and in fact is some factor e^2 compared to the elderly.

Finally we have the resulting model fit shown in figure 5. Here we see the resulting model mortality fit for a series of different ages with the observed residuals of our fit. Observe that 2010 to 2020 denotes unseen data for the model, and we see the predicted confidence intervals for these values. What is generally observed is how the fitted mortality fits well with observed values and that observed predictions are in the confidence intervals of our model except for young ages where there is more variance. Moreover, we can see that the residuals for all the variables are normally distributed around our fitted values, with no signs of heterogeneity or sparse-ness, as assumed by our linear model.



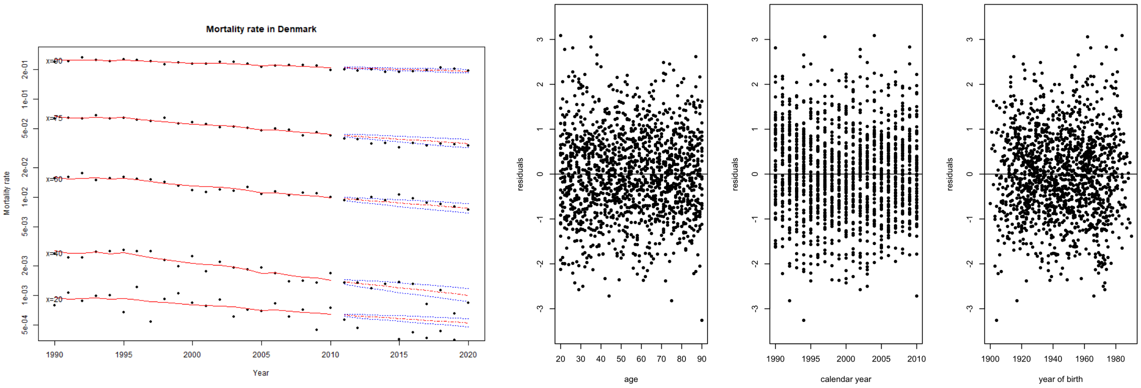
(a) Observed and predicted Mortality rate.

(b) Residuals for the model.

Figure 5: Describes the result from the fitted Poisson Lee-Carter model on Swedish female mortality data between 1990-2010 with ages between 20-90 years. Figure a) denotes the raw mortality rates as black points. The red line visualize the model value of the mortality rate before 2010 and after 2010 it visualize the predicted mortality rates. The blue lines describes the 5% and 95% percentiles.

Multi-population

In this part we have fitted a Poisson Lee-Carter model to the Danish male population between 1990-2010 and the ages 20-90 years. In figure 6 we can see the result of the fitted model. Observations regarding the fit are much in accordance with what we saw for our fit for Swedish data, but of course we see that our estimates differ between the models, with men having a higher rate of mortality compared to women. Specifically we see how the mortality for men aged 40 decreases much more rapid compared to Swedish, even as there is much more variance in this case.



(a) Observed and predicted Mortality rate.

(b) Residuals for the model.

Figure 6: Describes the result from the fitted Poisson Lee-Carter model on Danish male mortality data between 1990-2010 with ages between 20-90 years. Figure a) denotes the raw mortality rates as black points. The red line visualize the model value of the mortality rate before 2010 and after 2010 it visualize the predicted mortality rates. The blue lines describes the 5% and 95% percentiles.

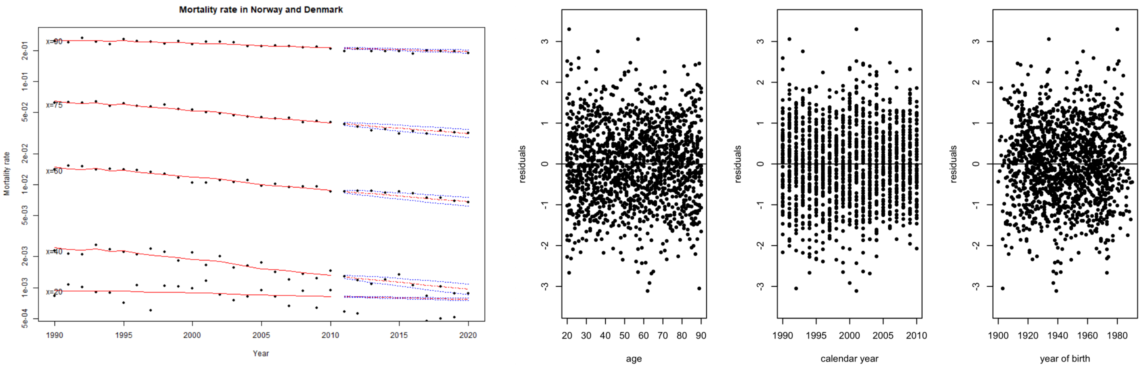
Joining Danish and Norwegian data we see results in figure 7. In figure 7 (a) we can see the prediction of the mortality rate $\hat{\mu}^{L-C,(0)}$ as the red line for some ages. Figure 7 (b) describe the residuals for the model. Again observations are much in accordance with previous fits, even as the model is fitted on multi-population data. Similar with the Danish model, a lot of trends are similar but do observe that the mortality of 40 year olds approach that of 20 year olds. Compared to the Danish data 20 year olds had less of a flatter trend compared to the joint dataset. This is to be excplored further in the final task of this assignment.

Continuing the the joint mixed-population dataset we introduce the ratios

$$\hat{\pi}(x) := \frac{\sum_{t=1990}^{2010} d_{xt}^{(1)}}{\sum_{t=1990}^{2010} r_{xt}^{(1)} \hat{\mu}_t^{(0)}(x)} \tag{5}$$

where $\hat{\mu}^{(0)}(x)$ is our joint predictor on mortality rate, $r_{xt}^{(1)}$ is the exposure on the Danish population and $d_{x,t}^{(1)}$ the number deaths for Danish males aged x at year t. This ratio tells us how much the estimated Nelson-Aalen mortality rate for the Danish male population differ between the fitted mortality rate in the joint Danish and Norwegian males population. A ratio of 1 means that there is no difference between the mortality rates.

In figure 8 we see this ratio plotted for a series of different ages. What we can observe is that the ratio is smaller or equal to 1 for all ages, meaning that the Danish male population



(a) Observed and predicted Mortality rate.

(b) Residuals on the model.

Figure 7: Describes the result from the fitted Poisson Lee-Carter model on joint Danish and Norwegian male mortality data between 1990-2010 with ages between 20-90 years. Figure a) denotes the raw mortality rates as black points. The red line visualize the model value of the mortality rate before 2010 and after 2010 it visualize the predicted mortality rates. The blue lines describes the 5% and 95% percentiles.

have a lower mortality rate than the joint Danish and Norwegian male population. Which means that Norwegian males dies in a faster pace than Danish males. The difference is larger for smaller ages, where the biggest difference is for ages around 20 to 40 years, and less so for higher ages. This is much accordance to our observations comparing figure 6 and 7 where specifically we see a large difference between the evolution for 20 year olds

Introducing the estimator

$$\hat{\mu}_t^{L-C,(1)}(x) = \hat{\pi}(x)\hat{\mu}_t^{L-C,(0)}(x) \tag{6}$$

where $\hat{\mu}_t^{L-C,(1)}(x)$ is the estimated mortality rate for the Danish male population with age x and year t, $\hat{\mu}_t^{L-C,(0)}(x)$ is the mortality rate for the joint Danish and Norwegian male population and $\hat{\pi}(x)$ is the ratio between the mortality rates.

The result for the estimated $\hat{\mu}_t^{L-C,(1)}(x)$ can be seen as the red line in figure 9. The green line is the mortality rate estimated directly on the Danish male population without using a joint population. From the figure we can see that the estimated mortality rates differ a bit between the methods, where the mortality rate often is a bit larger for the method which was based on the joint population. We do observe that the observations now are much more in accordance to figure 6 and 7 , specifically for age 20.

This plot visualizes some of the problems in combining populations. Combining the data we

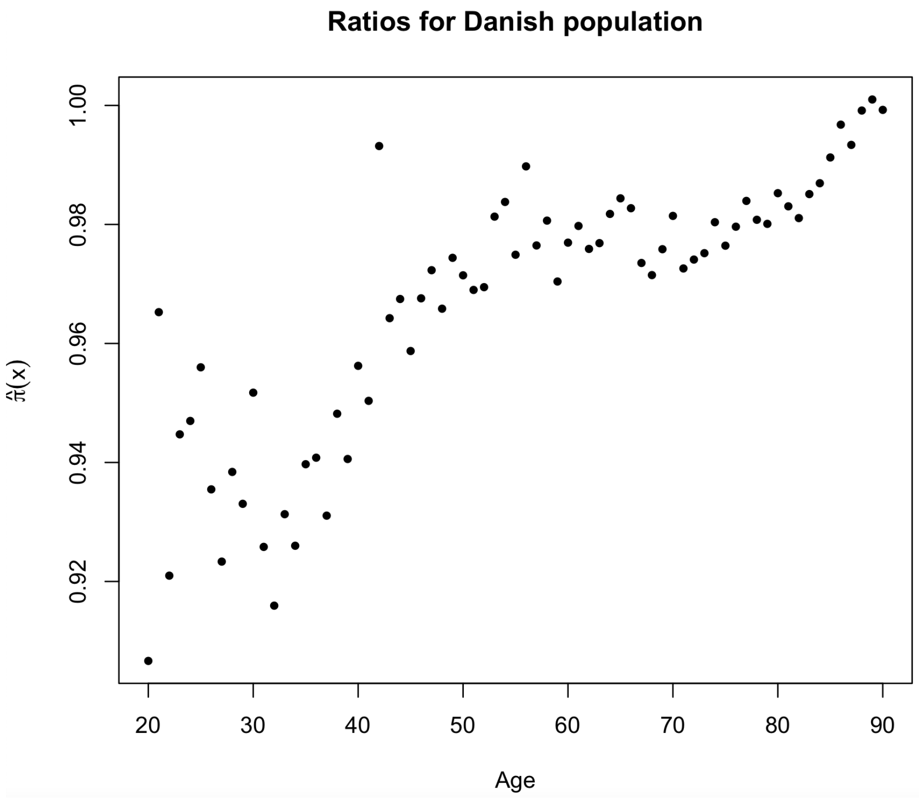


Figure 8: Ratios which are used in the alternative method of achieving a mortality rate in figure 9. Describes the relationship between the modeled mortality rate on jointed Danish and Norwegian male data and observed mortality rate in Denmark.

get less variance and better joint estimates but sometimes there are discrepancies between the subpopulations. In this case that subpopulation is the age 20 where we see difference in time trends. There is no obvious solution here,either you take higher variance model the population seperately or you join them for more narrow predictions. The choice should be decided on what is important to measure the model.

As a conclusion we have shown different ways one can model and forecast mortality according to the Lee-Carter Model. Doing so we have interpreted the different coefficients and variables. Generally we see that life expectancy seem to increase over time. Specifically we see that there is a decline in mortality for younger ages, something which should be good for us writing this report (aged 25-35). This does not guarantee future life expectancy, but we can say that we have it better than what people our age had 5 years ago. As all things in life, the past does not predict the future, and since you only die once, we have only observed small portions of the mortality that is to come.

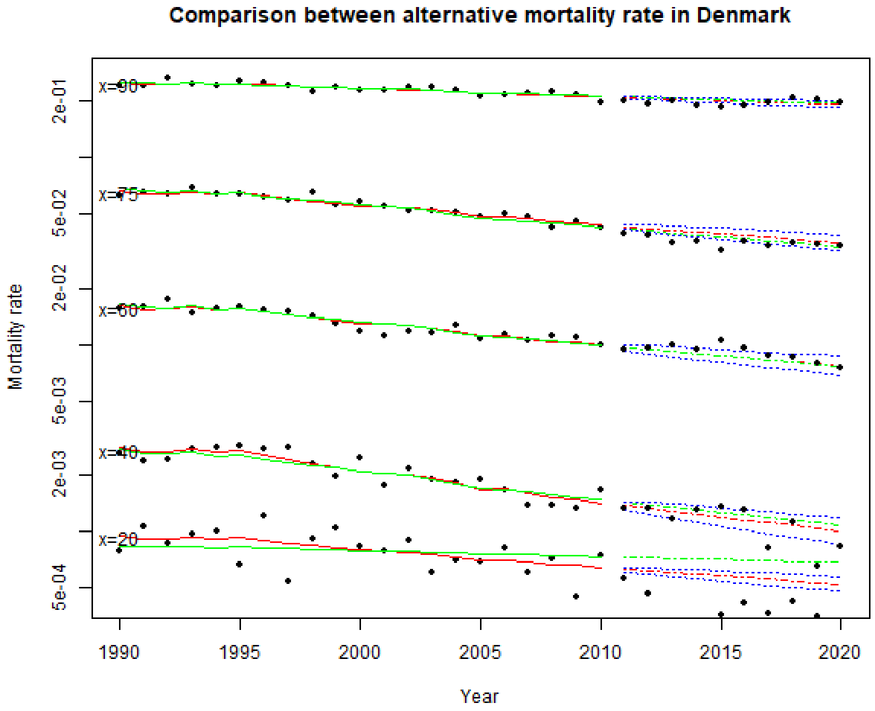


Figure 9: Describes the result from the fitted Poisson Lee-Carter model on Danish male mortality data between 1990-2010 with ages between 20-90 years compared to an alternative method for achieving Mortality rate for this group. The raw mortality rates are denoted as black points. The blue lines describes the 5% and 95% percentiles of the joint mortality. The red line is the mortality rate for Danish Males as described in equation 6 while the green the joint mortality of both populations.

References

[1] O. O. Aalen. *Survival and Event History Analysis: A Process Point of View*. Springer-Verlag, 2008.

[2] N. Brouhns, M. Denuit, and J. K. Vermunt. A poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and economics*, 31(3):373–393, 2002.

[3] R. D. Lee and L. R. Carter. Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671, 1992.

[4] R. Norberg. Basic life insurance mathematics. *Lecture notes, Laboratory of Actuarial*

Mathematics, University of Copenhagen, 2002.

[5] R. Rau, C. Bohk-Ewald, M. M. Muszyńska, and J. W. Vaupel. Visualizing mortality dynamics in the lexis diagram. 2017.