The sections included in this project are

- introduction,
- Lee-Carter model,
- data,
- model selection,
- forecasting, and
- conclusion

1. Introduction

Lee-Carter model introduced by Lee and Carter (1992) has long been a popular long-term mortality projection model since debut. The Lee-Carter model is a a two-factor model, extracting two age-specific elements for every age and a time-varying effect for every fitted time from the central death rates. In the projection phrase, the model keeps the age-specific element of each age and projects the future time-varying index by a time series model. This project will concentrate on the model selection and forecasting for the time-varying index.

2. Lee-Carter model

A brief introduction of actuarial concept is given before the introduction of the Lee-Carter model. The central death rate at age x in year t, $m_{x,t}$, is defined as $D_{x,t}$, the number of deaths aged x last birthday at the date of death during year t divided by $E_{x,t}$, the average population aged x last birthday during year t. Commonly, there are two approximations to the central death rate $m_{x,t}$ with the mortality rate of an individual aged x at time t, $q_{x,t}$. The first approach, $q_{x,t} = 1 - \exp(-m_{x,t})$, is based on the assumption of constant force of mortality within each integer age. The second approach, $q_{x,t} = m_{x,t}/(1+0.5m_{x,t})$, is under the assumption of uniform distribution of deaths (UDD) within each integer age. In this project, the former approximation is adopted for the data transformation between $m_{x,t}$ and $q_{x,t}$.

The Lee-Carter model is given by

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

where

- both a_x and b_x are age-specific constants,
- k_t is the time-varying index,
- $\epsilon_{x,t}$ is the error term and is assumed to follow a normal distribution with mean zero and to be independent of age x and time t, and
- the mortality data from time t_L to t_U and from age x_L to x_U for age are used for fitting.

As Lee and Carter (1992) explained, under their model e^{a_x} can depict the general shape of the central death rate for age x across the projection period, while b_x can reveal the relative changes in response to t since $d\ln(m_{x,t})/dt = b_x dk_t/dt$. The parameter estimation for the Lee-Carter model is not unique, and is subject to two constraints, $\sum_t k_t = 0$ and $\sum_x b_x = 1$. The first constraint is a natural constraint, which leads a_x to be the average of $\ln(m_{x,t})$ over time t. That is,

$$\hat{a}_x = \frac{1}{t_U - t_L + 1} \sum_{t=t_L}^{t_U} \ln(m_{x,t}).$$

The original paper suggested using the singular value decomposition (SVD) method to find $\{b_x\}$ and $\{k_t\}$ which minimize the sum of least squared errors. Alternatively, the second constraint gives the estimation of k_t by

$$\hat{k}_t = \sum_{x=x_L}^{x_U} [\ln(m_{x,t}) - \hat{a}_x]$$

for each year t, and $\hat{b}_x$ can be obtained by regressing $[\ln(m_{x,t}) - \hat{a}_x]$ on $\hat{k}_t$ without the constant term being involved for each age x.

The forecasted central death rate at age x and year T, denoted as $\hat{m}_{x,K}$, is $\exp(a_x + b_x k_T)$, $T > t_U$. Hence, the forecasted mortality rate at age x and year T, denoted by $\hat{q}_{x,K}$, is

$$\exp(-\exp(\hat{a}_x + \hat{b}_x \hat{k}_T)).$$

Lee and Carter (1992) gave an approach to constructing the confidence interval for $\ln(m_{x,T})$ for age x and year T. The standard deviation of the logarithm of the projected central death rate at age x and year T ($T > t_U$), denoted by s.d.$(\ln(\hat{m}_{x,T}))$, is

$$\hat{b}_x \sqrt{\mathrm{Var}(\hat{k}_T)},$$

where $\mathrm{Var}(\hat{k}_T)$ is the variance of the projected k_t at time T.

3. Data

The mortality data used in this project is from Human Mortality Database (2013). The mortality data is collected from Japanese males from year 1950 to 2009. The age range applied in this project is from 25 to 84, a 60-age span.

4. Model selection

This section focuses on the model diagnostic for the time-vary element k_t through various plots like the sample autocorrelation and sample partial autocorrelation. The sample autocorrelation at lag k, denoted as r_k, is defined as

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}.$$

The partial autocorrelation at lag k, denoted as ϕ_{kk}, is defined as

$$\phi_{kk} = Corr(Y_t, Y_{t-k}|Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k+1}).$$

Figure 1, 2 and 3 give the picture of the time varying element k_t on the aspect of trend, sample autocorrelation, sample partial autocorrelation. As we can conclude from these three plots, k_t is not a stationary time series with high autocorrelation in the first few lags. As a result, the first difference of k_t is then examined.

Figure 4, 5 and 6 serve the purpose of the diagnostic of the first difference of k_t. As you can see from in 4, the first difference of k_t is relatively much more stationary than the original time series. The sample autocorrelation and partial autocorrelation in Figure 4 and 6, respectively, also support that the first difference of k_t has low autocorrelation between different lags.

5. Model fitting

Four ARIMA model are selected to fit the time varying element k_t, considering the non-stationarity of k_t and stationarity of the first difference of k_t. They are are

- ARIMA(1,1,0),
- ARIMA(2,1,0),
- ARIMA(0,1,1) and
- ARIMA(1,1,1).

The output from R for all four model is presented in the Appendix. To review the goodness of fitting, the statistics of least square errors and AICs (see in Table 1) of the four models indicate that the ARIMA(2,1,0) has the best fitting result, however the difference in these two statistics are minor. In addition, the theoretical autocorrelation of the four models is compared with the sample correlation also suggest that ARIMA(2,1,0) has similar autocorrelation pattern as the sample autocorrelation. ARIMA(2,1,0) model has the property that the autocorrelation starts an exponential decay from the third lag.

FIGURE 1. The time series of k_t for year 1950 to year 2009

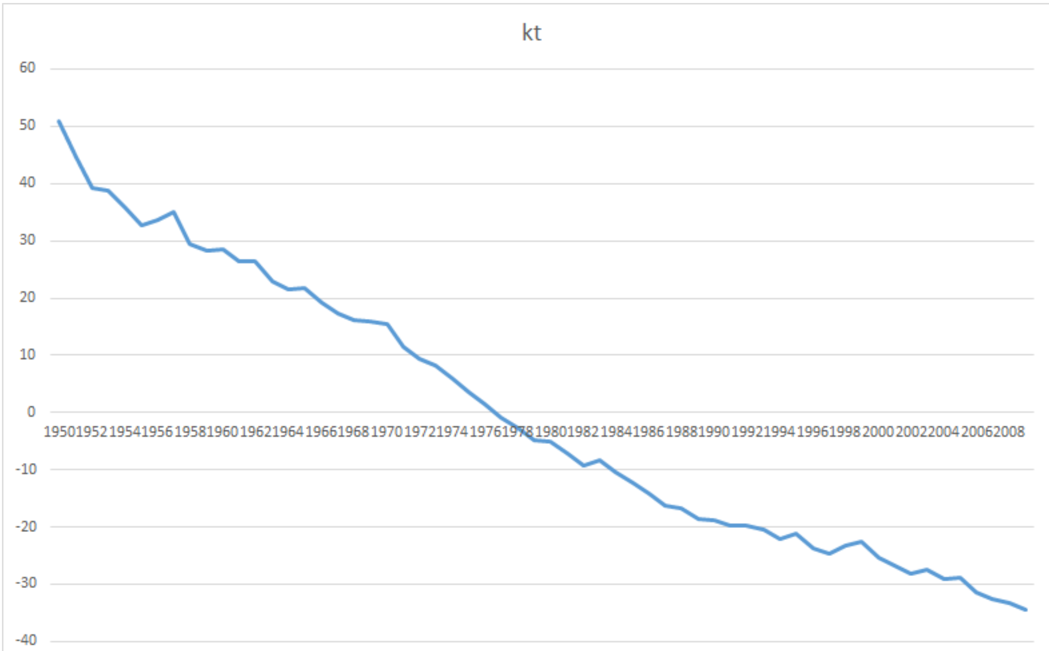


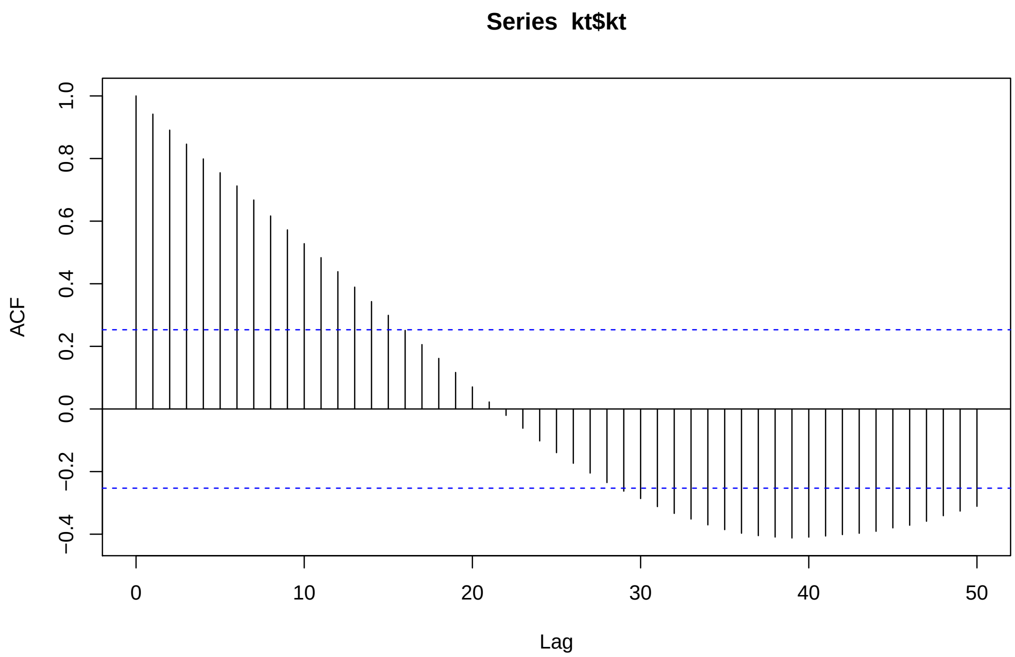
TABLE 1. Least square errors and AICs for ARIMA(1,1,0), ARIMA(2,1,0), ARIMA(0,1,1) and ARIMA(1,1,1)

Item	ARIMA(1,1,0)	ARIMA(2,1,0)	ARIMA(0,1,1)	ARIMA(1,1,1)
LSE	2.522	2.387	2.521	2.515
AIC	228.01	226.91	227.99	229.86

TABLE 2. Theoretical autocorrelation V.S. sample correlation

Lag	ARIMA(1,1,0)	ARIMA(2,1,0)	ARIMA(0,1,1)	ARIMA(1,1,1)	Sample correlation
1	-0.0164	-0.0489	0.0292	0.5357	-0.0140
2	0.0003	-0.2595	0.0000	-0.0156	-0.1920
3	0.0000	0.0288	0.0000	0.0005	0.2480
4	0.0000	0.0663	0.0000	0.0000	0.0980
5	0.0000	-0.0117	0.0000	0.0000	-0.0320

FIGURE 2. The autocorrelation of the time series of k_t for year 1950 to year 2009

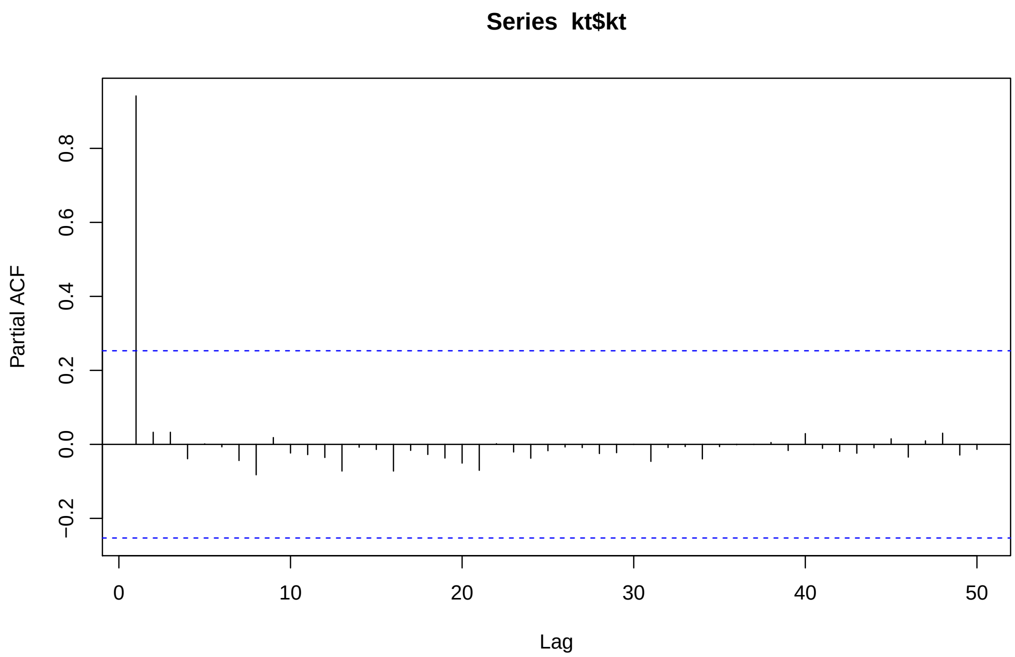


6. FORECASTING

The forecasting difference between these four time series model are minor in the forecast phase as we can see from Figure 7, even if we enlarge the forecast portion in the right side of Figure 7.

Based on the discussion in the previous section, only ARIMA(2,1,0) model is selected to forecast the mortality rate in this section. With 8 and 9 for the forecast and confidence interval for k_t and mortality rates, respectively. The acceleration of decreasing speed in k_t is different for certain period (i.e. the k_ts before 1985 dropping faster than those after). However, the forecasted time-varying element inherited the characteristics of both the pre-1985 and post-1985 mortality data, making the time-varying element change more steep than the most recent trend. In addition, this time-varying element is universal for all ages in study, but the experience of mortality improvement in each age is different. The mortality improvement is more significant in order ages, and Figure 9 also support that projected mortality rates under the Lee-Carter model are more consistent with the historic mortality rates in older ages than younger ones.

FIGURE 3. The partial autocorrelation of the time series of k_t for year 1950 to year 2009



7. Conclusion

There have been some research accomplishment regarding the changing of the constant term in the time-varying element, and thus the problem mentioned above that the k_ts before 1985 dropping faster than those after can be solved.

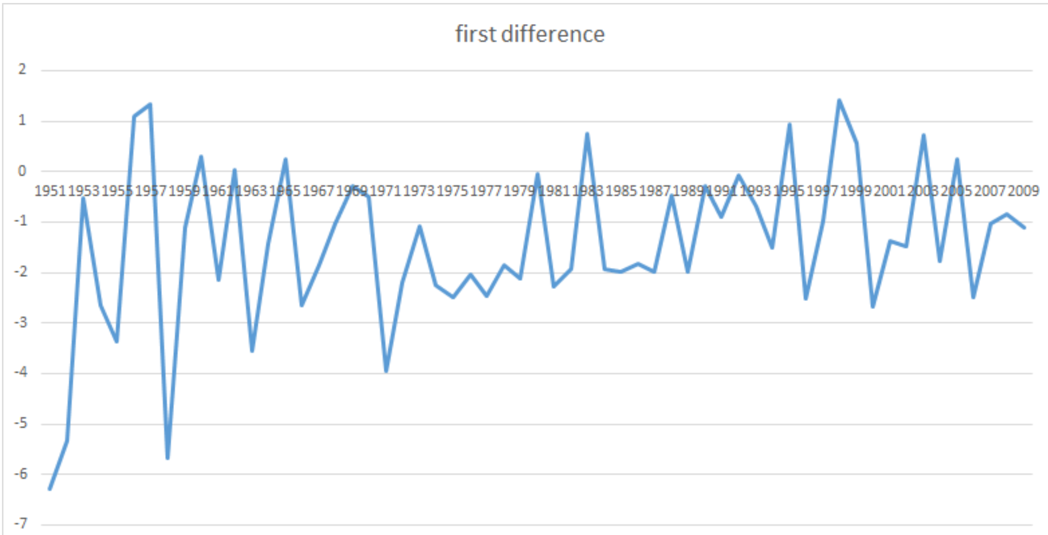
The Lee-Carter model may not a good choice to predict the mortality rates with a large scope of ages since the time-varying element is universal for all ages.

8. Appendix

8.1. ARIMA(1,1,0).

```
  Call:
arima(x = diff(kt$kt), order = c(1, 0, 0))

Coefficients:
          ar1  intercept
```

FIGURE 4. The first difference of the time series of k_t for year 1950 to year 2009




```
      -0.0164    -1.4432
s.e.   0.1407     0.2037

sigma^2 estimated as 2.522:  log likelihood = -111,  aic = 228.01
```

$$\nabla k_t + 1.4432 = -0.0164(\nabla k_{t-1} + 1.4432) + e_t$$

8.2. ARIMA(2,1,0).

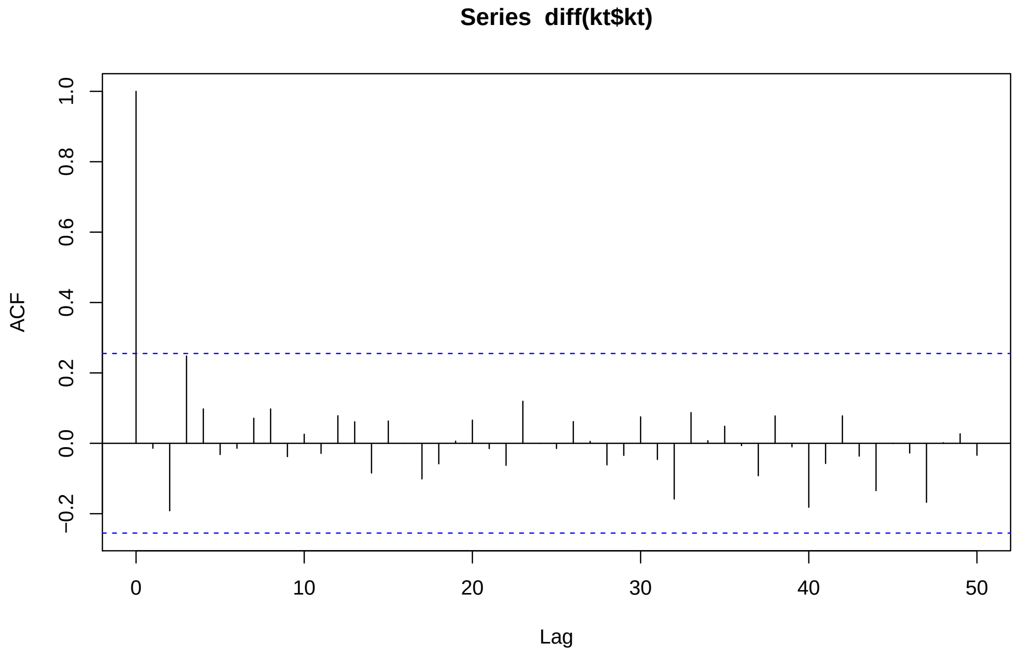
```
Call:
arima(x = diff(kt$kt), order = c(2, 0, 0))

Coefficients:
          ar1      ar2  intercept
      -0.0617  -0.2625    -1.4143
s.e.   0.1398   0.1468     0.1537

sigma^2 estimated as 2.387:  log likelihood = -109.46,  aic = 226.91
```

$$\nabla k_t + 1.4143 = -0.0617(\nabla k_{t-1} + 1.4143) - 0.2625(\nabla k_{t-2} + 1.4143) + e_t$$

8.3. ARIMA(0,1,1).

```
Call:
arima(x = diff(kt$kt), order = c(0, 0, 1))
```

FIGURE 5. The autocorrelation of the first difference of the time series of k_t for year 1950 to year 2009



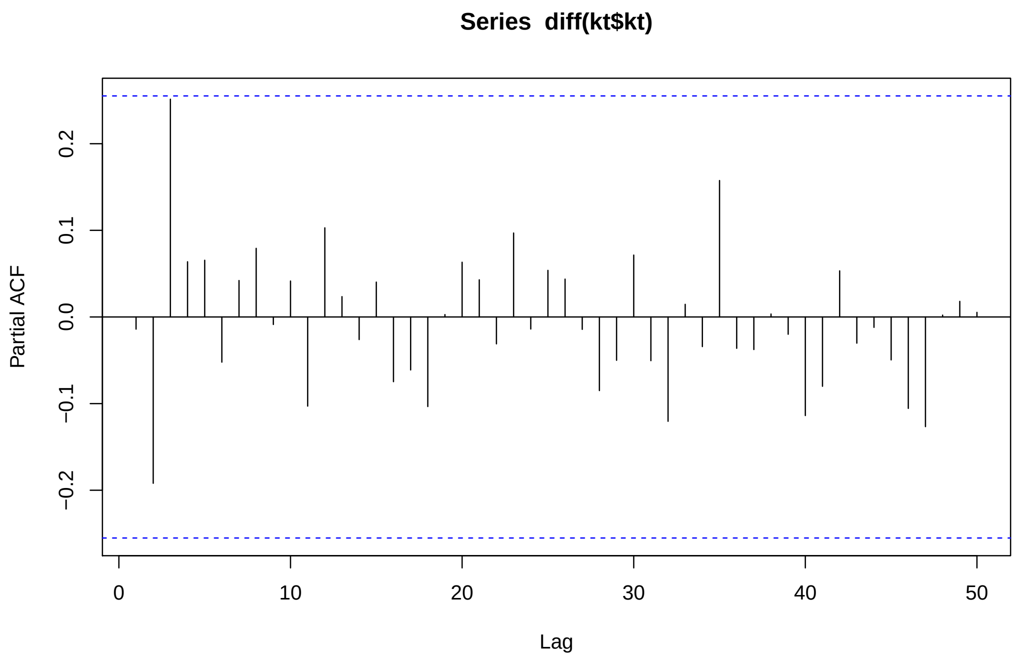

```
Coefficients:
          ma1  intercept
      -0.0292    -1.4422
s.e.   0.1854     0.2013

sigma^2 estimated as 2.521:  log likelihood = -111,  aic = 227.99
```

$$\nabla k_t + 1.4422 = e_t + 0.0292 e_{t-1}$$

8.4. **ARIMA(1,0,1).**

```
Call:
arima(x = diff(kt$kt), order = c(1, 0, 1))

Coefficients:
         ar1      ma1  intercept
      0.2646  -0.3223    -1.4372
```

FIGURE 6. The partial autocorrelation of the first difference of the time series of k_t for year 1950 to year 2009




```
s.e.  0.6499   0.6320     0.1916

sigma^2 estimated as 2.515:  log likelihood = -110.93,  aic = 229.86
```

$$\nabla k_t + 1.4372 = 0.2646(\nabla k_{t-1} + 1.4372) + e_t + 0.3223e_{t-1}$$

REFERENCES

[1] Lee, R.D. and Carter, L.R., *Modeling and forecasting* US *mortality*. Journal of the American Statistical Association, 87: 659-675, 1992.

[2] Human Mortality Database, University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany), Available at www.mortality.org or www.humanmortality.de. 2013

FIGURE 7. The forecast of the time-varying element k_t

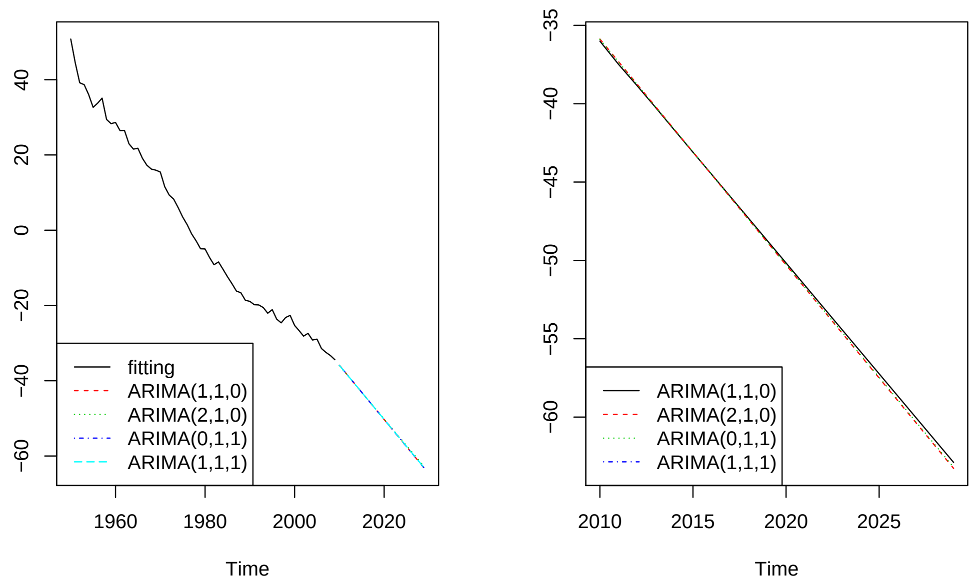


FIGURE 8. The forecast and 90% confidence interval of the time-varying element k_t

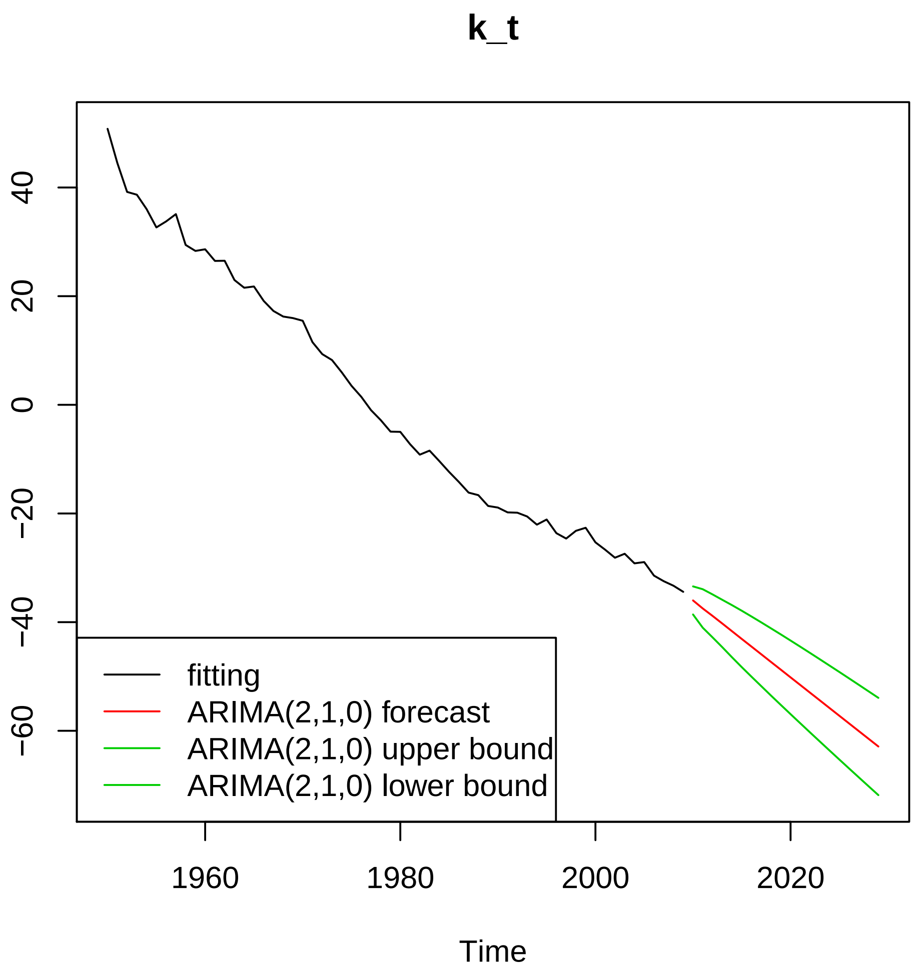


Figure 9. The forecast and 90% confidence interval of the mortality with selected ages