Project Report ~ LSTAT2170 : Times series analysis

Introduction

Times series are data collected based on time (may be discrete or continuous). Their analysis provides reliable information about the past situation and allows prediction on the future. This analysis in business, finance, marketing ... are great opportunities for stakeholders to grow and have great outcomes of they activities.

In this project, we will apply times series forecasting methods to the *car drivers* data. This data is a monthly number of car drivers killed and seriously injured in Great Britain from January 1969 to December 1984.

First, we will describe the structure of this data and try to find a suitable model to fit the data. And then, we will perform a prediction based on the fitted model that will be compare to a non-parametric prediction (the Holt-Winters method).

1 Data loading and preliminary analysis

1.1 Data viewing

Here, let's have a general look of the data set (Table 1).

Table 1: Car drivers killed and seriously injured in Great Britain
from January 1969 to December 1984

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1969	1687	1508	1507	1385	1632	1511	1559	1630	1579	1653	2152	2148
1970	1752	1765	1717	1558	1575	1520	1805	1800	1719	2008	2242	2478
1971	2030	1655	1693	1623	1805	1746	1795	1926	1619	1992	2233	2192
1972	2080	1768	1835	1569	1976	1853	1965	1689	1778	1976	2397	2654
1973	2097	1963	1677	1941	2003	1813	2012	1912	2084	2080	2118	2150
1974	1608	1503	1548	1382	1731	1798	1779	1887	2004	2077	2092	2051
1975	1577	1356	1652	1382	1519	1421	1442	1543	1656	1561	1905	2199
1976	1473	1655	1407	1395	1530	1309	1526	1327	1627	1748	1958	2274
1977	1648	1401	1411	1403	1394	1520	1528	1643	1515	1685	2000	2215
1978	1956	1462	1563	1459	1446	1622	1657	1638	1643	1683	2050	2262
1979	1813	1445	1762	1461	1556	1431	1427	1554	1645	1653	2016	2207
1980	1665	1361	1506	1360	1453	1522	1460	1552	1548	1827	1737	1941
1981	1474	1458	1542	1404	1522	1385	1641	1510	1681	1938	1868	1726
1982	1456	1445	1456	1365	1487	1558	1488	1684	1594	1850	1998	2079
1983	1494	1057	1218	1168	1236	1076	1174	1139	1427	1487	1483	1513
1984	1357	1165	1282	1110	1297	1185	1222	1284	1444	1575	1737	1763

1.2 Data plot

From this plot (Figure 1), we notice a trend and a seasonality in the data. Hence,

- from *1969* to *1973*, the trend tends to increase, and the seasonal part is in somehow irregular (box in gray on the plot)
- around *1973-1974*, their is a structural break where the trend goes down, with the seasonality, a little bit breaked.
- from *1975* to *1980*, the data shows a relative stable structure, with the trend and the seasonality that tend to be relatively constant. This fact remains during *1980-1982* (box in sky blue) before showing another break around *1983*

Overall, in first sight, the drivers killed and seriously injured during this period tend to *decrease* (which is obviously a good news : maybe a certain policies have been made).

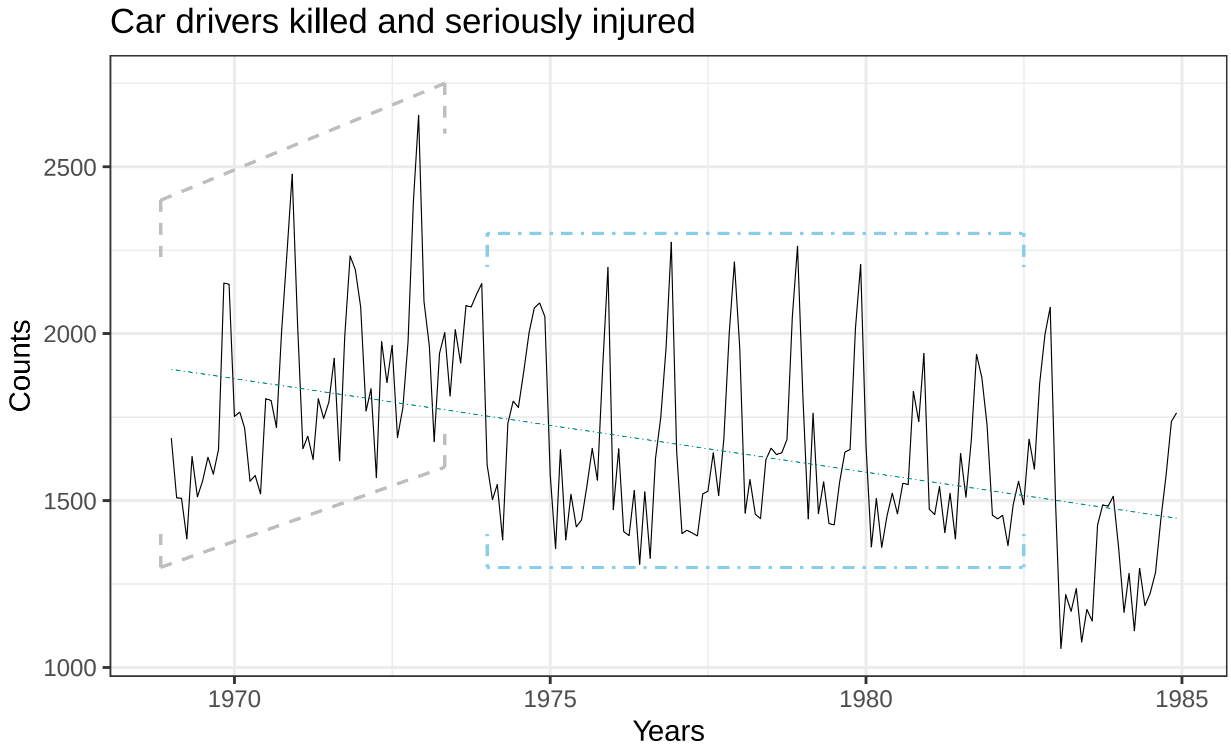


Figure 1: Car drivers killed and seriously injured in Great Britain from January 1969 to December 1984

To make a better analysis of this *car drivers* data, and because of the seasonality and trend in the data, we need to get the stationary part of the data. This is discuss in the next session

2 Visual inspection of the times series structure

2.1 Detrend and deseason of the data

As the data set presents a trend and a seasonality, let's apply a non parametric method to render it stationary. Be award that, due to the structure of the plot, we don't need any transformation to stabilize the variance as the variance among year (and overall years) seems stable.

Thus, as the data is a monthly data over years, a ***12*** order differential is supposed to treat the seasonality in the data and an order ***1*** will treat the trend of the data. Thus we could get the stationary part of the data.

What we have said above (Figure 2), can be observed in the plot below. Take a closer look at the plot and notice that the stationary part is *mean zero*.

With the non parametric method, we get a stationary part of the *car drivers* data. To which, we can check what theoretical model can best fit this stationary part of the data. This is done in the next session.

2.2 Autocorrelation and Partial autocorrelation analysis

To find the best model to fit the stationary part, let's analyse the autocorrelation and partial autocorrelation of the detrend and deseason data (Figure 3). Due to this seasonality and trend, we need to analyse the monthly dependent and also the yearly dependent.

From the autocorrelation and partial autocorrelation plots :

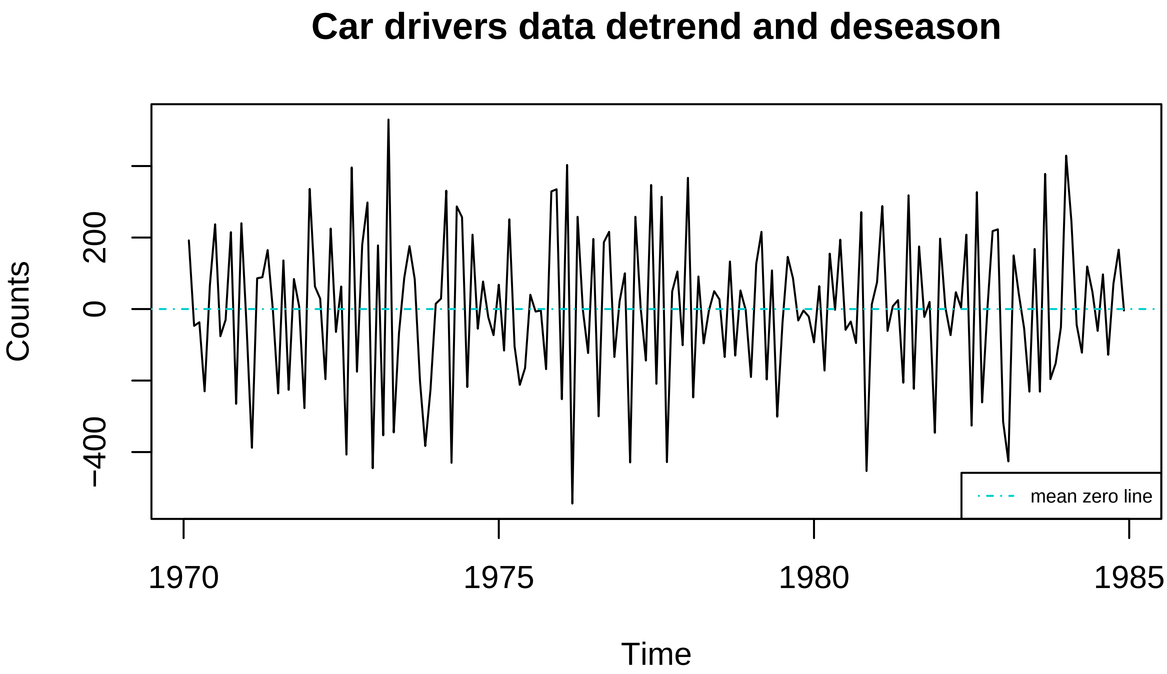


Figure 2: Stationary part of Car drivers data

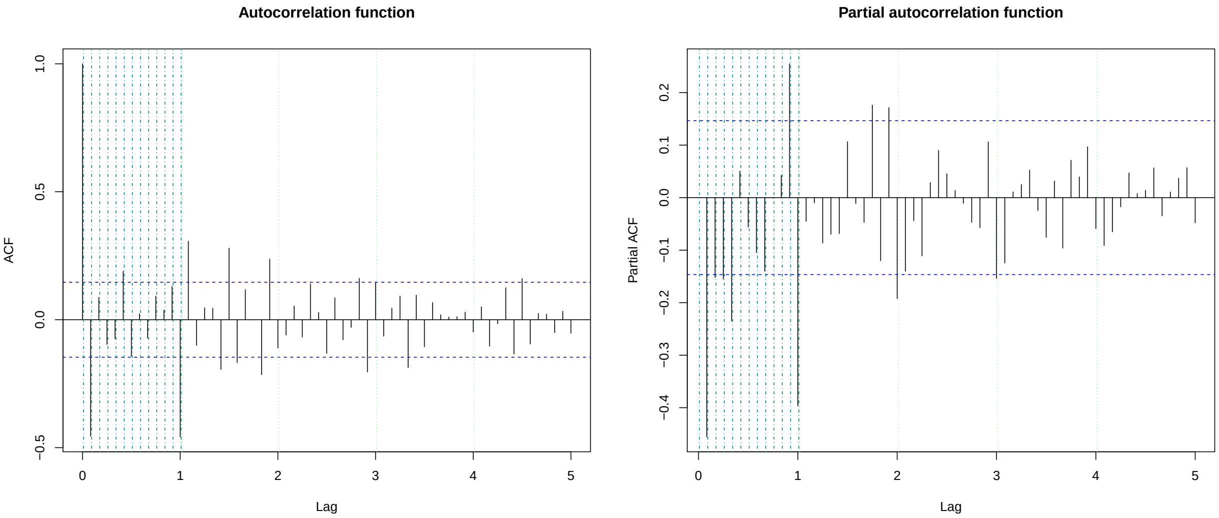


Figure 3: Autocorrelation and Partial function of the stationary component of the data

- the yearly component has a quick decay of the autocorrelation and a relatively slow decay of the partial autocorrelation with the two order significance. Hence, an $MA(2)$ or an $ARMA(1,1)$ may best fit this component.
- about the monthly dependent, we have some kind of exponential decay of the autocorrelation function with the partial autocorrelation that are mostly insignificant. This doesn't allow us to have a clear idea about a specific model that can fit this part. Thus, let's opt for an $ARMA(1,1)$ or $ARMA(1,2)$.

But we should not based the choice of the model on just the visual analysis. Let's analyse the *Akaike information criterion (AIC)* of each suggested model.

3 Model selection and parameters adjustment

Since we have assume that an $MA(2)$ or an $ARMA(1,1)$ may fit the yearly component and an $ARMA(1,1)$ or $ARMA(1,2)$ for the monthly part, we will compare these models AIC throughout a $S-ARIMA(p,1,d) \times (P,1,Q)_{12}$ with max value of p, q, P et Q is 2.

3.1 Model selection

The objective here is to select the model with the smaller AIC and in some extend with the small number of parameters.

```
modele (p,d,q)x(P,D,Q)_saison :  0 1 1 x 0 1 1 _ 12 :  nb param:  2     AIC: 1.637841
modele (p,d,q)x(P,D,Q)_saison :  0 1 2 x 0 1 1 _ 12 :  nb param:  3     AIC: 1.175508
modele (p,d,q)x(P,D,Q)_saison :  0 1 2 x 0 1 2 _ 12 :  nb param:  4     AIC: 2.638954
modele (p,d,q)x(P,D,Q)_saison :  0 1 2 x 1 1 1 _ 12 :  nb param:  4     AIC: 2.694761
modele (p,d,q)x(P,D,Q)_saison :  1 1 1 x 0 1 1 _ 12 :  nb param:  3     AIC: 0
modele (p,d,q)x(P,D,Q)_saison :  1 1 1 x 0 1 2 _ 12 :  nb param:  4     AIC: 1.541156
modele (p,d,q)x(P,D,Q)_saison :  1 1 1 x 1 1 1 _ 12 :  nb param:  4     AIC: 1.587302
modele (p,d,q)x(P,D,Q)_saison :  1 1 1 x 1 1 2 _ 12 :  nb param:  5     AIC: 2.813602
modele (p,d,q)x(P,D,Q)_saison :  1 1 2 x 0 1 1 _ 12 :  nb param:  4     AIC: 0.1927573
modele (p,d,q)x(P,D,Q)_saison :  1 1 2 x 0 1 2 _ 12 :  nb param:  5     AIC: 1.325938
modele (p,d,q)x(P,D,Q)_saison :  1 1 2 x 1 1 1 _ 12 :  nb param:  5     AIC: 2.53819
```

Based on the output of the automatic selection criterion (based on the smaller *Akaike Information Criterion*), we can select two models with a slightly small AIC : $SARIMA(1,1,1) \times (0,1,1)_{12}$ with *AIC = 0* and $SARIMA(1,1,2) \times (0,1,1)_{12}$ with *AIC = 0.19.*

3.2 Parameters analysis

After we have selected the best models (based on the AIC), let's analyse the significance of the parameters of the models.

3.2.1 Model 1 : $SARIMA(1,1,1) \times (0,1,1)_{12}$

Table 2: Parameters of SARIMA(1,1,1)x(0,1,1) s=12

	coef	var.coef	p-value
ar1	0.249	0.013	0.03
ma1	-0.786	0.006	0.00
sma1	-0.928	0.013	0.00

For the $SARIMA(1,1,1)\text{x}(0,1,1)_{12}$, all the parameters are less than one and are significant (Table 2). But notice that the coefficient of the *moving average* part (*-0.928*) of the seasonality is relatively close to the region of non-stationarity.

3.2.2 Model 2 : $SARIMA(1,1,2) \times (0,1,1)_{12}$

Table 3: Parameters of SARIMA(1,1,2)x(0,1,1) s=12

	coef	var.coef	p-value
ar1	-0.890	0.002	0.000
ma1	0.364	0.022	0.015
ma2	-0.636	0.013	0.000
sma1	-0.899	0.008	0.000

For this model $SARIMA(1,1,2) \times (0,1,1)_{12}$, as for the previous one, all its parameters are significant too and the coefficients are not that close to the *non-stationarity* region (see Table 3).

Then, the question is which one to choose among these two ?? Indeed, the first model has less parameters, and it would be a great idea to choose that one, but let's analysis the residuals of the models and their prediction power. This in done in the next session.

4 Model validation

Here, we will analyse the residuals of the two models and see their prediction power.

4.1 Residuals analysis

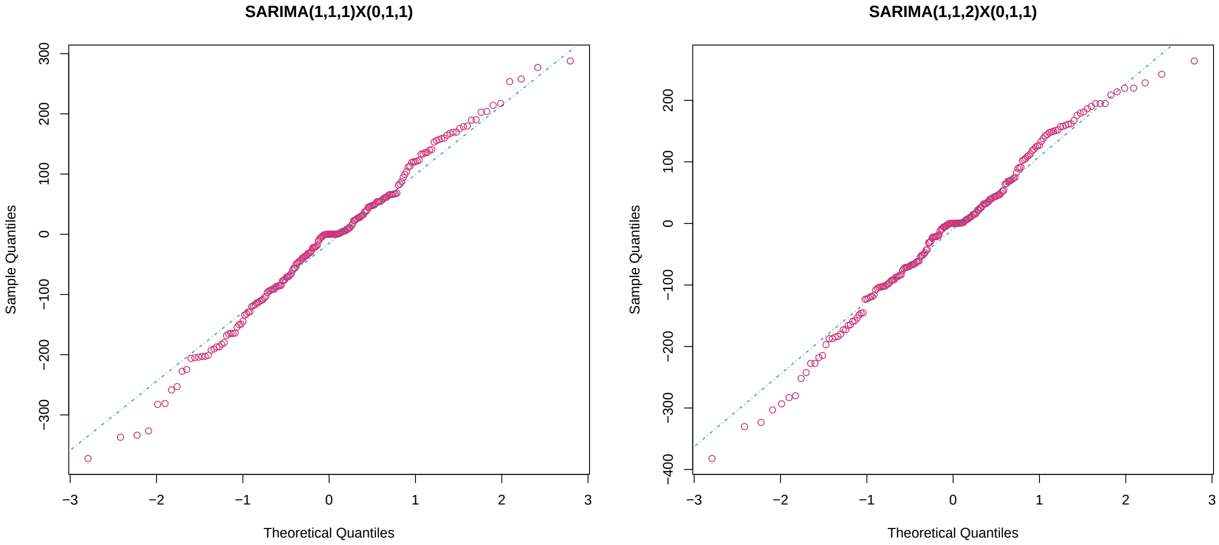


Figure 4: Normal Q-Q Plot of models

The normal QQplot (Figure 4) of the residuals of the two models shows that the residuals mostly aligned with the theoretical normal quantiles (As shown in the plots, the residuals are aligned with the red line). Thus, the model residuals are normally distributed. Now, is there any correlation among the residuals?

Through the Ljung-Box test and the autocorrelation of the standardized residuals (Figure 5 & 6), the residuals are independently distributed so uncorrelated for both models.

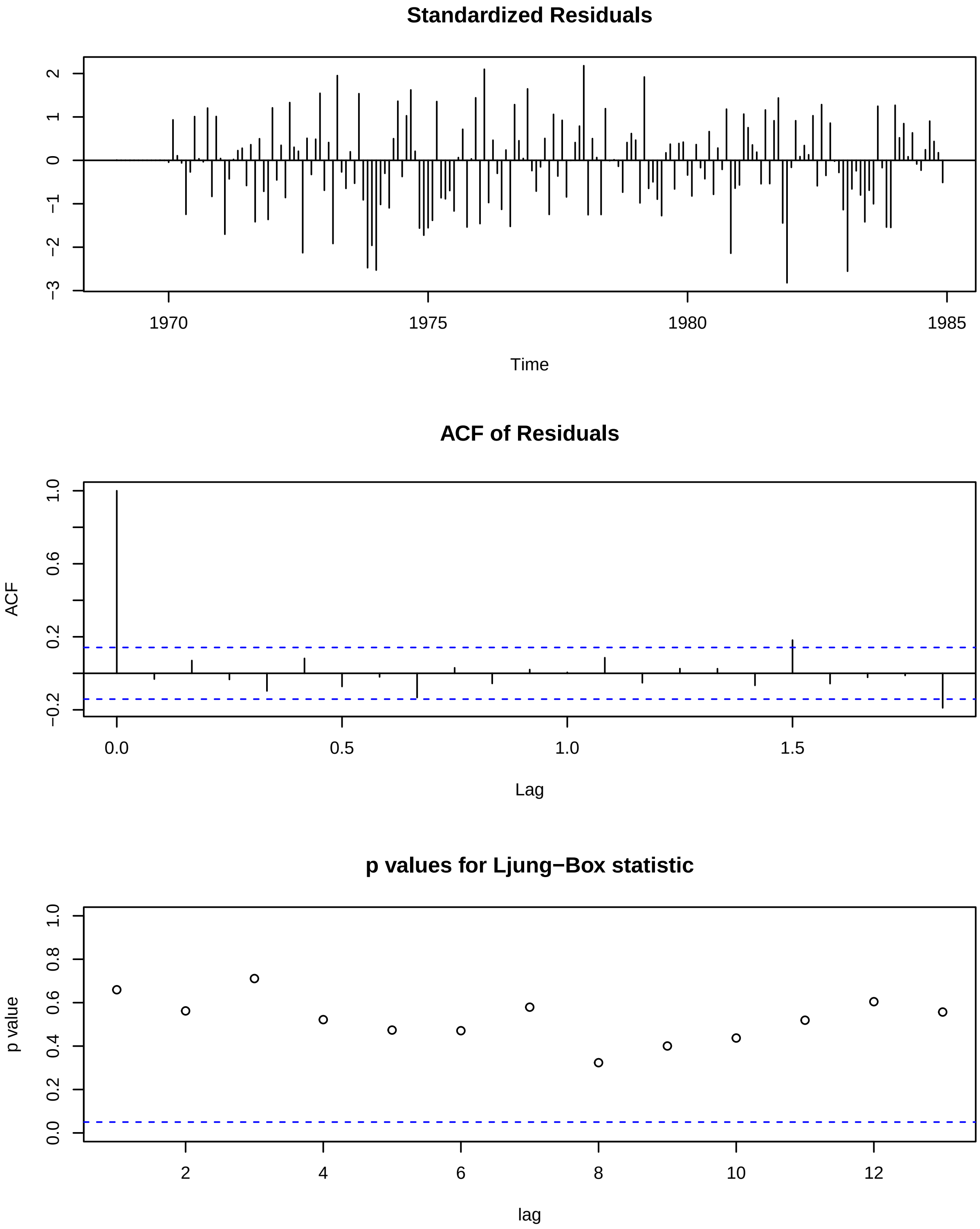


Figure 5: Residuals analysis of SARIMA(1,1,1)X(0,1,1) s=12

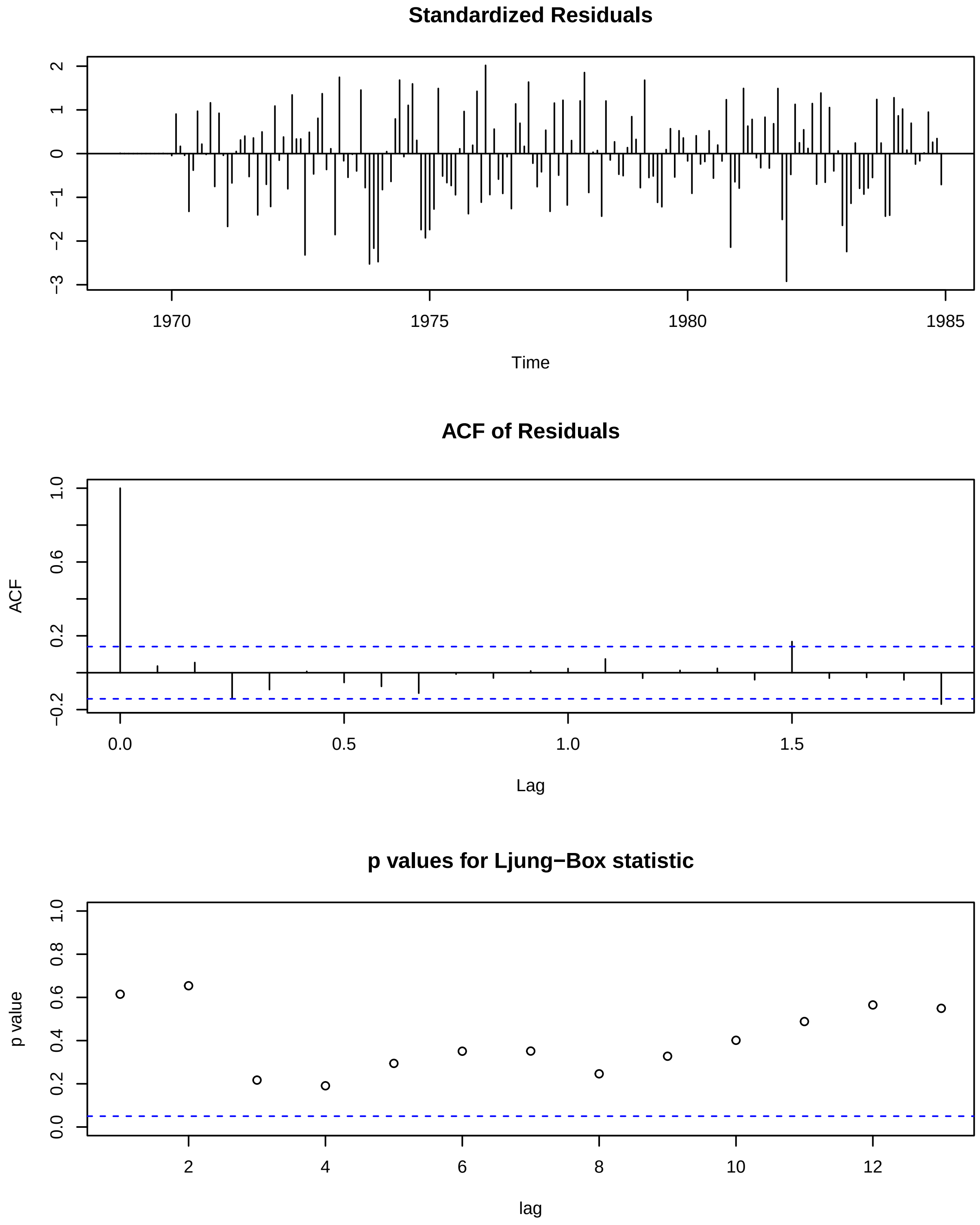


Figure 6: Residuals analysis of SARIMA(1,1,2)X(0,1,1) s=12

Sum up the residuals analysis, for both models $SARIMA(1,1,1)\times(0,1,1)_{12}$ and $SARIMA(1,1,2)\times(0,1,1)_{12}$, their residuals are normally distributed and uncorrelated. That is the good thing we are looking for, but still which one to choose ? Indeed, $SARIMA(1,1,1)\times(0,1,1)_{12}$ has the lower AIC, but is it going to be the best to fit the car drivers data ??

4.2 Prediction error analysis

For the $SARIMA(1,1,1)\times(0,1,1)_{12}$ model, the prediction error is : 1.968632×10^4 and 2.0745978×10^4 for $SARIMA(1,1,2)\times(0,1,1)_{12}$

Though the prediction error analysis, the model 1 : $SARIMA(1,1,1)\times(0,1,1)_{12}$ is definitely the one with less prediction error.

Thus, based on all the analysis above, the model $SARIMA(1,1,1)\times(0,1,1)_{12}$ is the one with: - the smaller AIC, - less *parameters* (so *parsimonious*), - less prediction error - and its residuals are *independently normally distributed* so *uncorrelated* at 5% level of confidence.

It's then the one that can best describe the phenomenon among our data set.

Model specification : Our chosen model is : $SARIMA(1,1,1)\times(0,1,1)_{12}$

Let's :

- X_t be the observed car drivers data at time t
- μ_t, the trend of the time series
- s_t the seasonality part
- $\varepsilon_t \sim WN(0,\sigma^2)$, an innovation part of the data
- Y_t the stationary part (with *mean zero*)of X_t: obtained after removing the trend and the seasonality out of X_t

Thus, we have

$$X_t = \mu_t + s_t + \varepsilon_t$$

and

$$\begin{aligned} Y_t &= \nabla^1\nabla^1_{12}X_t \\ &= (1-B)(1-B^{12})X_t \\ &= (1-B^{12}-B+B^{13})X_t \\ &= X_t - X_{t-1} - X_{t-12} + X_{t-13} \end{aligned}$$

Y_t being an $ARMA$ process, we then have :

$$\begin{aligned} (1-\alpha B)Y_t &= (1-\beta B)(1-\gamma B^{12})\varepsilon_t \\ Y_t - \alpha Y_{t-1} &= (1-\gamma B^{12}-\beta B+\gamma\beta B^{13})\varepsilon_t \\ &= \varepsilon_t - \gamma\varepsilon_{t-12} - \beta\varepsilon_{t-1} + \gamma\beta\varepsilon_{t-13} \\ \implies Y_t &= \alpha Y_{t-1} + \varepsilon_t - \gamma\varepsilon_{t-12} - \beta\varepsilon_{t-1} + \gamma\beta\varepsilon_{t-13} \end{aligned}$$

With the estimated parameters, we finally get :

$$Y_t = 0.25\times Y_{t-1} + \varepsilon_t + 0.93\times\varepsilon_{t-12} + 0.79\times\varepsilon_{t-1} + 0.73\times\varepsilon_{t-13}$$

With X_t, we have :

$$\begin{aligned} \implies X_t = {} & X_{t-1} + X_{t-12} - X_{t-13} + 0.25\times X_{t-1} - 0.25\times X_{t-2} - 0.25\times X_{t-13} + 0.25\times X_{t-14} \\ & + \varepsilon_t + 0.93\times\varepsilon_{t-12} + 0.79\times\varepsilon_{t-1} + 0.73\times\varepsilon_{t-13} \end{aligned}$$

with $\varepsilon_t \sim \mathcal{N}(0,\sigma^2 = 17432)$ (under *5% confidence level*)

5 Prediction

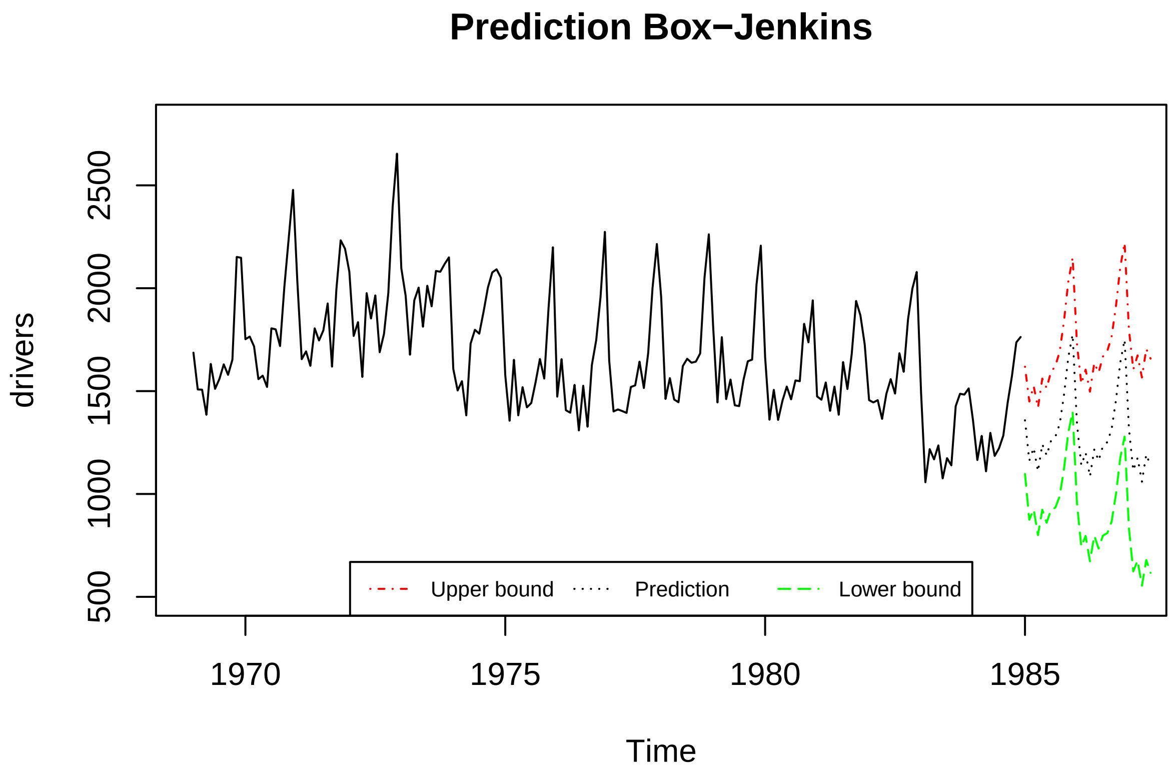
For the prediction, we will adopt two approaches, the parametric and the Holt-Winters approach.

Why these two approach ?

The parametric approach in the prediction that takes into account the structural breaks in the data (or all information available in the data). Where the non parametric one, takes into account just the later observation to predict the future. Nevertheless, looking closely at our data structure (with all observed structural breaks) the parametric approach seems to be the one that can predict the future that will best fit the realities that will be observed.

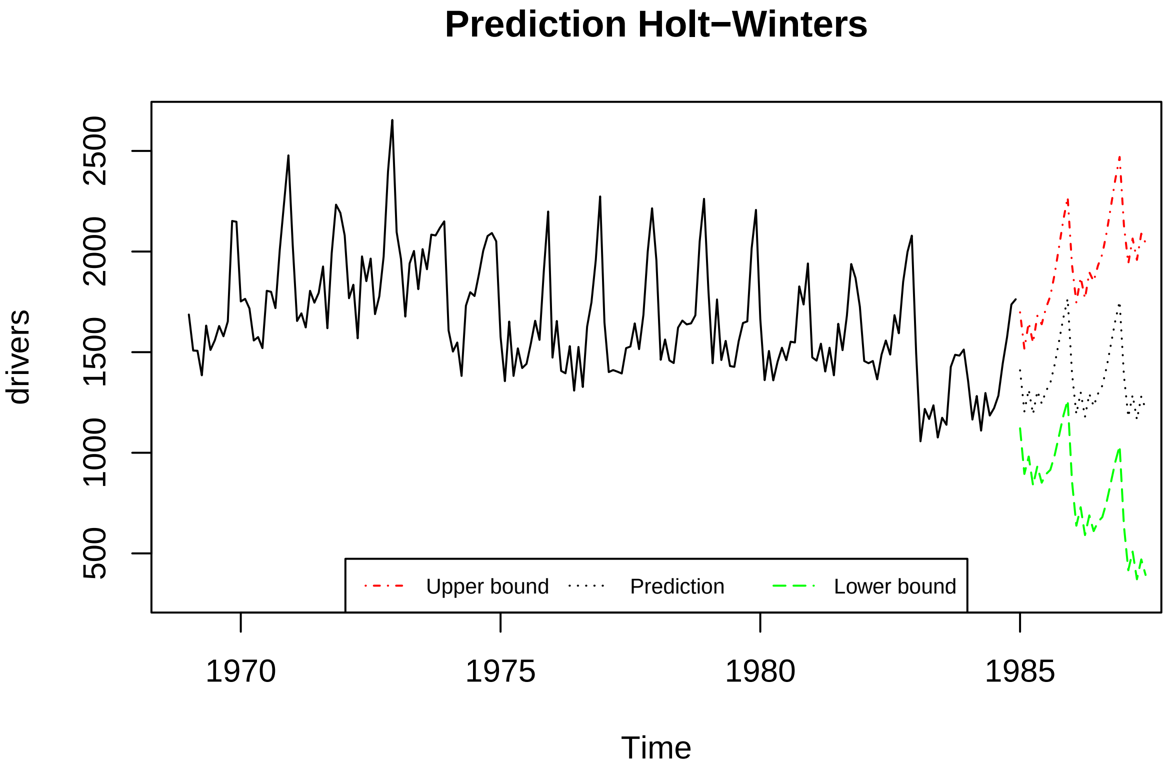
Also, recall that since the *Holt-Winters* approach take into account the later observations, we could get a prediction that goes in same lines with the later observations (don't forget that around the end of the observed data, there is some kind of decay of the number of drivers killed and seriously injured).

5.1 Box-Jenkins approach



As stated, this prediction method gives a prediction that is close to the reality observed from de data, with a prediction intervals that are relatively small.

5.2 Holt-Winters approach



The *Holt-Winters* approach, present a prediction with large prediction interval. Indeed, with this prediction interval, we have large *space of intervention* (if possible to say so), so that the reality will be within this interval.

But looking closely to the structure of the data, we notice that the number of injured drivers is decreasing, so in the future, our expectation is that the number keeps decreasing.

5.3 Comparaison between Box-Jenkins and Holt-Winters

Comparing the two methods (Figure 7), the prediction interval given by *Holt-Winters* is larger than the one given by the *parametric* approach. Such thing because of the fact that the parametric approach takes into account the structure of the data. Then, the fact that by the end of the observed data, there is such a change in the structure of the data lead to such prediction. Also, notice that at some pick (by the end of 1986 and 1987), these two approach give the same prediction (line in gray)

Conclusion

Sum up all we have seen above, the data about the killed and seriously injured car drivers in Great Britain shows the evolution of these numbers from January 1969 to December 1984. From the analysis of this data, we have seen that their numbers trend to decrease during the concerned years, but still with a seasonality. This seasonality can be justify by some periods of the year where accidents are most likely to occur (say around the ends of the years). This data have been modeled by a $SARIMA(1,1,1) \times (0,1,1)_{12}$ with its residuals been independently and normally distributed with *mean* $= 0$, $\sigma^2 = 17432$. A prediction based on parametrics approach and non parametrics approach have been made to predict to future evolution of these numbers.

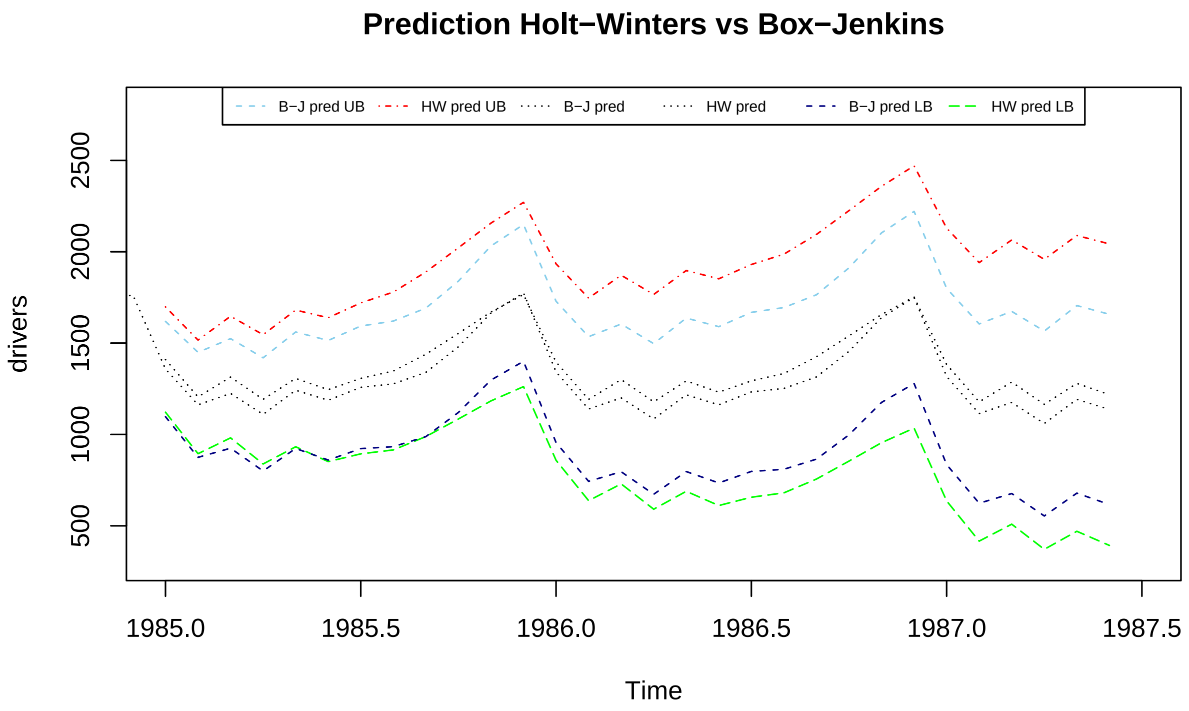


Figure 7: Prediction comparison (Box-Jenkins vs Holt-Winters)