

A Compendium of Reproducible Research about Time Series Analysis

Patrick Wessa

May 4, 2008

Abstract

This document can be used as an introductory, interactive case study about Time Series Analysis based on decomposition, multiple regression and exponential smoothing (including the Holt-Winters model). Section 2 describes the problem and section 3 introduces theoretical concepts that are of importance in applied analysis. Section 4 treats the problem of decomposing a time series into its underlying components (trend, seasonality, and a random component) from an experimental (“try, see, and reflect”) perspective. In section 5 emphasis is on “ad hoc” Forecasting of the time series under investigation, based on regression (section 5.1) and exponential smoothing techniques (section 5.2).

At the same time this document is an illustrated tutorial about Wessa.net and FreeStatistics.org. It illustrates how Reproducible Research can be integrated in Statistics Education based on a Compendium that contains all references to archived information that makes the underlying research reproducible and reusable.

1 Introduction

The computations contained in this document have been computed with the R language [1, software] which is embedded in the R Framework that is freely available on the internet at wessa.net [16, website]. The R Framework allows us to create, publish, and maintain statistical software within a feasible open-access business model (as described in [19, article]). One of the main advantages of using this system is that it allows for the creation of research that can be reproduced and reused by anyone with an internet connection. The computations that were used to write this document have been archived in a newly developed repository at FreeStatistics.org¹.

This document is meant to be reproducible. Therefore it contains citations of all resources that are necessary to reproduce/reuse the underlying computations. This includes articles, spreadsheets, websites, software (R modules), and archived computations.

¹This website contains the Compendium Platform for Reproducible Research and was partially funded by the OOF 2007/13 grant of the K.U.Leuven Association.

2 Case: the Market of Health and Personal Care Products

2.1 Problem

A company sells an exclusive variety of pharmaceutical and health-related products through numerous retail stores in the USA. The production manager needs accurate information about the monthly US retail sales for the purpose of production planning and maintaining an adequate inventory of produced items (to be shipped to customers on demand). The primary objective of this analysis is to:

- gain insight into the dynamics of the Retail Market of Health and Personal Care products
- examine the role of the long-run trend and seasonality
- create extrapolation forecasts without the need of exogenous variables
- evaluate the impact of past events (such as promotions, strikes, etc...)

2.2 Data

The monthly time series under investigation (c.q. “HPC” time series) was stored in the computational archive [2, computation] of FreeStatistics.org. The first observation corresponds to January, 2001 and the last observation to December 2007. The retail sales are measured in millions USD and have not been adjusted for seasonality nor inflation.

The stored time series is readily available and can be instantly viewed or used in statistical analysis as is illustrated by these steps:

1. visit the webpage where the time series has been stored:
<http://www.freeststatistics.org/blog/date/2008/Mar/02/t1204472634kobk5s74i81tzo2.htm>
2. click the “Reproduce” button at the top of the webpage (a new window opens and the archived computation is re-executed)
3. click the “Descriptive Statistics” link (in the right panel of the webpage containing the statistical software)
4. click the “Central Tendency” link
5. observe how the HPC time series is contained in the Data field
6. click the “Compute” button
7. observe how various measures of central tendency are computed for the HPC time series

3 Theoretical concepts

3.1 Equi-distant Time Series

Within the context of this document a time series is defined as a series of chronologically ordered values about some variable of interest that have been observed at regular time intervals. The term “equi-distant” does not imply that the time intervals between observations is exactly equal. Rather it is assumed that we can treat the time intervals as if they were equal because this allows us to employ easier types of analysis.

Here are two examples of non-equi-distant time series:

- quotes and trades on the stock market
- tax changes imposed by the Government

Here are a few examples of equi-distant time series:

- monthly sales data (even if the number of working days is not the same for every month)
- daily returns on the stock market (even if there are no trades on holidays and during the weekend)
- monthly tax levels (sampled at the end of month - even if taxes change the first day of the month)
- monthly (total) number of observed sparrow eggs in a local community (even if many months contain zeroes)

Non-equi-distant time series can be converted to equi-distant time series by means of:

- summing (over regular time intervals)
- averaging (over regular time intervals)
- selection (such as an opening or closing price)
- interpolation (of missing periods)

In any case it is important to understand that the so-called “sampling frequency” causes artifacts in time series that should be treated appropriately. For instance, some business cycles or seasonal effects may be artificially caused by the process that converts a non-equi-distant into an equi-distant time series. In addition, the diagnostic tools that are used to investigate the dynamical properties of time series are (to some extent) artifacts of the sampling frequency. The statistical methods that employ non-equal time intervals are beyond the scope of this document.

3.2 The Autocorrelation Function

The Autocorrelation Function (ACF) of a time series Y_t relates sequential correlations (on the y-axis) $\rho_k = \rho(Y_t, Y_{t-k})$ for $k = 1, 2, \dots, K$ to the time lag k (on the x-axis). Often this is presented in graphical form because it allows to quickly detect patterns that are typical of the dynamical properties of the underlying time series. Note that the time lag k is sometimes called the “order” of autocorrelation.

The sample autocorrelations are computed by the formula

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^T (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^T (Y_t - \bar{Y})^2} \text{ with } \bar{Y} = \frac{1}{T} \sum_{t=1}^T Y_t.$$

The 95% Confidence Intervals (of individual autocorrelation coefficients) are displayed as two horizontal lines, based on the approximated standard deviation $\sigma_{\rho_k} \simeq \frac{1}{\sqrt{T}}$.

In practice the ACF can be used to:

- describe/summarize autocorrelation at various orders
- identify non-seasonal and seasonal trends
- identify various types of typical patterns that correspond to well-known forecasting models
- check the independence assumption of the residuals of regression and forecasting models

4 Decomposition of Time Series

The decomposition of a time series may provide insight into the relative importance of the long-run trend and seasonality. The production manager may be interested in the seasonal component because it shows how the total variance of sales depends on seasonal fluctuations. If the rate of production is nearly constant and if it is necessary to deliver goods on demand (retailers don't want to keep a large inventory) then it is obvious that the company needs to stock overproduction (in months with low demand) to meet the orders in months with high demand (and insufficient production capacity).

The variance of the seasonal component may be an important factor that determines the size of inventory (at least if we assume a constant market share). The behavior of trends and business cycles on the other hand, may be important factors in making strategic decisions on the medium or long run.

4.1 Classical Decomposition of Time Series by Moving Averages

Model Classical decomposition of time series can be performed by Moving Averages. The definitions of Moving Averages are treated in more detail in section 5.2, and a spreadsheet implementation of the classical decomposition is discussed in [22]. At this stage it is sufficient to define

- the time series under investigation as Y_t ,

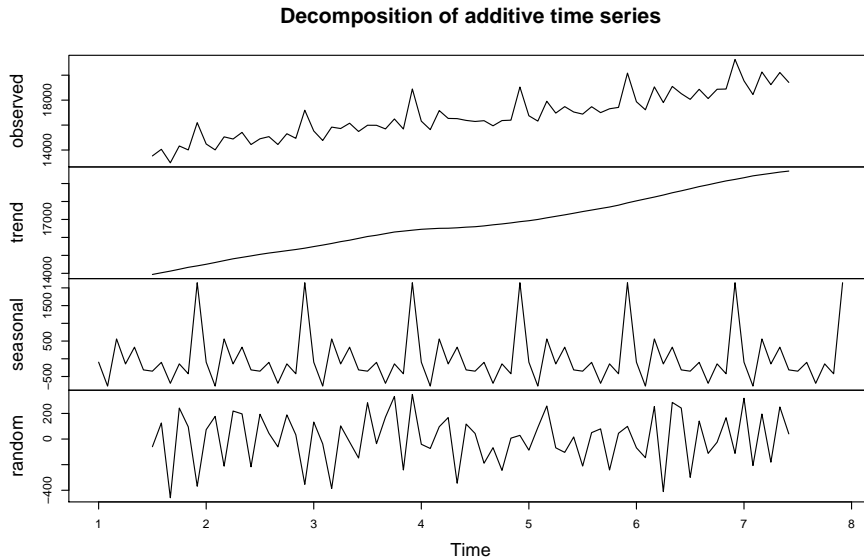


Figure 1: Classical Decomposition - US Retail Sales of Health and Personal Care

[link of reproducible computation](#)

- the long-run trend as L_t ,
- the seasonal component as S_t ,
- and the error as e_t

Now the additive model can be written as $Y_t = L_t + S_t + e_t$ for $t = 1, 2, \dots, T$ and the multiplicative model as $Y_t = L_t * S_t * e_t$ for $t = 1, 2, \dots, T$.

Analysis The HPC time series was analyzed by the “Classical Decomposition R module” that is available on-line [14, software] and makes use of a script² that is based on the well-known R language [1, software]. The result of the additive, classical decomposition analysis [3, computation] (click to reproduce) clearly shows that a positive trend is present (see Figure 1). In addition, a strong seasonal component (showing a distinctive, regular pattern) is present which leads us to believe that it should be possible to generate predictions about the time series based on historical information.

A detailed look at the components that are listed in the table of the archived analysis [3, computation], reveals that the trend component is much more important than seasonality: for example, the 7th row (c.q. July 2001) shows that the trend component has a value of 13937.96 whereas the seasonal component is only -349.25. According to this analysis the trend accounts for roughly 97% of the predicted value. This only leaves about 3% to be explained by seasonality.

²Note that the underlying R code is readily available in each on-line R module. If the R module is executed you will be able to view the input R code and the resulting output from the R engine by clicking the appropriate hyperlinks in the result page.

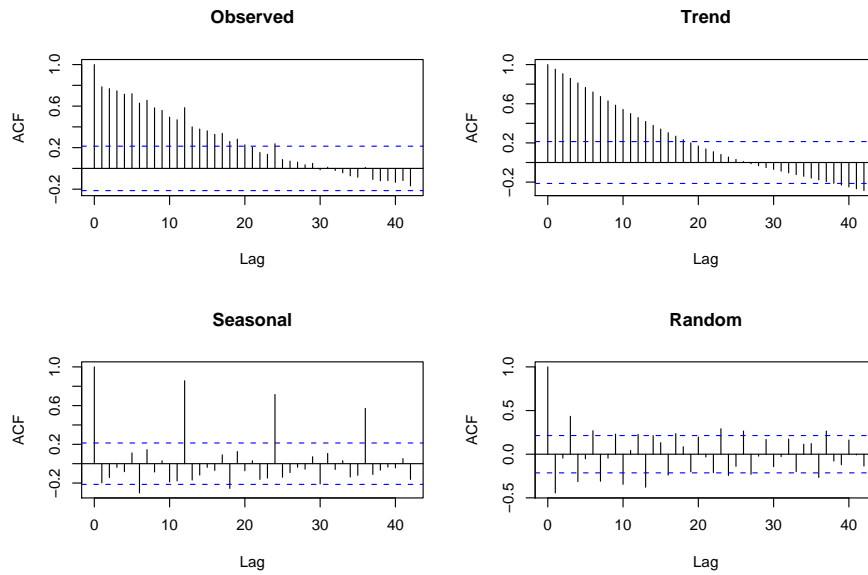


Figure 2: ACF of Classical Decomposition - US Retail Sales of Health and Personal Care

[link of reproducible computation](#)

Interpretation A much better way to identify the previously mentioned properties is to make use of the Autocorrelation Function (ACF). The ACF of the Classical Decomposition in Figure 2 clearly shows some remarkable features:

- The trend component exhibits a slowly (linearly) decreasing series of autocorrelation coefficients. This is the typical pattern for a long-run trend³.
- The seasonal component shows slowly (linearly) decaying autocorrelation coefficients (all of which are significantly positive) at time lags $k = 12, 24, 36, \dots$. This is the typical pattern of a so-called seasonal trend (c.q. strong seasonality).
- Finally, the error component⁴ displayed in Figure 2 shows a (regular) series of non-zero autocorrelations (significant at the 5% type I error level). This is important because it implies that the prediction errors of the Classical Decomposition model are not independent (they are in fact autocorrelated). In other words: past prediction errors contain systematic information that can be used to predict future prediction errors - hence, they can also be used to improve the forecasts of the time series.

Conclusion There are two conclusions from this analysis:

³It does not matter if the trend is positive or negative - the ACF pattern is the same in both cases.

⁴In the R module this is called the “random” term.

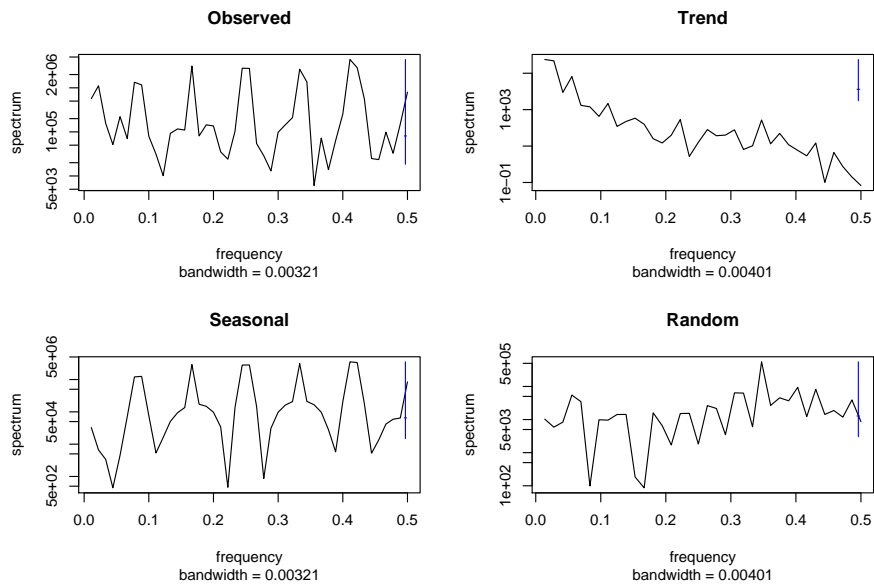


Figure 3: Spectrum of Classical Decomposition - US Retail Sales of Health and Personal Care

[link of reproducible computation](#)

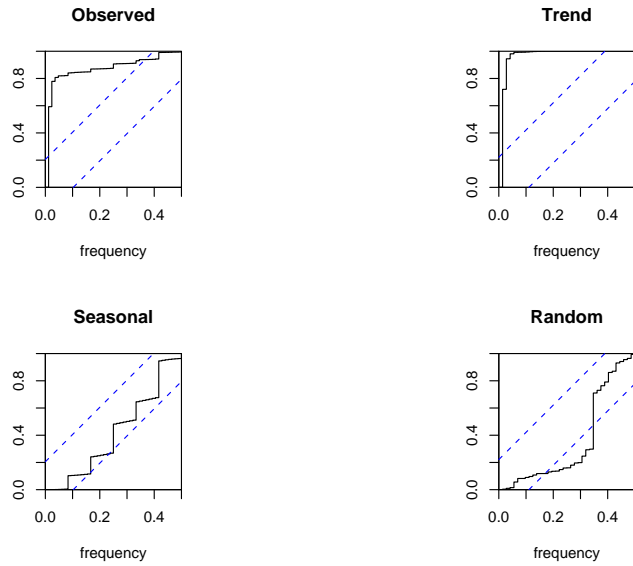


Figure 4: Cumulative Periodogram of Classical Decomposition - US Retail Sales of Health and Personal Care

[link of reproducible computation](#)

- A forecasting model that is based on Moving Averages (as defined in section 5.2) is incomplete and should not be trusted by the production manager to make predictions. We will verify this statement in section 5.2.2.
- The model clearly indicates the presence of a long-run trend and a strongly seasonal pattern that can be predicted (but we need a more sophisticated model).

Assignment Compare the additive and multiplicative models.

- Which model is better?
- When would you (theoretically) expect a large difference between the additive and multiplicative model⁵? Apply additive and multiplicative decomposition to the Airline time series [7, computation] (click to reproduce) to empirically verify the theoretical answer⁶.

4.2 Seasonal Decomposition of Time Series by Loess

Model reference Seasonal Decomposition by Loess is more sophisticated than the previous method and is described in full details in [21, article].

Analysis The HPC time series was analyzed by the R module “Decomposition by Loess” that is available on-line [15, software]. Based on the default parameters, the Decomposition by Loess analysis [4, computation] (click to reproduce) clearly confirms the positive long-run trend (compare Figure 1 and 5). Similarly, the seasonal component shows a distinctive, regular pattern just like in the previous model.

Assignment Compare the analysis of Classical Decomposition and Seasonal Decomposition by Loess. Answer the following questions for both methods:

- Is the trend component more important than the seasonal?
- Describe the typical pattern of the Autocorrelation Function for all components (see Figure 6)
- Interpret Figures 7 and 8 in your own words
- Is the second (c.q. more sophisticated method) doing a better job?⁷

⁵Hint: think about how the seasonal pattern may become stronger (on the long run) as the number of consumers in the HPC market grows. In other words, there may be a relationship between the long-run trend and the seasonal component. This question is also closely related to the concept of “Heteroskedasticity”.

⁶The Airline Passenger time series is - unlike the HPC time series - characterized by the fact that seasonality becomes stronger as the level of Airline passengers grows. There is a direct relationship between trend and seasonality which is a special case of “Heteroskedasticity”.

⁷Hint: the words “random” and “remainder” both relate to the prediction error (they are synonyms).

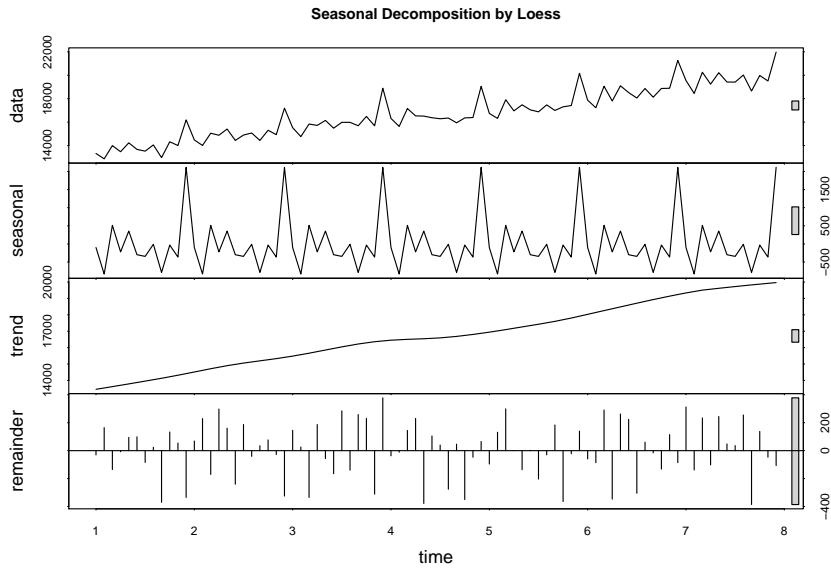


Figure 5: Seasonal Decomposition by Loess - US Retail Sales of Health and Personal Care

[link of reproducible computation](#)

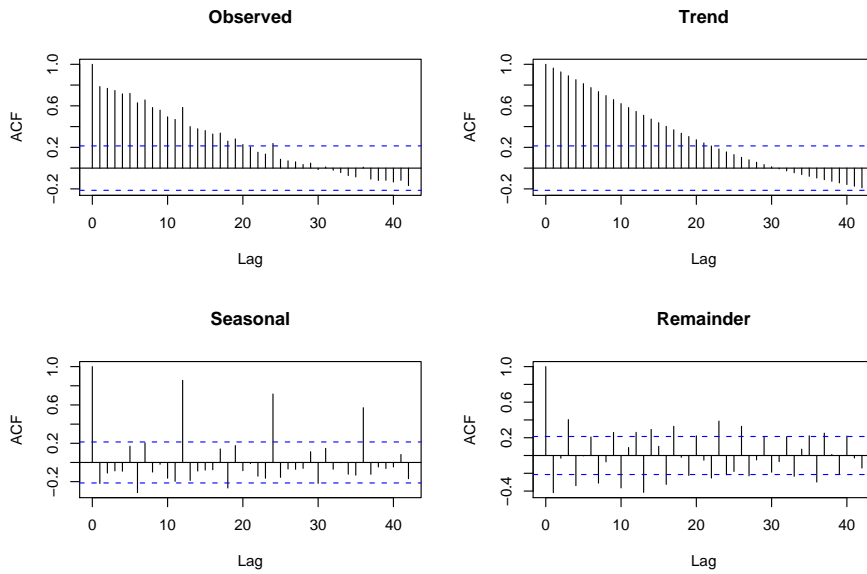


Figure 6: ACF of Decomposition by Loess - US Retail Sales of Health and Personal Care

[link of reproducible computation](#)

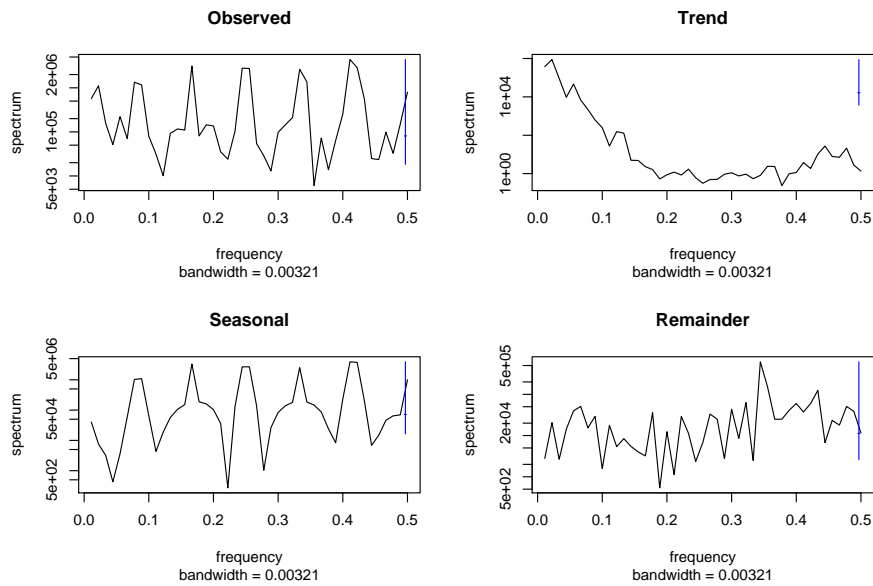


Figure 7: Spectrum of Decomposition by Loess - US Retail Sales of Health and Personal Care

[link of reproducible computation](#)

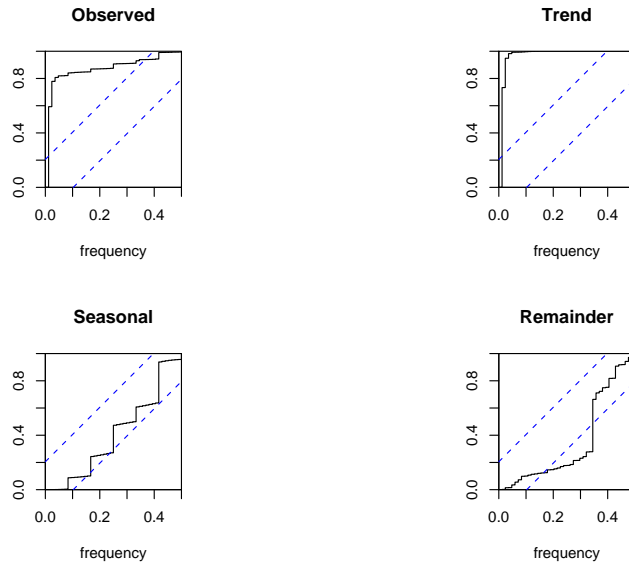


Figure 8: Cumulative Periodogram of Decomposition by Loess - US Retail Sales of Health and Personal Care

[link of reproducible computation](#)

- Recompute the Decomposition by Loess⁸ with a Seasonal Window = 6. Does the new result change the answer for the previous question? Explain.
- The R module for Decomposition by Loess does not display a histogram of the prediction error. Adapt the underlying R code of the R module to include a residual histogram and recompute the analysis by adding the following code segment at an appropriate place of the R code:

```

bitmap(file="test5.png")
myresid <- m$time.series[!is.na(m$time.series[, "remainder"]), "remainder"]
hist(as.numeric(myresid), main="Residual Histogram", xlab="Residual Value")
dev.off()

```

What can you conclude about the distribution of the prediction error? How does the residual distribution change when the Seasonal Window = 6?⁹

4.3 Decomposition by Structural Time Series Models

Structural Decomposition is yet another approach that extracts time series components about seasonality and the long-run trend. There are three flavors of Structural Decomposition available: the local level, the local trend, and the basic structural model.

4.3.1 The local level model

The local level M_t is updated according to the relationship $M_{t+1} = M_t + \xi_t$ with $\xi_t \sim N(0, \sigma_\xi^2)$. The time series Y_t is modeled¹⁰ by $Y_t = M_t + \epsilon_t$ with $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. There are two parameters to be estimated: σ_ξ^2 and σ_ϵ^2 .

4.3.2 The local trend model

In the local trend model the updating mechanism is described by $M_{t+1} = M_t + N_t + \xi_t$ with $\xi_t \sim N(0, \sigma_\xi^2)$ and by $N_{t+1} = N_t + \zeta_t$ with $\zeta_t \sim N(0, \sigma_\zeta^2)$. The time series Y_t is modeled by $Y_t = M_t + \epsilon_t$ with $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. There are three parameters to be estimated: σ_ξ^2 , σ_ζ^2 and σ_ϵ^2 . It is obvious that the local level model is just a special case of the local trend model.

4.3.3 The basic structural model

Model The previous Structural Models are not useful for the purpose of decomposition because there is no seasonal component. The third model however is a local trend model with an added seasonal component: $Y_t = M_t + S_t + \epsilon_t$ with $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. The seasonal component is defined by: $S_{t+1} = -S_t - \dots - S_{t-s+2} + \omega_t$ with $\omega_t \sim N(0, \sigma_\omega^2)$

⁸Hint: archive the computation and add the hyperlink in your reply so that the reader can verify your analysis.

⁹You can verify your adaptation of the R module by comparing your result with [5, computation].

¹⁰Note that this model is a special case of an ARIMA model with restricted parameters.

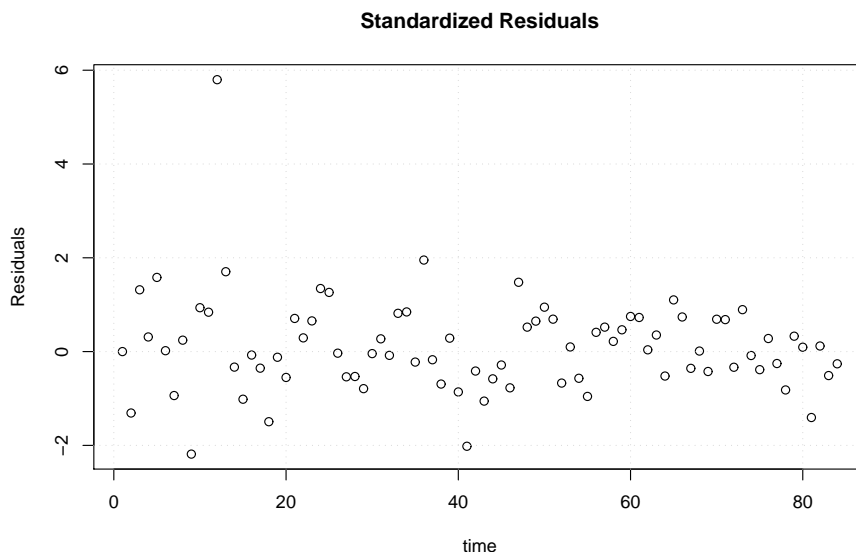


Figure 9: Residuals of Structural Time Series Model - US Retail Sales of Health and Personal Care

[link of reproducible computation](#)

Analysis Figure 9 shows that the basic structural model seems to fit the HPC time series well. The standardized residuals exhibit a mildly regular pattern and a single outlier. In Figure 10 the typical patterns of the ACF can be clearly seen in all components. The model is not perfect because the ACF of the standardized residuals suggests the presence of a mild form of residual autocorrelation.

More advanced analysis Readers who are familiar with spectral analysis may find Figures 11 and 12 interesting. The level component exhibits a typical pattern where emphasis is on the short frequency part of the spectrum (waves with long periods are dominant). The cumulative periodogram of this long-run component falls outside of the 95% Kolmogorov-Smirnov confidence intervals (two parallel, dashed lines). In addition, the presence of seasonality (in the seasonal component) is also very clear: the peaks in the spectrum and stair-jumps in the cumulative periodogram are both very typical for strong seasonality - observe how these peaks and stair-jumps coincide with the seasonal frequencies ($\frac{1}{12} = 0.08, \frac{1}{6} = 0.18, \frac{1}{4} = 0.25, \frac{1}{3} = 0.33$).

Interpretation The fact that the seasonal cumulative periodogram falls inside the confidence intervals does by no means imply that the seasonal pattern is not present. The same remark can be made about the ACF of the standardized residuals: most autocorrelation coefficients fall inside of the 95% confidence intervals, yet we diagnose a mild form of autocorrelation. How can that be if (almost) all autocorrelations are not significantly different from zero?

This question is often puzzled over - the answer however, is fairly simple:

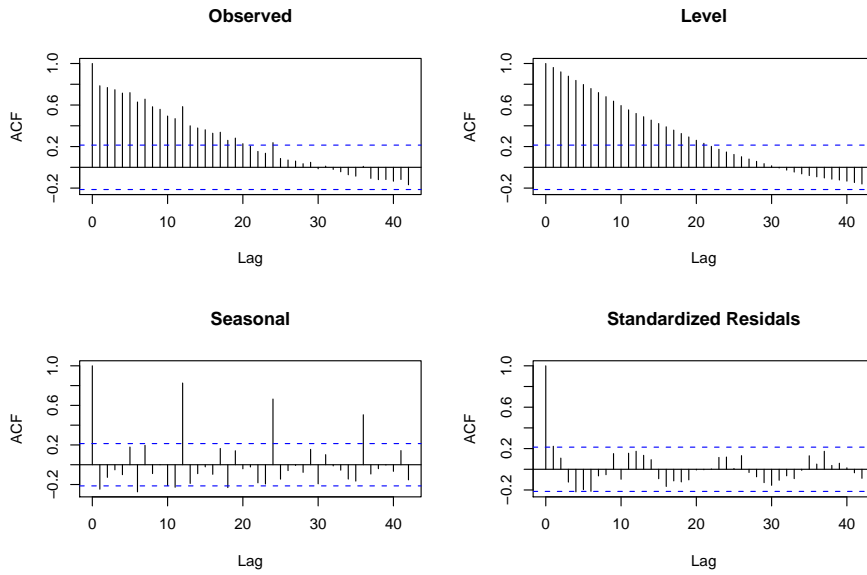


Figure 10: ACF of Structural Time Series Model - US Retail Sales of Health and Personal Care

[link of reproducible computation](#)

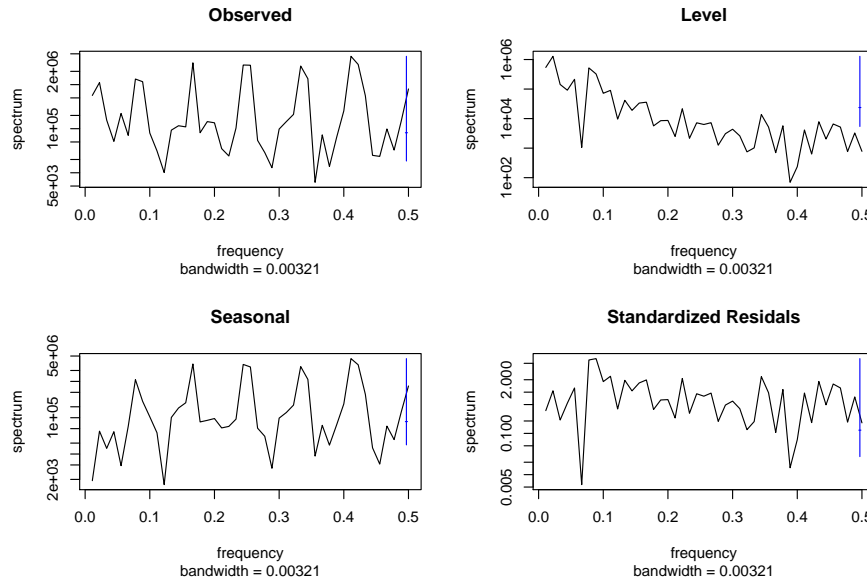


Figure 11: Spectrum of Structural Time Series Model - US Retail Sales of Health and Personal Care

[link of reproducible computation](#)

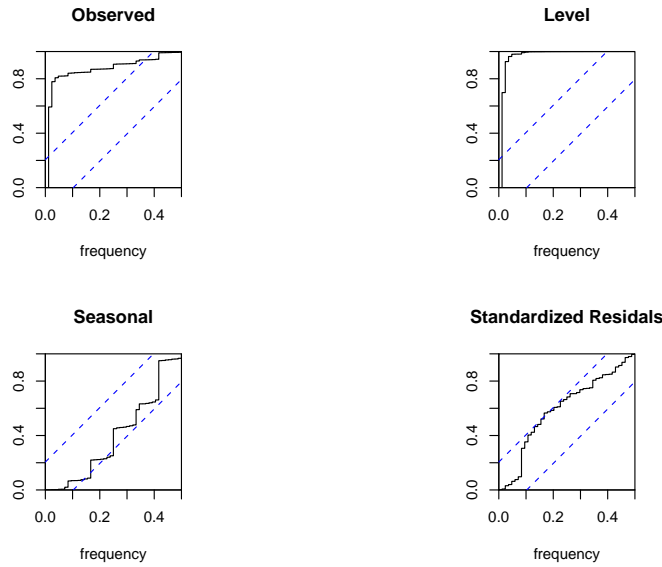


Figure 12: Cumulative Periodogram of Structural Time Series Model - US Retail Sales of Health and Personal Care
[link of reproducible computation](#)

the Kolmogorov-Smirnov intervals and the confidence intervals of the ACF must not be interpreted in terms of (simultaneous) significance of a group of autocorrelations or sinusoid waves. In other words, the confidence interval of the ACF relates to the hypothesis $H_0 : \rho_k = 0$ versus $H_1 : \rho_k \neq 0$ for $k = 1, 2, \dots, K$. It does not relate to the hypothesis $H_0 : \rho_k \neq f(\rho_{k-1}, \rho_{k-2}, \rho_{k-3}, \dots)$ versus $H_1 : \rho_k = f(\rho_{k-1}, \rho_{k-2}, \rho_{k-3}, \dots)$ where $f(\cdot)$ is a regular function that leads to regular patterns in the ACF¹¹.

Assignment Examine and compare all decomposition models based on the empirical results about the HPC time series.

- Make a comparison based on the residual ACF of each model
- Make a comparison based on the residual spectrum and cumulative periodogram of each model
- What advice can you give the production manager?

Assignment Examine the residual diagnostics of the structural time series model in Figure 13.

- Are the standardized residuals normally distributed? Explain.
- Identify the outlier and recompute the analysis after neutralizing the outlier (by interpolation). Do this by inserting a line into the R code:

¹¹A similar argument can be made about the cumulative periodogram.

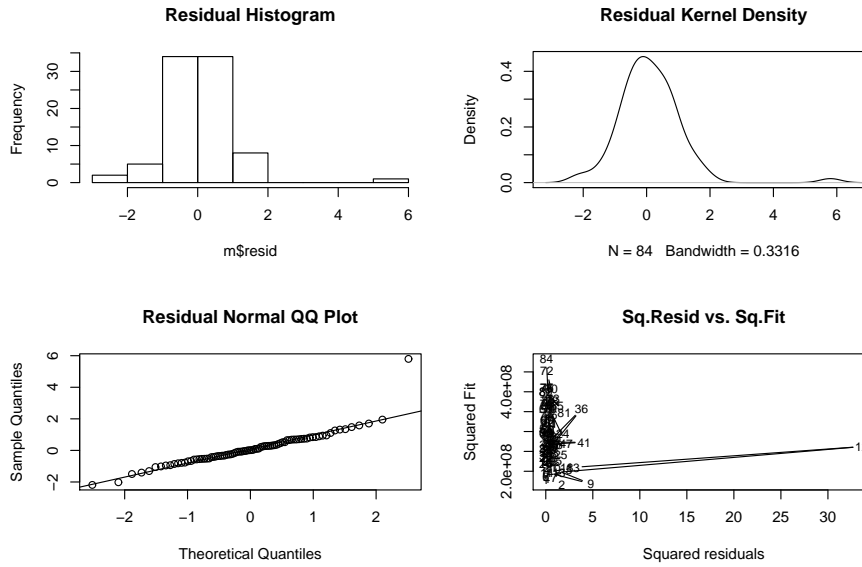


Figure 13: Residual Diagnostics of Structural Time Series Model - US Retail Sales of Health and Personal Care
[link of reproducible computation](#)

`x[???` = ???

where ? has to be replaced with meaningful code.

5 Ad hoc Forecasting of Time Series

The long-run trend and the seasonal component can both be used to make predictions $\hat{Y}_{t=j}$ of a time series (see sections 5.1 and 5.2). A model that generates such predictions (c.q. Forecasting model) is based on a formula that estimates future outcomes by a weighted sum of past observations $Y_{t<j}$, artificially constructed variables (deterministic components - see section 5.1), and previous predictions $\hat{Y}_{t<j}$ (for any period for which observations are not available). In the case where we only use historical information we can write:

$\hat{Y}_t = \alpha + \sum_{i=1}^{t-1} (\pi_i \hat{Y}_i)$ for $t \in \mathbb{N}$ and with $\hat{Y}_i = Y_i$ for any $i \leq T$. The weights π_i that are used in this formula are parameters that are computed by a statistical estimation algorithm.

In practice however, the estimation of a large number of parameters π_i is not feasible because:

- the estimation of a large number of parameters is computationally difficult
- for mathematical reasons we need more observations than parameters to make the estimation method work

- the “principle of parsimony” warns us about building models with too many parameters (“overfitting”)

Therefore, many suggestions have been made in scientific literature to define parsimonious prediction models where the π_i weights are generated by a function with only few parameters. This subject will be treated in more detail in section 5.2

The case where the Forecasting model is based on artificially constructed variables is considered in the next section.

5.1 Regression Analysis of Time Series

Model Assumptions Regression Analysis of time series is - within the context of this document - based on the following assumptions:

- there is a fixed linear trend
- the fixed seasonal effect is modeled as a shift of the constant term (for each seasonal period)

In other words, the exogenous variables that are employed in these regression models are artificially constructed and deterministic in nature. The seasonal component is modeled by so-called “seasonal dummies” which is similar to the basic structural model with $\sigma_\omega^2 = 0$. The trend is represented by time t .

Analysis The following model was fitted to the HPC time series: $Y_t = \alpha + \beta t + \sum_{i=1}^{s-1} \gamma_i D_{i,t} + e_t$ for $t = 1, 2, \dots, T$ where $e_t \sim N(0, \sigma_{e_t}^2)$.¹² The so-called seasonal dummies are binary variables where $D_{i,t} = 1$ if $t \bmod s = i$ and $D_{i,t} = 0$ otherwise.

The analysis was performed by making use of the multiple regression R module [17, software]. In normal circumstances the data have to be entered as a multivariate dataset. The easiest way to do this is to copy the multivariate dataset from a spreadsheet and paste it into the Data X textbox of the R module. In this case however, we only need a single column (containing the observations of the endogenous time series) because the seasonal dummies and the linear trend are automatically generated by the software (if the parameters are set appropriately - see Figure 14).

Interpretation The result is interesting for the production manager because the explicit regression formula allows one to make predictions about the HPC time series’ future. The explicit regression model¹³ can be obtained by substituting the “Greek” parameters by their estimates: $\hat{Y}_t = 15515.916666667 - 2132.84102182539D_{1,t} - 2859.02430555556D_{2,t} - 1507.77901785714D_{3,t} - 2251.39087301588M_{4,t} - 1687.43129960317D_{5,t} - 2357.90029761905D_{6,t} - 2422.36929563492D_{7,t} - 2104.69543650794D_{8,t} - 2896.45014880952D_{9,t} - 2143.91914682540D_{10,t} - 2478.67385912699D_{11,t} + 77.7547123015873 t$ (cfr. [8, computation], click to reproduce)

To better understand the structure of this model it is good to have a look at how the R engine computes the regression model. In the archived computation

¹²In this case $s = 12$ because we use monthly data

¹³Note that in the R module the seasonal dummies are denoted by M instead of D.

14483	
14011	
15057	
14884	
Names of X columns:	
HPC	
Sample Range: (leave blank to include all observations)	
From:	<input type="text"/>
To:	<input type="text"/>
Column Number of Endogenous Series (?)	
<input type="text" value="1"/>	
Fixed Seasonal Effects	
Include Monthly Dummies	
Type of Equation	
Linear Trend	
Chart options	
Width:	<input type="text" value="600"/>
Height:	<input type="text" value="400"/>
<input type="button" value="Compute"/>	

Figure 14: Data Entry - Multiple Regression R module
link of R module

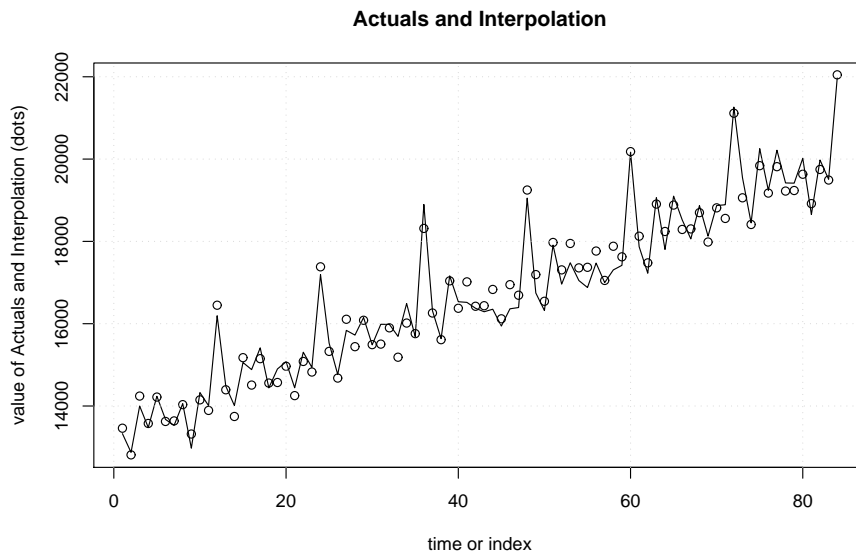


Figure 15: Fit of multiple regression - US Retail Sales of Health and Personal
Care

link of reproducible computation

[8, computation] there is a hyperlink (click to open the R output) that points to the raw output that was generated by the R server¹⁴. In this output one can clearly see how the data matrix of the regression was formed

```
> x
      HPC M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11  t
1  13328  1  0  0  0  0  0  0  0  0  0  0  0  1
2  12873  0  1  0  0  0  0  0  0  0  0  0  0  2
3  14000  0  0  1  0  0  0  0  0  0  0  0  0  3
4  13477  0  0  0  1  0  0  0  0  0  0  0  0  4
5  14237  0  0  0  0  1  0  0  0  0  0  0  0  5
6  13674  0  0  0  0  0  1  0  0  0  0  0  0  6
7  13529  0  0  0  0  0  0  1  0  0  0  0  0  7
....
```

It is now obvious that the prediction for $t = 1$ is $\hat{Y}_t \simeq 15515.92 - 2132.84D_{1,t} + 77.76t$ or simply $\hat{Y}_1 \simeq 15515.92 - 2132.84 + 77.76$. Similarly, we can compute the predicted value for any $t = 1, 2, \dots, T + H$ where H is the forecast horizon (Figure 15 shows the interpolation fit of the regression model). The interpretation of the γ_i parameters is the amount by which the constant term is in/decreased in month i .

Predictions made for months $i = 12, 24, 36, \dots$ don't have a dummy term. For instance, the forecast for $t = 12$ is $\hat{Y}_t \simeq 15515.92 + 77.76t$ or simply $\hat{Y}_{12} \simeq 15515.92 + 77.76 * 12$. The estimated constant term $\hat{\alpha} \simeq 15515.92$ can be interpreted as the retail sales figure in the month of December¹⁵ while disregarding the long-run trend effect. Hence, all seasonal effects are expressed in relationship with the month of reference (December). Since all $\hat{\gamma}_i < 0$ we may conclude that the sales figures in January-October are lower than in the December month of the same year. The parameter $\hat{\beta} \simeq 77.76$ implies that each month the retail sales increase by 78 million USD (on average).

The diagnostics about the residuals (Figures 16, 17, and 18) however, clearly suggest that there are several problems¹⁶.

Conclusion The regression model has clear advantages because it allows us to measure the effect of occurrences of arbitrary events that can be represented

¹⁴Readers who have R installed on their computer might also consider to paste the R input code (click to open the R input) into the R interpreter.

¹⁵Actually we know that the 12th month is the month of December because the HPC time series starts in January 2001.

¹⁶Not all the residual diagnostics are shown in this document. Refer to the archived computation to investigate all results.

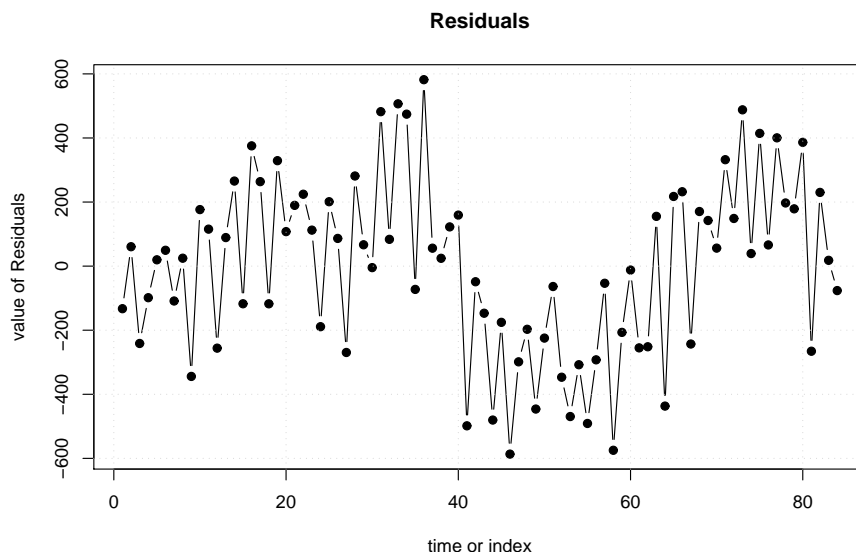


Figure 16: Residuals of multiple regression - US Retail Sales of Health and Personal Care

[link of reproducible computation](#)

by a dummy variable. For instance, the production manager may not only be interested in seasonal and trend effects - but also in the effect of sudden health care reforms that change the expenditures in the retail market. One only needs to introduce an additional dummy variable that represents the event under study. A few examples may illustrate the use of such a dummy variable (also called “intervention variable” η_t) in the extended model $Y_t = \alpha + \beta t + \sum_{i=1}^{s-1} \gamma_i D_{i,t} + \omega \eta_t + e_t$ for $t = 1, 2, \dots, T$ where $e_t \sim N(0, \sigma_{e_t}^2)$:

- During a short one-week trial (in month $t = j$) various Public Health Agencies launched a campaign to decrease unnecessary antibiotic use. This intervention could be coded as $\eta_j = 1$ and $\eta_{t \neq j} = 0$. We would expect that the estimated intervention effect $\hat{\omega} < 0$.
- Structural reforms of the health insurance system (as of $t \geq j$) makes a series of new drugs and health-related services available for a large population. This could be coded as a “step variable” $\eta_{t < j} = 0$ and $\eta_{t \geq j} = 1$. We would expect that the estimated intervention effect $\hat{\omega} > 0$.

To include an additional dummy variable into the model one may simply create a spreadsheet that contains the endogenous variable together with the intervention variables. An example is available in GoogleDocs [13, spreadsheet] (click to open).

Assignment Investigate the output of the archived computation and answer the following questions.

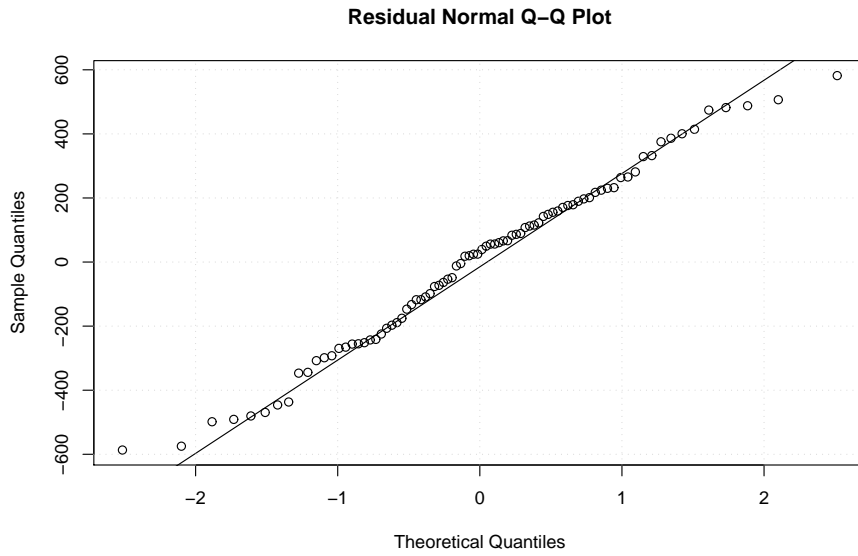


Figure 17: Residual Normal QQ Plot of multiple regression - US Retail Sales of Health and Personal Care
[link of reproducible computation](#)

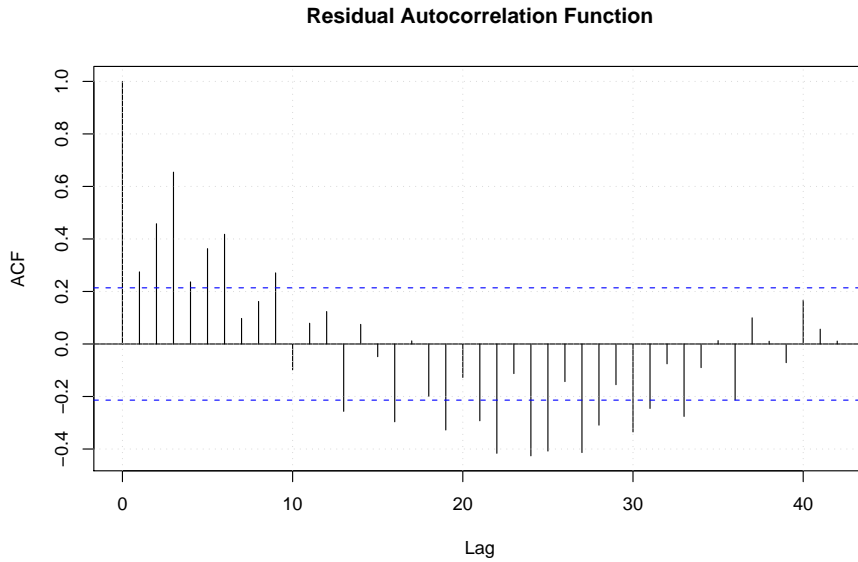


Figure 18: Residuals ACF of multiple regression - US Retail Sales of Health and Personal Care
[link of reproducible computation](#)

- Are the parameters of the regression model significantly different from zero?
- Does the answer to the previous question change if you would use a type I error of only 1%?
- Which two months have to lowest HPC retail sales?
- Which is the second best month (in terms of sales)?
- Does the regression model have a good fit? Explain.
- Carefully examine the various residual diagnostics of regression model. Are the underlying model assumptions satisfied?
- In section 4.3.3 we identified an outlier at $t = 12$. Does the multiple regression model confirm that Y_{12} is in fact an outlier? Explain your answer based on residual diagnostics and a regression computation with an added dummy variable¹⁷.

5.2 Smoothing Models

Various types of exponential smoothing models are available for forecasting purposes. The main difference between smoothing models and the previously described regression models is that smoothing relates to past observations whereas the classical regression model assumes deterministic exogenous variables.

Some readers might wonder why it is not possible to specify a regression equation of the form

$$\begin{cases} \hat{Y}_t = \alpha + \beta t + \sum_{i=1}^{s-1} \gamma_i D_{i,t} + \sum_{i=1}^K (\phi_i \hat{Y}_{t-i}) & K < s \\ \hat{Y}_t = \alpha + \beta t + \sum_{i=1}^K (\phi_i \hat{Y}_{t-i}) & K \geq s \end{cases}$$

The answer is that the classical regression model assumes deterministic (so-called “controlled”) exogenous variables - an assumption that is clearly violated when the endogenous variable is used at the right hand side of the regression equation. The classical assumptions can be relaxed (to allow for the lagged endogenous time series to appear as explanatory variable in the equation) but the problem is that the Ordinary Least Squares estimation technique is no longer appropriate for such models. The solutions to this problem are not treated in this document.

5.2.1 The Mean Model

The most naive forecasting model that could be considered is the so-called Mean Model where the time series under investigation Y_t is modeled by $Y_t = \alpha + e_t$ with a constant mean¹⁸ $\alpha \in \mathbb{R}$ for $t = 1, 2, \dots, T$. Obviously, this model will not perform well in the presence of trends because it attributes equal weights to past observations. Another factor to take into consideration is the choice of the constant α . The constant can be estimated in various ways (e.g. arithmetic

¹⁷Hint: if you compute a new regression model (with an additional dummy variable) you might have a look at the significance of the $\hat{\gamma}$ parameter.

¹⁸Actually, we can use any measure of central tendency - not just the arithmetic mean.

mean, median, midrange, ...) and they all have different (robustness) properties. A detailed, illustrated discussion of this naive model is available in another tutorial [20, document].

5.2.2 Single Moving Average

The Single Moving Average model is an extension of the Mean Model that attributes zero weights to observations of a distant past. Formally, we can define the Moving Average $M_t = \frac{Y_t + Y_{t-1} + Y_{t-2} + \dots + Y_{t-N+1}}{N}$ with $t = 1, 2, \dots, T$. The parameter N defines the number of the most recent observations that is thought to contain useful information about the time series' level.

Obviously, the extrapolation prediction must be based on observed information. Hence, $\hat{Y}_t = M_{t-1} = \frac{Y_{t-1} + Y_{t-2} + Y_{t-3} + \dots + Y_{t-N}}{N}$. The prediction on the long-run is simply $\hat{Y}_{t+h} = \hat{Y}_t$ for $h = 1, 2, \dots$

Assignment Show that the Single Moving Average is not an appropriate forecasting model by generating computations in a spreadsheet:

- Cut off the last 12 observations of the HPC time series.
- Compute the Single Moving Average with $N = 12$
- Compare the generated forecast with the true observations (that were cut off).

5.2.3 Centered Moving Average

The Centered Moving Average cannot be used for forecasting purposes because it uses past and future observations to obtain the smoothed values: $M_t = \frac{Y_{t+\frac{N-1}{2}} + Y_{t+\frac{N-1}{2}-1} + \dots + Y_t + \dots + Y_{t-\frac{N-1}{2}+1} + Y_{t-\frac{N-1}{2}}}{N}$ for any odd N and $t = 1, 2, \dots, T$.

Sometimes it is appropriate to choose a N value that is even. For instance, in the presence of monthly seasonality the choice $N = 12$ is interesting because it allows us to interpret M_t as the level of the time series (where seasonal fluctuations have been smoothed). Therefore we must consider the Centered Moving Average when N is even:

$$M_t = \frac{Y_{t+\frac{N-1}{2}-1} + Y_{t+\frac{N-1}{2}+1}}{2N} + \frac{Y_{t+\frac{N-1}{2}-2} + Y_{t+\frac{N-1}{2}}}{2N} + \dots + \frac{Y_{t-1} + Y_{t+1}}{2N} + \dots + \frac{Y_{t-\frac{N-1}{2}} + Y_{t-\frac{N-1}{2}+2}}{2N} + \frac{Y_{t-\frac{N-1}{2}-1} + Y_{t-\frac{N-1}{2}+1}}{2N}.$$

From the above it is clear that if the Moving Average method is used to decompose a time series with $N = s$ then the first and last $N/2$ observations are lost. Verify this in the archived computation [3, computation] (click to reproduce)!

5.2.4 Single Exponential Smoothing

Model This model uses a “smoothing constant” α and a recursive equation to generate a one-step-ahead prediction: $\hat{Y}_{t+1} = A_t$ where $A_t = \alpha Y_t + (1 - \alpha)A_{t-1}$ with $0 < \alpha \leq 1$. In other words the prediction is a weighted sum (c.q. “interpolation”) of the previous observation and the previous prediction: $\hat{Y}_{t+1} = \alpha Y_t + (1 - \alpha)\hat{Y}_t$. Another - more conventional - way to express the same model is: $\hat{Y}_{t+1} = \alpha Y_t + (1 - \alpha)(Y_t - e_t)$

$$\begin{aligned}\hat{Y}_{t+1} &= \alpha Y_t + Y_t - e_t - \alpha Y_t + \alpha e_t \\ \hat{Y}_{t+1} &= Y_t - (1 - \alpha) e_t\end{aligned}$$

which illustrates that the model learns from the previous observation and the previous prediction error¹⁹.

The recursive nature of this model is obvious and implies that the predictions are “exponentially” weighted moving averages with a discount factor $(1 - \alpha)$:

$$\begin{aligned}\hat{Y}_{t+h} &= \alpha Y_t + (1 - \alpha) A_{t-1} \\ \hat{Y}_{t+h} &= \alpha Y_t + (1 - \alpha) (\alpha Y_{t-1} + (1 - \alpha) A_{t-2}) \\ \hat{Y}_{t+h} &= \alpha Y_t + (1 - \alpha) (\alpha Y_{t-1} + (1 - \alpha) (\alpha Y_{t-2} + (1 - \alpha) A_{t-3})) \\ &\dots \\ \hat{Y}_{t+h} &= \alpha Y_t + (1 - \alpha) \alpha Y_{t-1} + (1 - \alpha)^2 \alpha Y_{t-2} + (1 - \alpha)^3 A_{t-3} \\ \hat{Y}_{t+h} &= \alpha Y_t + \sum_{i=1}^{\infty} (1 - \alpha)^i \alpha Y_{t-i} \\ \hat{Y}_{t+h} &= \alpha \sum_{i=0}^{\infty} (1 - \alpha)^i Y_{t-i}\end{aligned}$$

If the smoothing constant $\alpha = 1$ then the model reduces to the so-called “Random Walk” model $\hat{Y}_{t+h} = Y_t$ for $t = 1, 2, \dots, T$ and $h \in \mathbb{N}_0$.

Analysis The single exponential smoothing model was fitted to the HPC time series. The Exponential Smoothing R module [18, software] computes an estimate for $\hat{\alpha}$, interpolation forecasts $\hat{Y}_{t \leq T}$, residual diagnostics, and extrapolation forecasts $\hat{Y}_{t > T}$. The result [9, computation] (click to reproduce) shows that the smoothing constant is $\hat{\alpha} \simeq 0.28$ and the forecast $\hat{Y}_{T+h} \simeq 20228$ (million USD). The forecast on the long-run is a “flat line” (see Figure 19) because the model is (seemingly) not able to cope with trends. In addition, the model is - for obvious reasons - unable to predict seasonal fluctuations.

From the residual diagnostics (the residual ACF, spectrum, and cumulative periodogram) that are displayed in Figure 20 it is clearly seen that a strongly seasonal pattern is present in the prediction errors. This means that the model should be improved by adding the long-run trend and seasonal component to the forecasting model.

Interpretation The single exponential smoothing model has been criticized because it seems to be unable to adequately model trend-like behavior. This criticism is not justified - depending on how “trend-like behavior” is defined. To understand this let us consider the special case where $\alpha = 1$ which corresponds to the Random Walk model. In finance, the efficient market hypothesis (EMH) states that stock market prices behave like a Random Walk: all relevant information is correctly reflected in the last known stock prices. If the EMH is true then it implies that the best prediction that can be made about future prices is the current price: $Y_t = Y_{t-1} + e_t \Rightarrow \hat{Y}_t = Y_{t-1}$. If we carefully examine stock price time series then we observe that all of them exhibit long-run trends (which can be easily verified through decomposition or analysis of the ACF or spectrum). In other words, there exists an important economic theory that states that the stock market time series can only be modeled by (a special case of) the single exponential smoothing model even though they

¹⁹Sometimes this model is also called ARIMA(0,1,1) which is consistent with the work of Box & Jenkins where the more general class of ARIMA forecasting models is discussed.

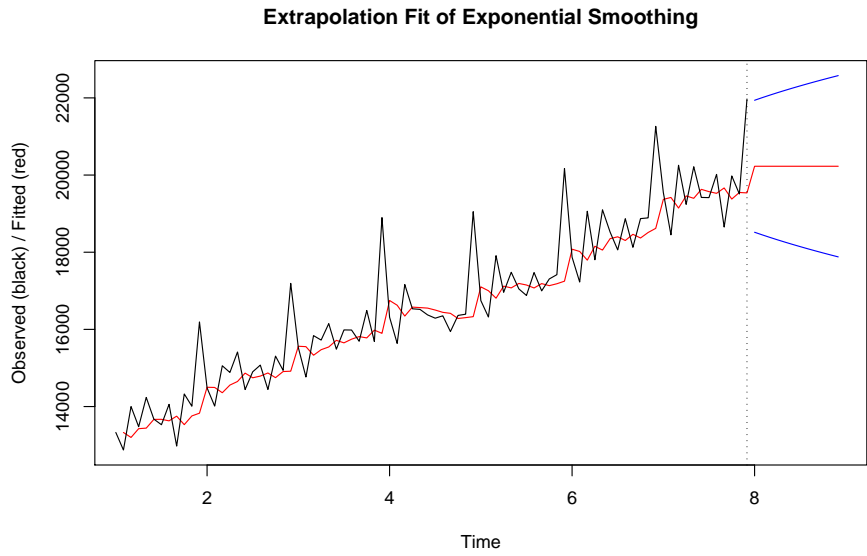


Figure 19: Extrapolation Forecast of Single Exponential Smoothing - US Retail Sales of Health and Personal Care
[link of reproducible computation](#)

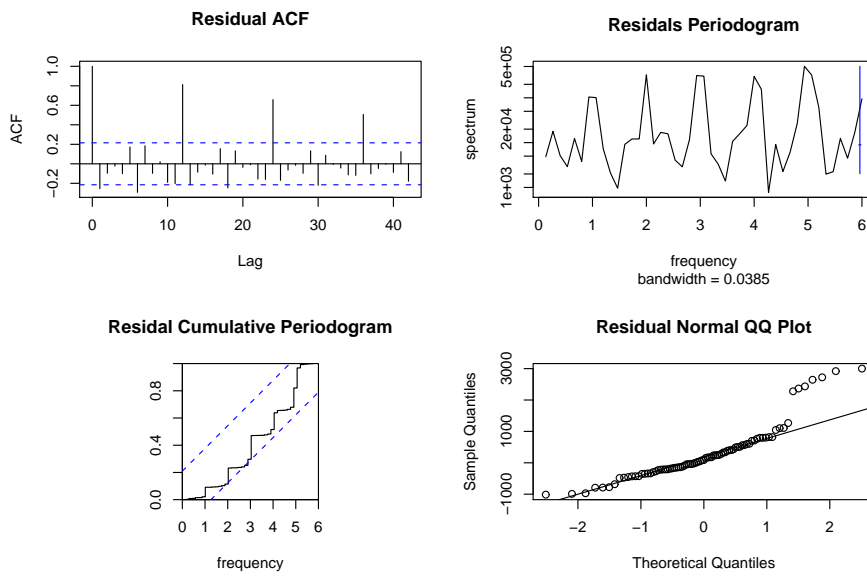


Figure 20: Residual Diagnostics of Single Exponential Smoothing - US Retail Sales of Health and Personal Care
[link of reproducible computation](#)

all contain long-run trends. Ultimately, this implies that some types of long-run trends cannot be predicted - in these cases a single exponential smoothing model might not be a bad choice. From the empirical evidence that what we have seen about the HPC time series however, it is obvious that the long-run trend is positive - hence, we expect the forecast to behave similarly. If we make an explicit distinction between the “level” of the time series and its incremental “trend” then we may fairly state that single exponential smoothing only models the “level” (and not the “trend”). If we treat both components as one then we need to make a distinction between trends that are predictable (like in retail sales) and trends that are unpredictable (like efficient stock markets).

Another remark about the trend in a single exponential smoothing model is related to the fact that it is possible to model a “deterministic” trend by simply adding a constant term to the equation (which is estimated together with the smoothing constant). The reason why this is the case and how this is interpreted, is beyond the scope of this document.

Conclusion The HPC time series should not be modeled by single exponential smoothing. An “incremental” trend and a seasonal component needs to be added (which requires a more sophisticated model).

Assignment Examine the archived computation and answer the following questions:

- Are the residuals normally distributed?
- Compare this model with the mean model. Which is better?²⁰

5.2.5 Double Exponential Smoothing

Model This model applies single exponential smoothing twice (with two different smoothing constants²¹ α and β) which leads to $\hat{Y}_{t+h} = A_t + hB_t$ where

$$\begin{cases} A_t = \alpha Y_t + (1 - \alpha)(A_{t-1} + B_{t-1}) \\ B_t = \beta(A_t - A_{t-1}) + (1 - \beta)B_{t-1} \end{cases} \quad \text{with } 0 < \alpha \leq 1 \text{ and } 0 < \beta \leq 1.$$

Analysis The double exponential smoothing model was fitted to the HPC time series. The Exponential Smoothing R module [18, software] computes an estimate for $\hat{\alpha}$, $\hat{\beta}$, interpolation forecasts $\hat{Y}_{t \leq T}$, residual diagnostics, and extrapolation forecasts $\hat{Y}_{t > T}$. The result is shown in [10, computation] (click to reproduce) and leads to much better (c.q. more realistic) predictions (see Figure 21) than with the previous model.

The problem with this model is that it still does not fit our HPC time series well. This is due to the fact that it doesn’t include a seasonal component which can be clearly seen in the residual ACF, spectrum and cumulative periodogram in Figure 22.

²⁰Hint: your answer can be based on empirical evidence and a theoretical argument.

²¹Note that in some textbooks this model is defined with only one parameter - in other words: $\alpha = \beta$. We don’t want to impose this restriction on the parameters.

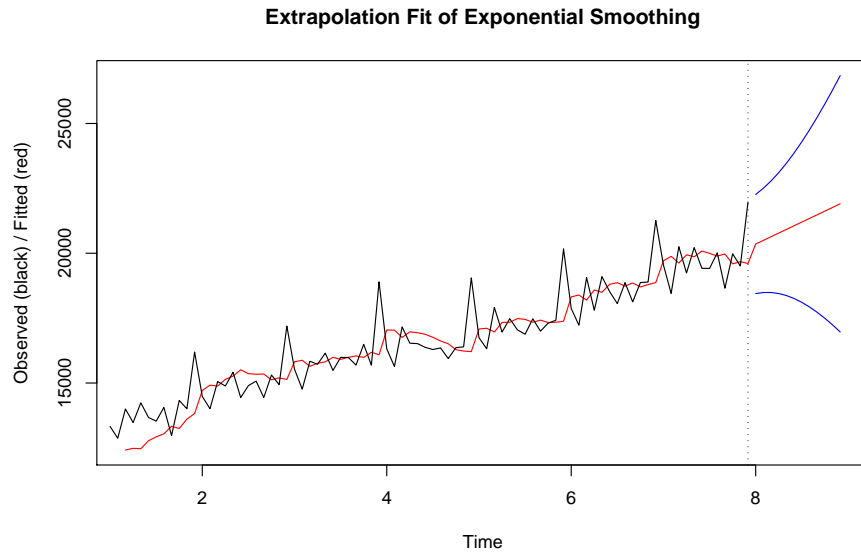


Figure 21: Extrapolation Forecast of Double Exponential Smoothing - US Retail Sales of Health and Personal Care
[link of reproducible computation](#)

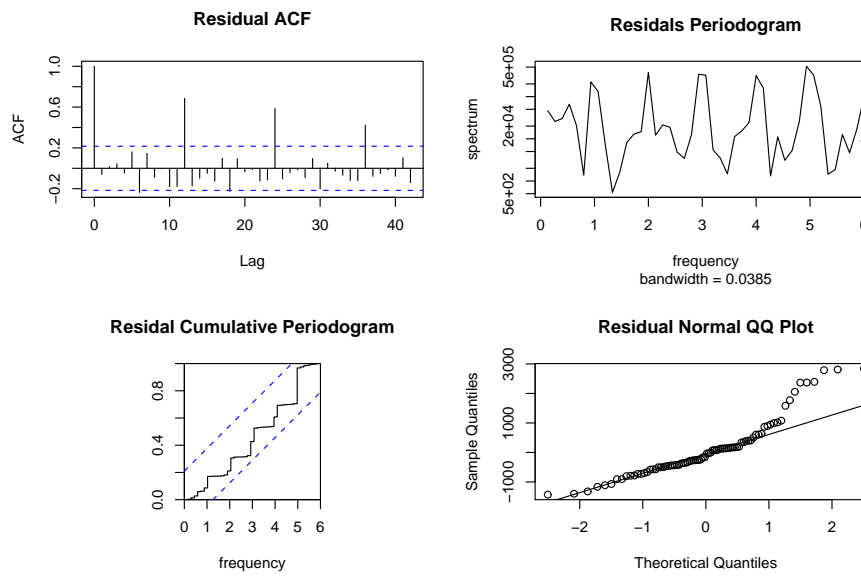


Figure 22: Residual Diagnostics of Double Exponential Smoothing - US Retail Sales of Health and Personal Care
[link of reproducible computation](#)

Assignment The residuals of the double exponential smoothing model are clearly not normally distributed. Answer the following questions (based on Figures 21 and 22):

- Is the distribution of the residuals skewed to the right or the left?
- Explain why the residuals are not normally distributed (and how you could solve the problem).

5.2.6 Triple Exponential Smoothing (Holt-Winters model)

Model The Holt-Winters model includes a level, an (incremental) trend, and a seasonal component. For each component there is a parameter (α , β , and γ) that characterizes the underlying dynamics of the time series. The model comes in two flavors, depending on the relationship between the level and seasonality²².

The additive Holt-Winters model is defined by the following relationships:

$$\hat{Y}_{t+h} = A_t + hB_t + S_{t+h-s}$$

$$\text{where } \begin{cases} A_t = \alpha(Y_t - S_{t-s}) + (1 - \alpha)(A_{t-1} + B_{t-1}) \\ B_t = \beta(A_t - A_{t-1}) + (1 - \beta)B_{t-1} \\ S_t = \gamma(Y_t - A_t) + (1 - \gamma)S_{t-s} \end{cases}$$

The multiplicative Holt-Winters model is defined by the following relationships:

$$\hat{Y}_{t+h} = (A_t + hB_t) S_{t+h-s}$$

$$\text{where } \begin{cases} A_t = \alpha(Y_t/S_{t-s}) + (1 - \alpha)(A_{t-1} + B_{t-1}) \\ B_t = \beta(A_t - A_{t-1}) + (1 - \beta)B_{t-1} \\ S_t = \gamma(Y_t/A_t) + (1 - \gamma)S_{t-s} \end{cases}$$

Single exponential smoothing ($\beta = \gamma = 0$) and double exponential smoothing ($\gamma = 0$) are both special cases of the additive Holt-Winters model.

Analysis The triple exponential smoothing model was fitted to the HPC time series by the use of the Exponential Smoothing R module [18, software]. The result [11, computation] (click to reproduce) shows much improved extrapolation forecasts (see Figure 23) and residual diagnostics (see Figure 24).

In the ACF, spectrum and cumulative periodogram there is no evidence of residual seasonality which implies that the model adequately computes seasonal effects. In addition, it can be concluded from the Normal QQ plot that the residuals are normally distributed. The model is not perfect because the residual ACF still exhibits a non-random (regular) autocorrelation pattern $\hat{\rho}_{k \in \{3,6,9,\dots\}} > 0$ and $\hat{\rho}_{k \in \{1,4,7,\dots\}} < 0$.

Assignment Examine the multiplicative Holt-Winters model [12, computation] (click to reproduce) and compare the results with the additive model. Which model is better?

²²Again, the difference between the additive and multiplicative models is related to the Heteroskedasticity problem - see section 4.1

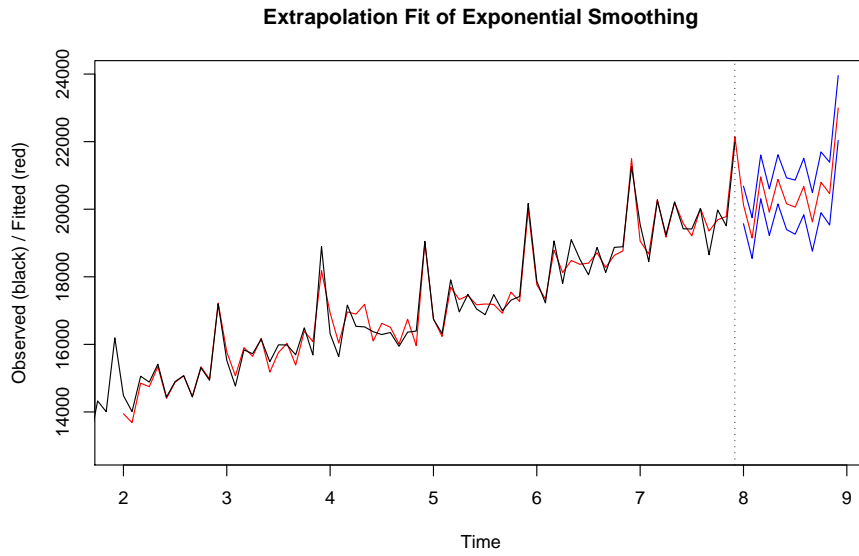


Figure 23: Extrapolation Forecast of Additive Holt-Winters Smoothing - US Retail Sales of Health and Personal Care
[link of reproducible computation](#)

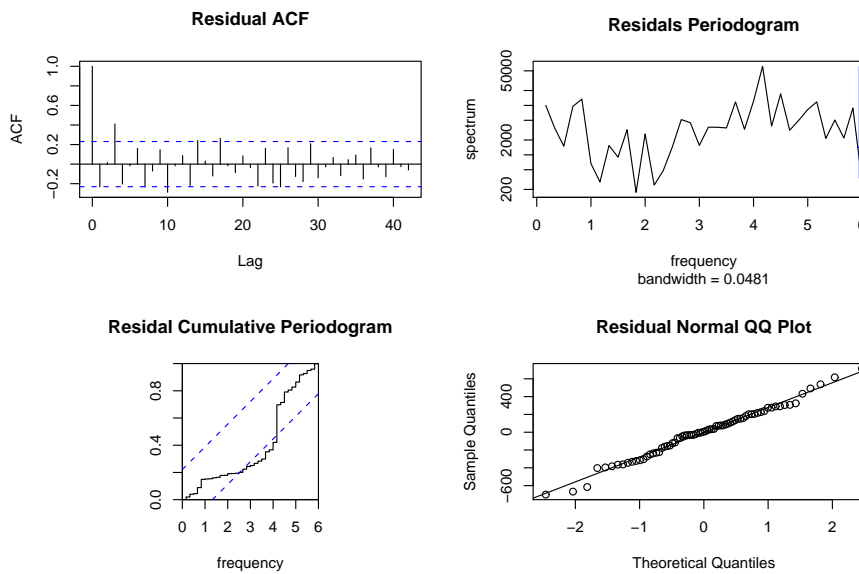


Figure 24: Residual Diagnostics of Additive Holt-Winters Smoothing - US Retail Sales of Health and Personal Care
[link of reproducible computation](#)

References

- [1] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- [2] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL <http://www.freestatistics.org/blog/date/2008/Mar/02/t1204472634kobk5s74i81tzo2.htm>, Retrieved Sun, 02 Mar 2008 16:44:08 +0100
- [3] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL <http://www.freestatistics.org/blog/date/2008/Mar/02/t1204475598g12zn9oc3bgyk55.htm>, Retrieved Sun, 02 Mar 2008 17:33:28 +0100
- [4] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL <http://www.freestatistics.org/blog/date/2008/Mar/06/t12048034000a6t92ft1rhvoic.htm>, Retrieved Thu, 06 Mar 2008 12:36:43 +0100
- [5] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL <http://www.freestatistics.org/blog/date/2008/Mar/06/t1204806748irjzv611b0ldfyh.htm>, Retrieved Thu, 06 Mar 2008 13:32:39 +0100
- [6] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL <http://www.freestatistics.org/blog/date/2008/Mar/08/t1204976128fhkkc79lopfua8.htm>, Retrieved Sat, 08 Mar 2008 12:35:48 +0100
- [7] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL <http://www.freestatistics.org/blog/date/2007/Oct/18/mmymb1ceu780zm21192701323.htm>, Retrieved Thu, 06 Mar 2008 13:50:59 +0100
- [8] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL <http://www.freestatistics.org/blog/date/2008/Mar/08/t12049839076t3dgtgn3c3py8w.htm>, Retrieved Sat, 08 Mar 2008 14:45:26 +0100
- [9] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL <http://www.freestatistics.org/blog/date/2008/Mar/10/t1205171047i3xo7gva8f85ux9.htm>, Retrieved Mon, 10 Mar 2008 18:44:15 +0100
- [10] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL <http://www.freestatistics.org/blog/date/2008/Mar/10/t1205171502777xpto22vqrg3p.htm>, Retrieved Mon, 10 Mar 2008 18:52:07 +0100
- [11] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL

<http://www.freeststatistics.org/blog/date/2008/Mar/10/t1205171842e0ry9c7xankjola.htm>,
Retrieved Mon, 10 Mar 2008 18:57:28 +0100

- [12] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL <http://www.freeststatistics.org/blog/date/2008/Mar/10/t120517200847h2f9y4t7wnhq9.htm>, Retrieved Mon, 10 Mar 2008 19:00:12 +0100
- [13] <http://spreadsheets.google.com/ccc?key=pV35do1d8bhGwAYPDmHv1CQ&hl=en>
- [14] Wessa P., (2008), Classical Decomposition (v1.0.1) in Free Statistics Software (v1.1.22-r4), Office for Research Development and Education, URL http://www.wessa.net/rwasp_decompose.wasp/
- [15] Wessa P., (2008), Decomposition by Loess (v1.0.0) in Free Statistics Software (v1.1.22-r4), Office for Research Development and Education, URL http://www.wessa.net/rwasp_decomposeloess.wasp/
- [16] Wessa, P. (2008), Free Statistics Software, Office for Research Development and Education, version 1.1.22-r4, URL <http://www.wessa.net/>
- [17] Wessa P., (2008), Multiple Regression (v1.0.25) in Free Statistics Software (v1.1.22-r4), Office for Research Development and Education, URL http://www.wessa.net/rwasp_multipleregression.wasp/
- [18] Wessa P., (2008), Exponential Smoothing (v1.0.2) in Free Statistics Software (v1.1.22-r4), Office for Research Development and Education, URL http://www.wessa.net/rwasp_exponentialsMOOTHING.wasp
- [19] Wessa P. (2008), A framework for statistical software development, maintenance, and publishing within an open-access business model, Computational Statistics 2008, (click to open)
The original publication is available at www.springerlink.com (DOI 10.1007/s00180-008-0107-y)
- [20] Wessa P. (2008), A Compendium of Reproducible Research about Descriptive Statistics and Linear Regression, URL <http://www.wessa.net/download/tutorial.pdf>
- [21] R. B. Cleveland, W. S. Cleveland, J.E. McRae, and I. Terpenning (1990) STL: A Seasonal-Trend Decomposition Procedure Based on Loess. Journal of Official Statistics, 6, 373, (click to open)
The original publication is available at www.jos.nu
- [22] Duke University, Spreadsheet implementation of seasonal adjustment and exponential smoothing, URL <http://www.duke.edu/rnau/411outbd.htm>