University of St.Gallen

Time Series Economics Assignment II

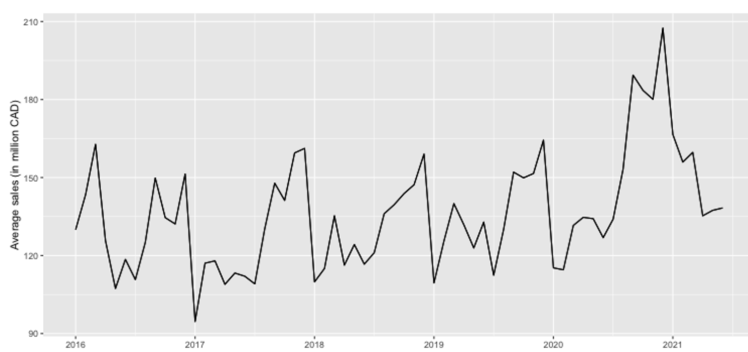
Fabio Luigi Eugenio Bernasconi [redacted]
Maria Kyprouli [redacted]
Lauritz Storch [redacted]

May 8, 2022

1. Forecasting Chocolate Sales

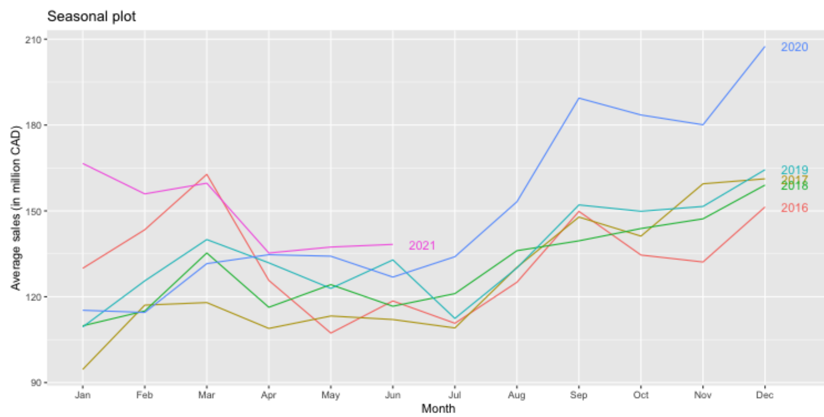
a) **The file “MonthlyChocolateSales2021.xlsx” contains informationon average monthly manufacturer sales of chocolate and chocolate confectionery in Canada from 2016 to 2021 (in million CAD). Load the file into R and turn the sales data into a time series object using the “tseries” library. Plot the data and comment. Can you make sense of increases and decreases in chocolate sales? Does the series appear stationary?**

The plot shows the information on the average monthly manufacturer sales of chocolate and chocolate confectionery in Canada from 2016 to 2021 (in million CAD). The data is fluctuating approximately around the values 100 and 160 from the years 2016 to 2020. The amplitude of the fluctuations increases overall in the time span 2020-2021 and spikes at approximately 210. One can observe a cyclical pattern for each year with troughs at the beginning and some small fluctuations until the mid of a year. In the second half of each year, one can observe an increase in sales with a peak in December. Sales sharply decrease in-between years. The peaks in December are probably caused due to Christmas. Apart from that, a close look at the data indicates that increases around other events like Easter, Halloween, and Thanksgiving also might influence chocolate sales in March, October, and November. Taking 2020 out, chocolate sales appear to be seasonal and not stationary. From 2020 onwards, one can observe an overall increment in chocolate sales and larger fluctuations around the previously mentioned events. This change in pattern might be COVID-related.

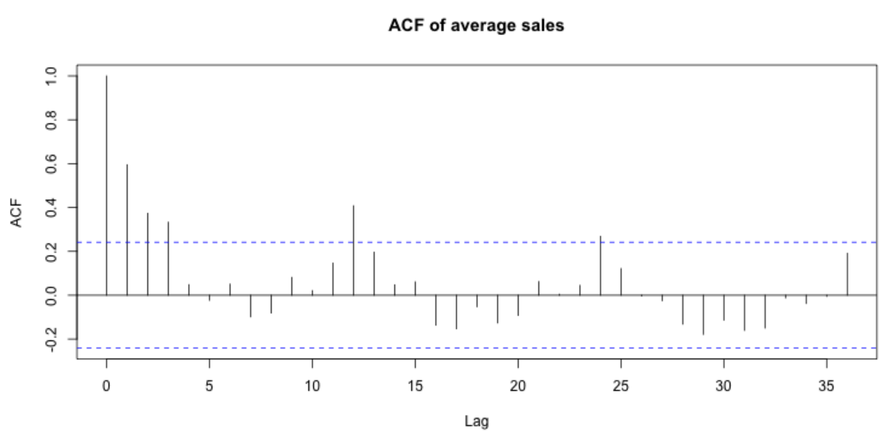


b) **Create a seasonal plot of the data (you can e.g. use the function “ggseasonplot”, choose the frequency appropriately) and plot the ACF of the time series for a number of $h = 36$ lags. Comment.**

Looking at the plots, one can observe that we have a similar seasonality pattern for each year. Taking 2020 out, sales increase from July onwards in an almost identical shape until December and then sharply decrease in January followed by an increment until March. From March until July, the plots move randomly around an average value. Most importantly, kinks appear yearly in March, September, and December. Although we observe differences in fluctuation size and overall increments in chocolate sales during 2020 and 2021, the general pattern also holds for both years and, hence, visually proofs our conjecture of a seasonal, non-stationary process.



Furthermore, the ACF plot shows significant spikes in the lags 1, 2, 3, 12, and 24. Hence, the seasonality that we observed previously is not significant with exception of the significant spikes in the 12th and 24th lag which indicate a strong correlation with the same month in the previous two years. Furthermore, the ACF indicates that neither an MA process nor an AR process is a valid candidate for representing seasonality in the data.



c) **Fit an ARIMA model using the "auto.arima" command of the "forecast" library. Check the model diagnostics and comment.**

As previously discussed, the seasonal plot and ACF plot indicate that we have non-stationarity. Therefore, an ARIMA(p,d,q) model is used to obtain stationarity for an ARMA(p,q) process by differencing d-times. In our case, we use an

$$\text{ARIMA} \quad \underbrace{(0,1,1)}_{\text{Non-seasonal part}} \; \underbrace{(0,1,1)_{m=12}}_{\text{Seasonal part}}$$

where $m = 12$ is the 12th-month observation. Since our model is an ARIMA$(0,1,1)(0,1,1)_{m=12}$ the model includes the seasonal spikes in the 12th month. The model coefficients are calculated in R and are given as follows

	coeff
ma1	-0.28^{*}
	(0.14)
sma1	-0.74^{**}
	(0.28)
AIC	413.61
AICc	414.10
BIC	419.52
Log Likelihood	-203.81
Num. obs.	53

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: $ARIMA(0,1,1)(0,1,1)_{12}$ model coefficients

Moreover, after plotting for the residuals diagnostics from ARIMA$(0,1,1)$, we can argue that lags of the error terms are uncorrelated (not reject the null hypothesis for Ljung-Box test: p-value 0.7006), meaning that residuals are white noise.

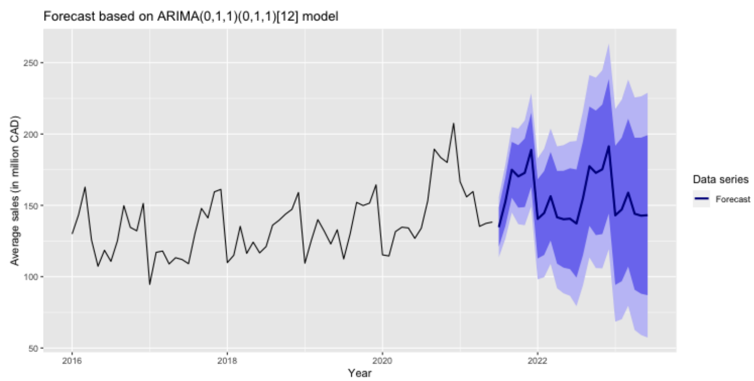
	Test	Results
1	Ljung-Box test:	$Q^* = 8.1416$, df = 11, p-value = 0.7006

Table 2: Ljung-Box test for residuals of ARIMA(0,1,1)(0,1,1)$_{m=12}$ model

d) **Calculate $h = 24$ step ahead forecasts for the time series using the function "forecast" of the library "forecast". Does your forecast live up to your expectations?**

We also have to keep in mind that the auto.arima model gives us a $SARIMA(0, 1, 1)$ model by default. If we switch the settings such that seasonality is zero (which would be wrong in the case of our data), we get an ARIMA(1,1,1) with flat predictions which would be clearly not sufficient in the sense of predicting our future.

The pattern remains the same as expected if we include seasonality in our forecasts. We can observe the same pattern, which remains cyclical as previously described in a), and it has a similar fluctuation size as in the years 2016 to 2020. A closer look shows that the forecast is a forward iteration and, therefore, is equal in both forecasted years. The error term forecast is slightly increasing over time.

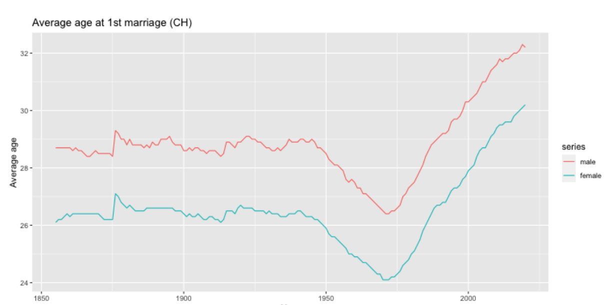


2. Census Data Analysis

a) **The file "CHMarriageAges.xlsx" contains information on the average age of women and men at first marriage in Switzerland from 1855 to 2020. Load the file into R and turn the data into a time series object using the "tseries" library. Plot the data and comment. Can you make sense of the development of the average marriage ages? Do the series appear stationary? How would you describe their joint behaviour?**

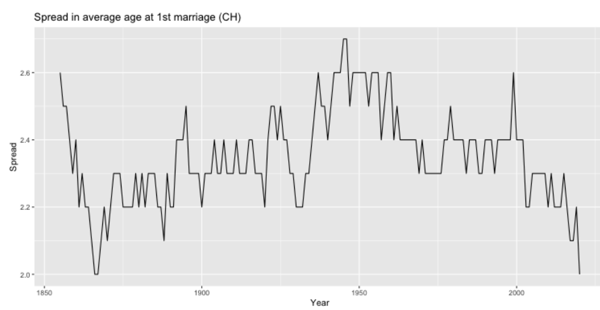
The data provided are the average age of women and men at first marriage in Switzerland from 1855 to 2020. A first look at the data plot shows that the average age remained almost constant from 1855 until 1945 for both genders. With the end of World War II in 1945, the average age decreases by approximately 2 years until 1975 for both genders. A potential reason for this could be that soldiers returned after the end of the war and thus marriage happened at earlier ages at this point. According to the report of the Swiss statistical authority, the marriage rate rose particularly significantly among the younger age groups; the proportion of women already married at 25 years of age rose from 35% in 1930 to nearly 60% in 1970 (Swiss Federal Statistical Office, 1998). After 1975, one can observe an upturn in the average age. The trend reversed more quickly than it developed with the sharp increase in the average age of first marriage, in two decades. The trend toward earlier marriages was wiped out at the end of the 19th century. This persistent rise could be due to the structural development of modern societies, the extension of studies, and the later entering the labor market of both males and females.

Furthermore, one can observe that the graphs of both genders move almost identically but females marry on average two years younger than males.

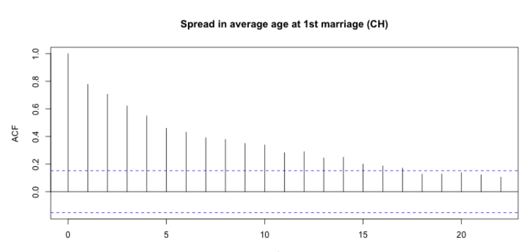


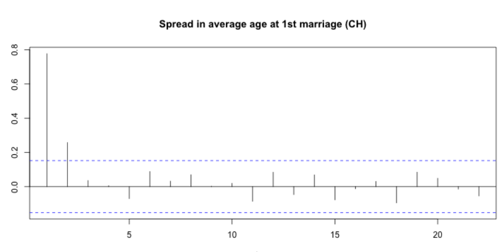
b) **Calculate the spread of the two times series (average male marriage age average female marriage age), plot it and calculate ACF and PACF. Comment. Name an ARMA and an ARIMA model that could suit the data according to the plots. You may consider taking differences in the spread and checking the (P)ACF. You can check your results with the function "auto.arima" from the forecast library.**

Looking at the spread plot of males minus females, one can observe that the data is fluctuating within 2.0 and 2.7 years. There are no mentionable conspicuous outliers. The plot peaks around 1950 and hits its trough around 1875 and 2020 simultaneously. The spread seems to follow a random walk which can be indicated by the fact that the next value is not far off the previous value. Hence, it could be a unit root process.

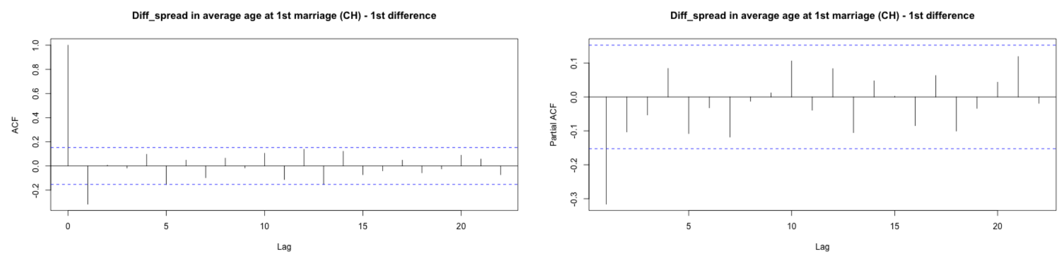


From plotting ACF and PACF for the spread of the two time series, we can observe that the ACF plot demonstrates a gradual decreasing pattern, while the PACF plot appears a spike on lag 1. Thus, we could argue that we can have an AR process with a large coefficient which again confirms our suspicion of a unit root. Furthermore, the alternating sign of the PACF indicates some positive MA(1) component in the process. Indeed, according to auto.arima, the most appropriate process is an ARIMA$(0, 1, 1)$ model. In other words, differencing one time gives an MA(1) process.





After calculating the differences in the spread, we see that only the ACF plot has a spike at lag 1. The only common significant spike appears at lag 0 for both plots. The first difference is a stationary process as we would expect in the case of a unit root process. According to auto.arima, the most appropriate process is an ARIMA$(0,0,1)$.



Looking at the table, we observe identical results for the ARIMA$(0,1,1)$ model of the spread and the ARIMA$(0,0,1)$ model of the spread difference which makes sense since the ARIMA model of the spread is performing the same difference calculation then we did before running the auto.arima function on $diff\_spread$.

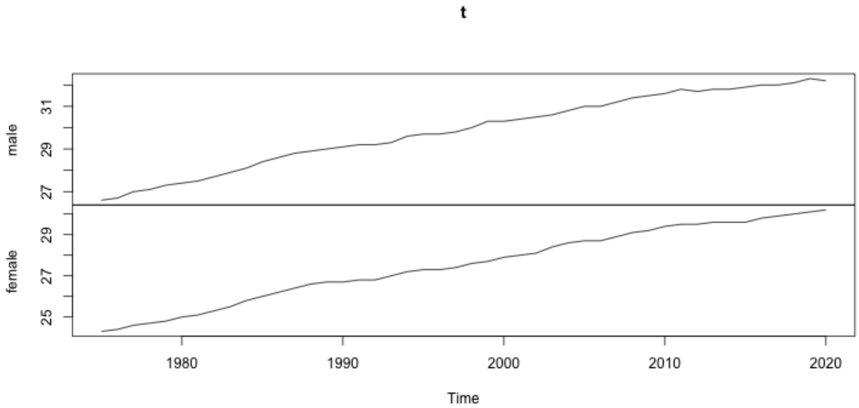
	ARIMA$_{spread}(0,1,1)$	ARIMA$_{spread_diff}(0,0,1)$
ma1	-0.36^{***}	-0.36^{***}
	(0.07)	(0.07)
AIC	-354.57	-354.57
AICc	-354.49	-354.49
BIC	-348.36	-348.36
Log Likelihood	179.28	179.28
Num. obs.	165	165

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: The ARIMA$(0,1,1)$ model for variable *spread*, ARIMA$(0,0,1)$ model for variable $spread\_diff$

c) **Focus now on the time series from 1975 onwards (to keep structural breaks in the data from distorting the results). Conduct an Augmented Dickey Fuller test on the two time series (male and female) using the R library "urca". Explain which set-up you choose in the testing procedure (lag length and regression setting) using economic and econometric arguments. Comment on the results. Are both time series stationary? If both of them are non-stationary, do they have the same order of integration?**

A first look at the time series plot for both genders indicates that our data is trending. The graph is almost linearly increasing over time and both graphs move almost identically.



The question on hand is whether we indeed have a trend component or a unit root with drift term? Therefore, we have the following equation

$$\Delta y_t = \alpha_0 + \underbrace{(\phi - 1)}_{\gamma} y_{t-1} + \alpha_2 t + \epsilon_t$$

where we test the null $H_0 : \alpha_2 = \gamma = 0$. If we cannot reject the null hypothesis, our process has a unit root with a drift component. If we reject the null hypothesis, then we can conclude that y_t is stationary around a deterministic time trend. To test for unit roots we implement the *augmented Dickey-Fuller test*. To conduct this test, we use the regression setting "*trend*" to include the intercept and time trend in our regression and check for the significance of the latter one. Adding lags affects the significance of the test negatively. Therefore, we start by setting our lags = 12 which is our maximum, and then during a second step ADF test consider only the lags that were significant in the first run. We observe that we do not reject our null hypothesis e.g. $\phi_{3,male} = 1.2389 < \phi_{3,critical}$ and $\phi_{3,female} = 1.5627 < \phi_{3,critical}$.

Value of test-statistic:	0.655	1.0854	1.2389
Critical values	1 pct	5 pct	10 pct
τ_3	-4.15	-3.50	-3.18
ϕ_2	7.02	5.13	4.31
ϕ_3	9.31	6.73	5.61

Table 4: Augmented Dickey Fuller test with 12 lags for males

Value of test-statistic:	-0.7267	3.3466	1.5627
Critical values	1 pct	5 pct	10 pct
τ_3	-4.15	-3.50	-3.18
ϕ_2	7.02	5.13	4.31
ϕ_3	9.31	6.73	5.61

Table 5: Augmented Dickey Fuller test with 12 lags for females

In our second step, we only consider the lags that were significant according to the first ADF test (0 for both series). Even if we take the minimum number of significant lags, we still do not observe rejection e.g. $\phi_{3,male} = 3.7504 < \phi_{3,critical}$ and $\phi_{3,female} = 2.3668 < \phi_{3,critical}$.

Value of test-statistic:	-0.8762	31.0255	3.7504
Critical values	1 pct	5 pct	10 pct
τ_3	-4.15	-3.50	-3.18
ϕ_2	7.02	5.13	4.31
ϕ_3	9.31	6.73	5.61

Table 6: Augmented Dickey Fuller test with 0 lags for males

Value of test-statistic:	-1.1196	45.1683	2.3668
Critical values	1 pct	5 pct	10 pct
τ_3	-4.15	-3.50	-3.18
ϕ_2	7.02	5.13	4.31
ϕ_3	9.31	6.73	5.61

Table 7: Augmented Dickey Fuller test with 0 lags for females

Concluding from the results of our augmented Dickey-Fuller test, we obtain that we do not reject our null hypothesis. Therefore, both processes have a unit root with a drift term. The processes are non-stationary and standard regression theory does not apply to our processes.

d) **What kind of econometric relationship do the two (restricted) series adhere to?**

In c) we have already confirmed that both processes have a unit root and hence are integrated of the same order (order 1). Now, to check whether the econometric relationship that the two series adhere to is one of cointegration (of order 1), we perform the Engle-Granger 2-step approach and regress the average age of males (in 1st marriage) on the average age of females (in 1st marriage).

If our noise terms also include a drift term then we have a striking result because inferences are invalid if they are built on such processes due to the spurious regressions problem. On the other side, if both processes are integrated in the same order then our residual sequence is stationary and we could obtain valid inferences.

Table 8

	Dependent variable:
	male
female	0.960***
	(0.007)
constant	3.422***
	(0.197)
Observations	46
R^2	0.998
Adjusted R^2	0.998
Residual Std. Error	0.086 (df = 44)
F Statistic	18,032.660*** (df = 1; 44)
Note:	*p < 0.1; **p < 0.05; ***p < 0.01

Table 9: Regression of male on female

We then use an augmented Dickey-Fuller test and set regression to "*none*" in R to include the intercept and exclude a time trend in our regression. We set our lags according to the Bayes information criterion. Our results indicate that we can reject $H_0 : \gamma \neq 0$ on a 1% significance level in the regression model

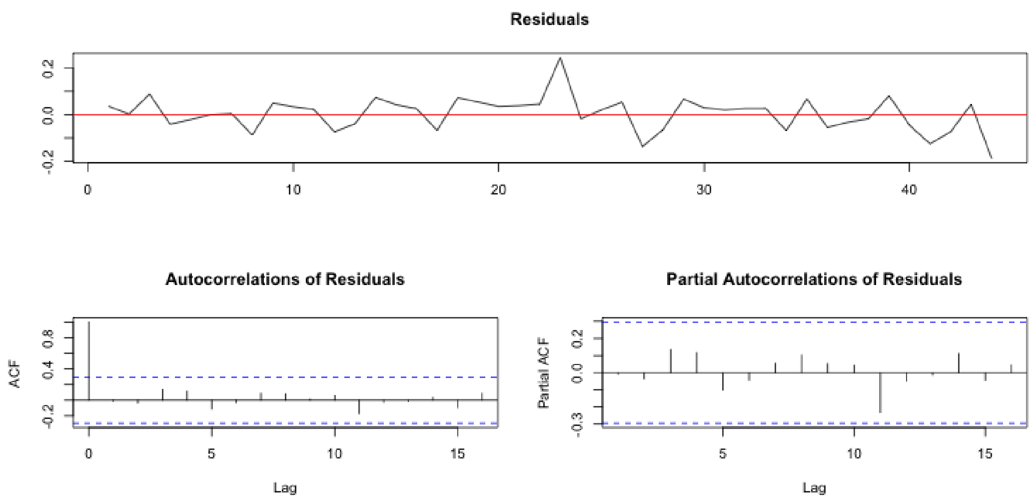
$$\Delta y_t = \gamma y_{t-1} + \epsilon_t$$

because $\tau_{1,\Delta y_t} = -2.9745 < \tau_{1,critical}$. Hence, Δy_t is a Random Walk without drift e.g. $\mathbb{E}[\Delta y_{t-1}] = 0$ and both processes are cointegrated of order one.

Value of test-statistic:		-2.9745	
Critical values	1 pct	5 pct	10 pct
τ_1	-2.62	-1.95	-1.61

Table 10: Augmented Dickey Fuller test on regression residuals

A visual check also confirms our findings.



References

Swiss Federal Statistical Office (1998), *Two centuries of Swiss demographic history, Graphic album of the 1860–2050 period*, ≪Swiss Statistics≫ series.