Final Project — Predicting the Price of Chocolate Chip Cookies over Time

Nicklaus Kim

Fall 2022

1 Introduction

One of the prominent and most widely studied areas for the application of time series analysis is the field of economics. Much research and investment has been committed into curating powerful and accurate analyses of time-dependent mechanisms such as supply and demand of a good/commodity, (un)employment and labor, and perhaps most directly, sales and prices. In this paper, we will look specifically at food price data, which currently seems apt as a basis for a time series analysis given the unpredictable state of prices of commodities and the economy as a whole today. One good source for such data is the Federal Reserve Economic Data (FRED) website, where we are able to find a great number of historical price data for all sorts of goods, from retail to other raw crops.[1] In particular, we will look at the average price of chocolate chip cookies (per pound) in U.S. cities. The choice of this particular item for our analysis was mostly due to the fact that it is a fun, lighthearted food item that seemed somewhat appropriate for this time of year (with the Christmas season and with it, Christmas cookies, fast approaching).

2 The Data

Our dataset consists of monthly recordings of the price per pound of chocolate chip cookies from 1980 to 2022 and can be readily visualized with a timeplot of the series (Figure 1). We can immediately notice a couple things from this plot: (1) there exists a relatively strong upward trend in the data — that is, the price has been steadily increasing over time; (2) there may potentially be some underlying seasonality since there appears to be quite a bit of fluctuation within individual years. These observations are especially evident if we look at the decomposition of the series (Figure 2), where we see quite clearly the trend and seasonality components that we must eventually separate from the "random

[1] These data are collected from the United States Bureau of Labor Statistics and similar sources.

noise" part of the data in order to conduct any useful analysis. Namely, we must have a *stationary* series x_t in order to do any model fitting or forecasting later.

We may partially be able to explain the two aforementioned patterns using some domain knowledge natural intuition we may have about the national economy and consumer tendencies. Most notably, the linear rise in price may be attributed to the fact that the American economy has consistently undergone staunch inflation during the time period covered by the data. We may ponder whether this inflationary behavior dominates the alternate possibility that the price of chocolate chip cookies *in particular* has been rising organically, whether due to some supply-and-demand reaction or a change in the quality or perception or "luxuriousness" of the product. (This would be something to investigate further by studying the prices of other, similar foods during the same period.) As for the seasonal trend, we may postulate that, much as is the case for many foods in the national economy, the price of cookies fluctuates from month to month in a cyclical or predictable way; they may be more or less expensive in some months compared to others, as a general, repeating pattern. For instance, as we alluded to earlier, perhaps it is possible that chocolate chip cookies are a somewhat festive or seasonal food item and so are more expensive during the winter months.

In addition, by looking at the autocorrelation (ACF) plot (Figure 3) and partial autocorrelation (PACF) plot (Figure 4), we can confirm that we certainly need to alter our data before applying a model. Currently, the data are being dominated by the trend and seasonality components, which does not make for a suitable or useful analysis and eventual forecasting. We will tackle each of these key ideas about transforming the data in turn, and in the end, fit an appropriate autoregressive moving average, or ARMA(p, q), model we can then use for making predictions on previously unseen data. To meet this end, our first course of action is to take the data and partition it into a training set, which we will use for the fitting and selection of models, and a test set, on which we will eventually make predictions. We remove the last several years of observations, from January 2019 to October 2022, and designate this as our test set; naturally, we keep the rest of the data, from January 1980 to December 2018, for training.

3 Regression

We can begin by first attempting to remove the clear (linear) trend currently present in the data. The relationship looks somewhat non-linear, but it also seems reasonable to just use a regular linear regression fit. Once we have found the regression line of best fit, we can transform our original data by subtracting this line from it. That is, we think of our original time series x_t as being composed of two parts, the linear trend component and (ideally) a random error or noise process:

$$x_t = \beta_0 + \beta_1 z_t + w_t.$$

It then just remains for us to estimate those regression coefficients β_0, β_1 for our data, which we accomplish by simply using the lm() functon in R, ending up with a relatively good fit to the data (Figure 5).

This method of strictly removing the *linear* trend has the added benefit of having better interpretability of the detrending of the data as opposed to, say, a quadratic or cubic fit. It is more straightforward to think of subtracting a line from the original data than say, a quadratic or cubic function. However, while we do not include the alternate analysis here, it may still be worthwhile to explore the use of a spline interpolation or similar method to fit the data points more closely.

4 Spectral Analysis

Next, we must investigate the fluctuations of the price particularly *within* individual years. That is, we would like to remove any strong seasonality present in the data. We begin by producing some preliminary (smoothed) periodograms (Figure 6) to look at the spectrum; we perform both a nonparametric and a parametric (autoregressive) fit. In both cases, there is really no indication of a seasonal pattern in the data, which is potentially discouraging at the onset.

At this point, however, we may call upon some of our intuition/knowledge of the subject matter and postulate the existence of a seasonal trend after all. As mentioned earlier, we have a suspicion that for most foods in general, the price will fluctuate from month to month or season to season, depending on different crop yields, goods transports, general consumer demand, etc. In particular, we may believe that the price of cookies is higher in the months leading up to the winter holidays, hinting at an annual cycle of some sort. In fact, with a bit of exploratory data analysis, we can see exactly this sort of relationship by comparing the timeplots for two separate years in our dataset (Figures 7, 8) and find that indeed, many of the months share a common pattern/directional change (increase/decrease). Therefore, we can attack the seasonality by way of a method similar to that of removing the linear trend; for each month, we subtract the overall mean price for that month as a way of effectively removing the overall seasonality for the data at large. Finally, we end up with a detrended, seasonality-removed dataset (Figure 9) that is ready to be fed into our predictive models.

5 Model Fitting

Now that we have modified our data appropriately, we are ready to fit some models in order to eventually produce some predictions on the testing data we set aside earlier. Our weapon of choice is the autoregressive moving average (ARMA) model, with model parameters p and q which denote the order of the AR and MA processes, respectively. In mathematical terms, this will be of the

form

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q},$$

where x_t is stationary, $\phi_p \neq 0$, $\theta_q \neq 0$, and $\sigma_w^2 > 0$.

In order to get an all-encompassing look at which model(s) may be successful, we choose to do a comprehensive "grid search" of the possible model parameters for $p = 0, \ldots, 9, q = 0, \ldots, 9$.

So, we fit all 100 of the possible $ARMA(p, q)$ models and calculate the Akaike Information Criterion (AIC) for each then look for the optimal model(s) by finding the model(s) with the lowest AIC. In the end, we end up seeing that the best model according to this metric is an $ARMA(7, 3)$, with an AIC of -2.212534. We note that when performing the 100 model grid search, some of the model fittings for some of the order combinations did not converge (as reference, the sarima() function in R was used in this step and specifically, led to an optim() error). However, of the ones that work with no errors, the $ARMA(7, 3)$ is indeed our model of choice.

We can next check the diagnostics of the $ARMA(7, 3)$ fit. We conduct a residual analysis of the fit (Figure 10) and see that the various tests and assumptions are being satisfied. The standardized residuals look to be adequately centered at zero with constant variance and follow a normal distribution. Therefore, we can posit that our model is valid and can thus be used for prediction in the next phase of our analysis.

6 Forecasting

We are now equipped to do some forecasting using our ARMA model. Using the $ARMA(7, 3)$ coefficients, we calculate the predicted values on the training set data and see that the predictions line up quite closely with the true values (Figure 11). In fact, we can also calculate the root mean-squared error (RMSE) of approximately \$0.09. Considering that the data values at hand — that is, the prices of the cookies — are within the range of roughly 1-4 dollars, this is a relatively good margin of error to achieve. This is a good indication that our model is sufficiently generating accurate predictions and can henceforth be applied to the previously unseen testing data (the 2019-present time period).

Next, we can apply the same steps to predict on the test set. We must first transform the testing data in the same way as we did for the training set. We can take the original values (of the testing set portion) of the time series, x_t, and subtract both the trend and seasonal cyclical aspect that were removed in the lead-up to the model fitting. Then, we simply use our model to generate the predictions, in the same say as we saw for the training data, and evaluate how close these are to the real, ground-truth price values. We can see again that visually, the predictions line up quite nicely with the true data (Figure 12). In addition, we again calculate the RMSE on the test data; we find it to be roughly 0.13. This test performance seems to indicate that it is safe to conclude that our chosen model is doing quite a good job at predicting the future price values of chocolate chip cookies.

7 Conclusion and Future Work

We wrap up our analysis by reviewing our general findings and proposing some other future directions one may pursue to get a further understanding of this dataset. We saw that most of the non-stationarity underlying the data is explained by the strong linear trend we removed. To reiterate an earlier point, it is possible that there exists a better method for fitting the curvilinear trend (perhaps a cubic spline or something similar), however for our purposes here a straight line fit performed well enough. After this was removed, the seasonal aspects, while not present when inspecting the periodogram, were influential as well. We were able to find that the data followed a moderate annual cycle, where certain months (Dec/Jan/Feb) saw increases in cookie price, while others (Jun/Jul) saw clear decreases. Once we subtracted the month-by-month means as a way of handling the underlying seasonal cycle, we were able to smoothly fit an $ARMA(7, 3)$ model to the final data. This model, in turn, produced quite accurate predictions of future prices for chocolate chip cookies.

Another future direction for an analysis of this data may be to difference the data. Since we are dealing with price data, differencing may offer an alternative view of the data and its fluctuations through time. Rather than the raw price from month to month, we can interpret the *change* in price across months, similar to how one might use the notion of returns when analyzing stock prices.

One conceptual question that may still remain is that of the origin of the seasonal aspect. That is, what might be the reason for the price being higher or lower in certain months than others? One idea for investigating this may be to conduct similar time-series analysis of other goods/foods and see whether any parallel or similar patterns emerge. In any event, the conclusions we may draw are limited to the scope of this analysis of chocolate chip cookies alone. What we can say for certain, while we may not know or understand fully the economic driving forces underlying, is that we have found a way to mathematically represent the steady upward linear trend and annually cyclical nature of the data and create a predictive ARMA model that can be reliably be used to forecast for the price of chocolate chip cookies in future months and years.

8 Appendix

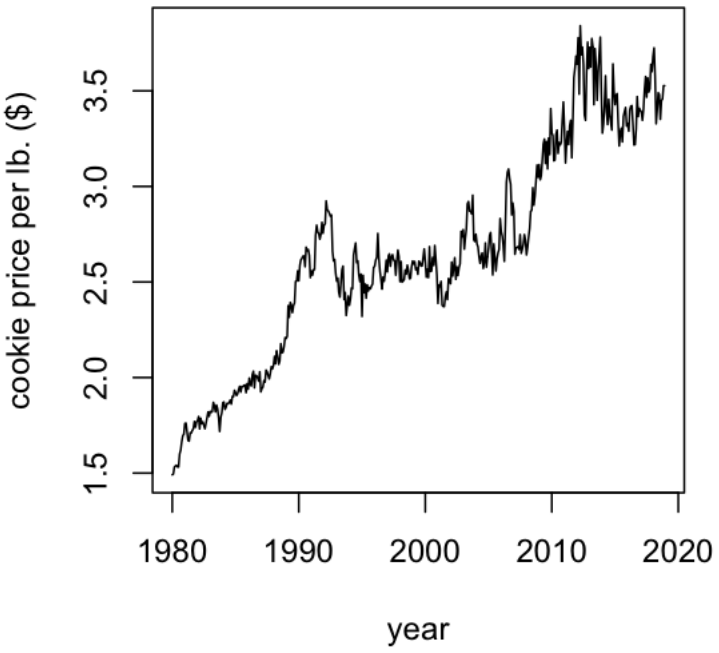

Figure 1: Timeplot of the (training) data

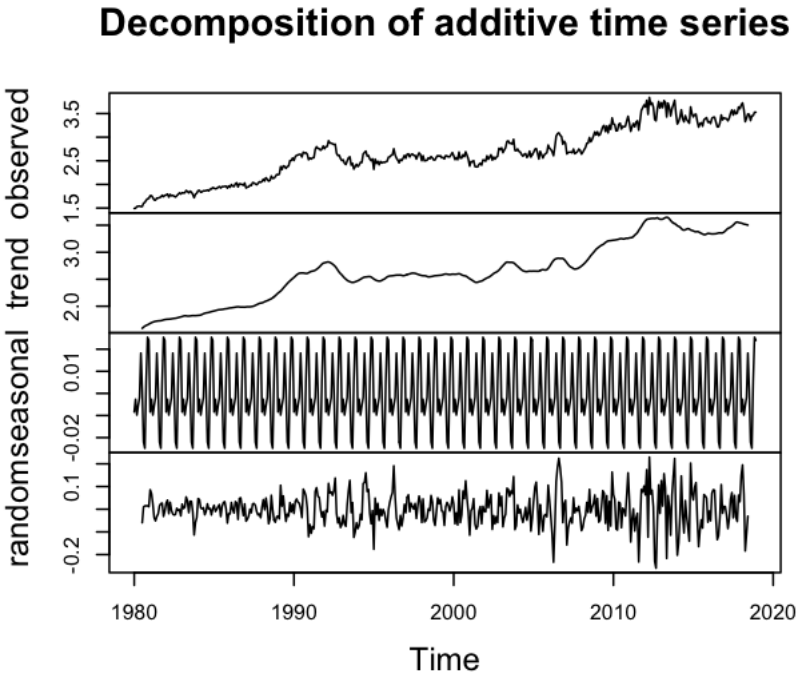

Figure 2: Plot of series decomposition

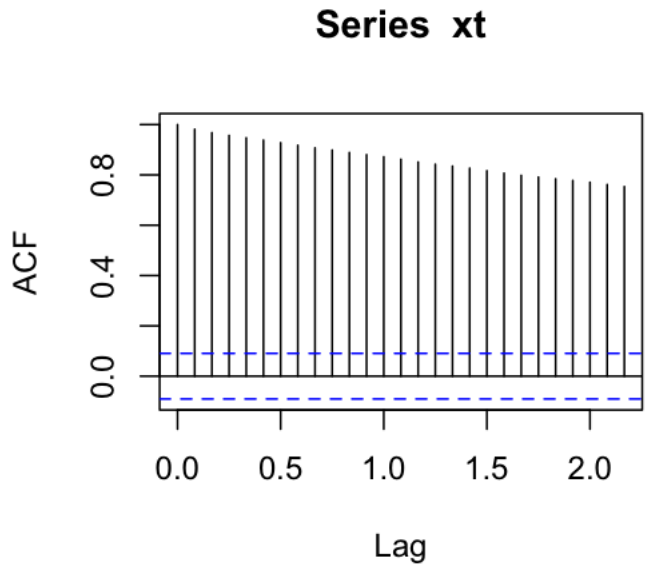

Figure 3: Autocorrelation (ACF) plot

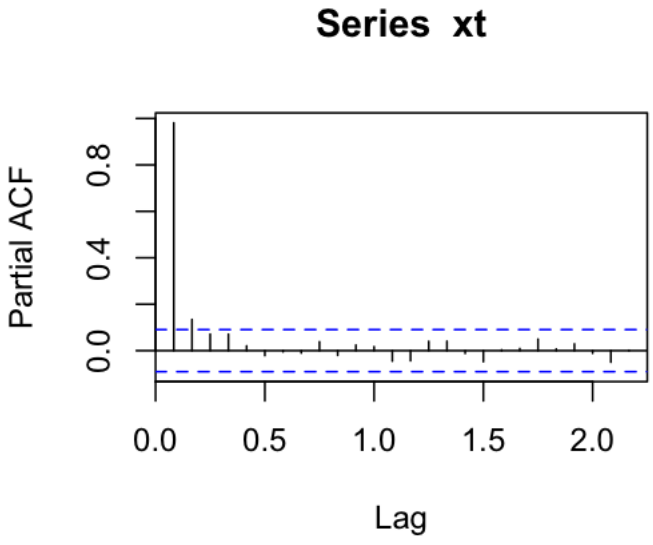

Figure 4: Partial autocorrelation (PACF) plot

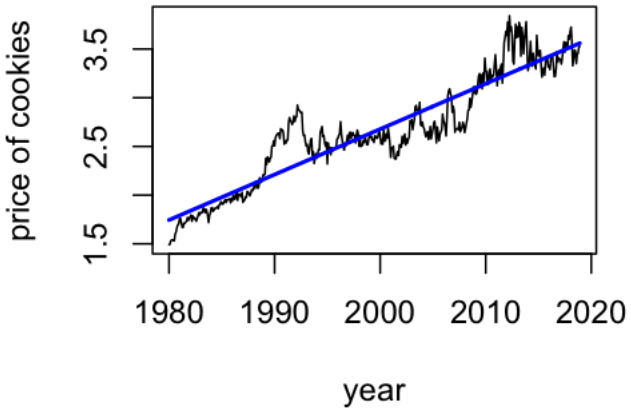

Figure 5: Data with regression line of best fit

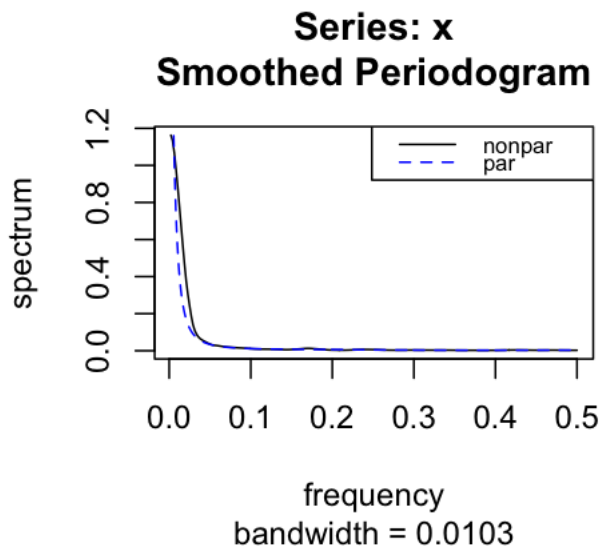

Figure 6: Periodogram for series

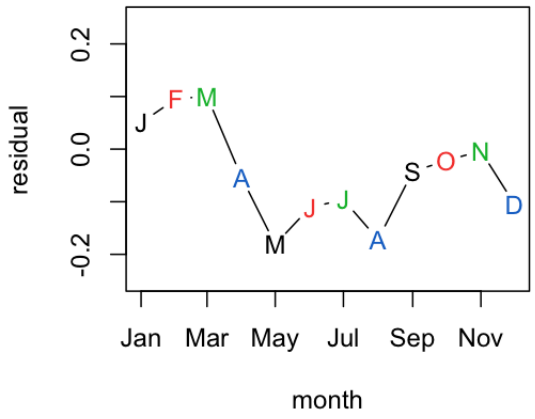

Figure 7: Prices by month in 2015

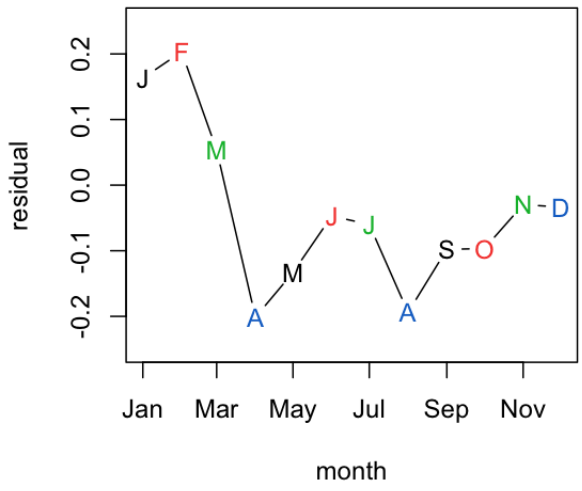

Figure 8: Prices by month in 2018

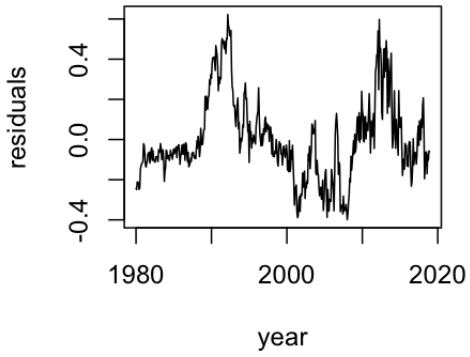

Figure 9: Series with trend + seasonality removed

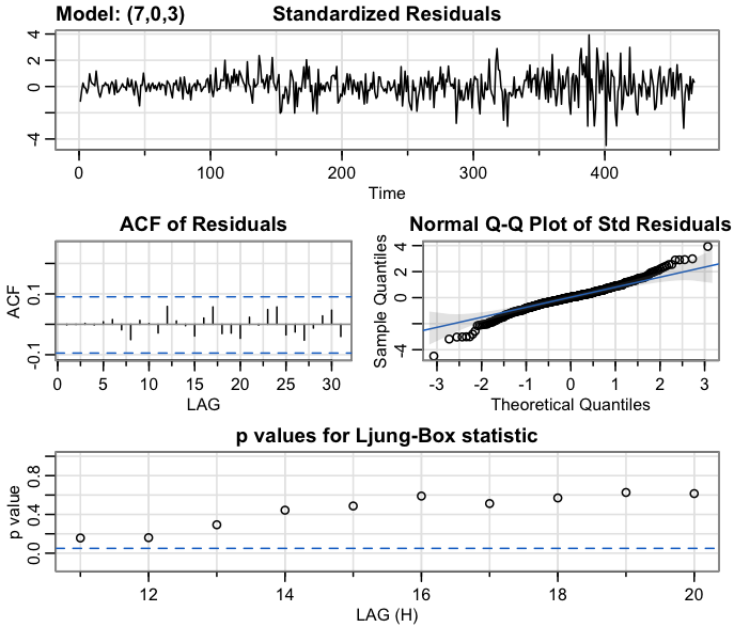

Figure 10: Diagnostic plots for ARMA(7, 3) model fit

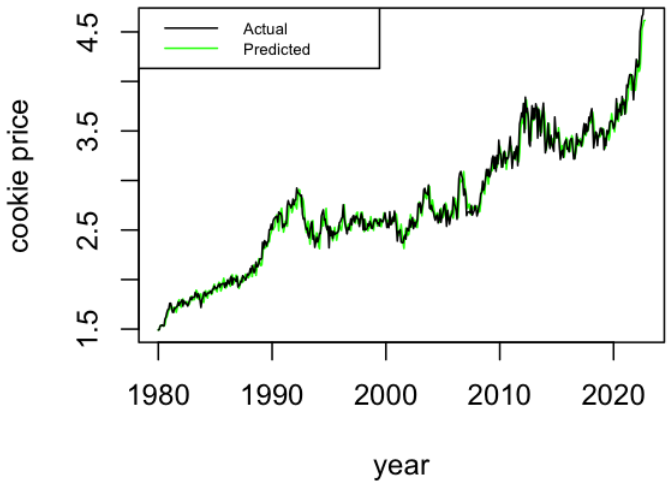

Figure 11: Predicted and actual prices for training data

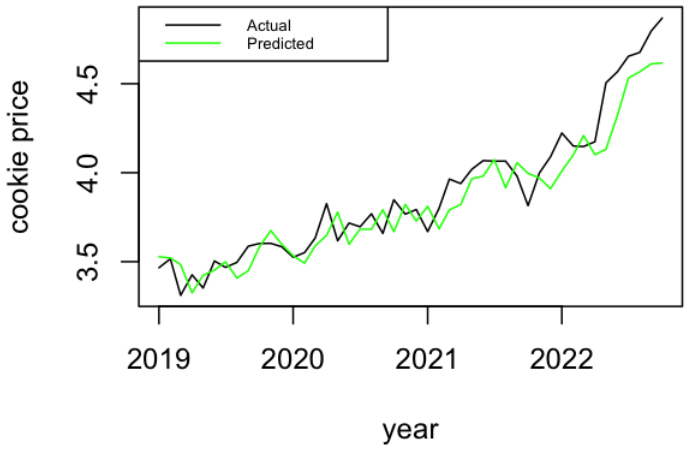

Figure 12: Predicted and actual prices for testing data

9 References

FRED Economic Data. *Average Price: Cookies, Chocolate Chip (Cost per Pound/453.6 Grams) in U.S. City Average*. https://fred.stlouisfed.org/series/APU0000702421

BLS Beta Labs. *BLS Data Viewer*. https://beta.bls.gov/dataViewer/view/timeseries/APU0000702421