# Time Series Forecasting

## Correlation of Variables

In order to give us a benchmark that we could compare to the other models used, we initially created a time-series forecasting model, specifically an ARMA model.

Before beginning to create our time model, we compared the correlation between the weekly sales and the other numerical variables present in the dataset. Unfortunately, there was very little correlation between the variables and the sales, with the highest correlation present between weekly sales and size. However even this was only a correlation of 24% and as nothing else was above 5% it was instead decided to create a time-series of just the date and weekly sales and to take an autocorrelation of that.

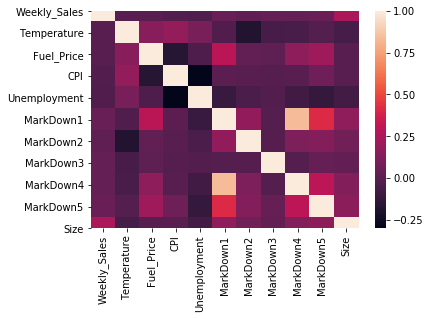


Fig x. Heatmap of Correlation between the numerical variables

## Autocorrelation and testing assumptions

To start with we transformed the data from a dataframe to a time-series by changing the index from a simple numerical one to a datetime index, based on the first day of each week’s sales. After taking the correlation of this series, compared to itself shifted by one week, the correlation was a much closer -42%. Taking this new series then we created an autocorrelation plot over a period of 60 lags order to highlight any statistically significant values that could be used when construction an ARMA model.

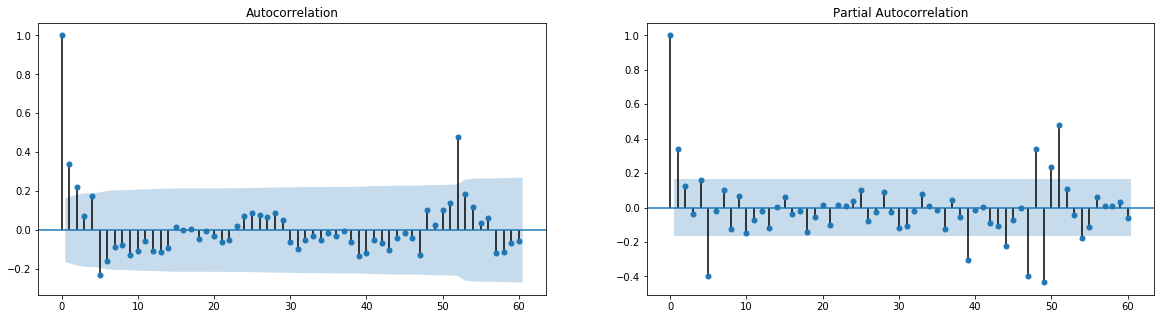


Fig x. ACF and PACF plot of the time series

From the graph it was apparent that there were significant values present at 1, 2, 5 and 52 lags, with these 4 values following outside of the confidence interval, demonstrating there’s a more than 95% chance that we can reject the null hypothesis that these values are due to random chance, and that is almost certain that the true autocorrelation is not equal to 0.

Next an Augmented Dickey-Fuller test was carried out on the time series. With the ADF test, the "null hypothesis" (the hypothesis that we either reject or fail to reject) is that the series follows a random walk. If this was true than it would make it much more difficult to find a suitable model to explain or predict any changes in the data. Therefore, if the test returns a low p-value (less than 5% as standard) it means we can reject the null hypothesis that the series is a random walk.

The test was carried out on the weekly sales time series and returned an extremely low p-value of 2.68e-7 which tells us that the data is stationary and therefore able to be successfully modelled using an ARMA model.

## AR, MA and Final ARMA Model

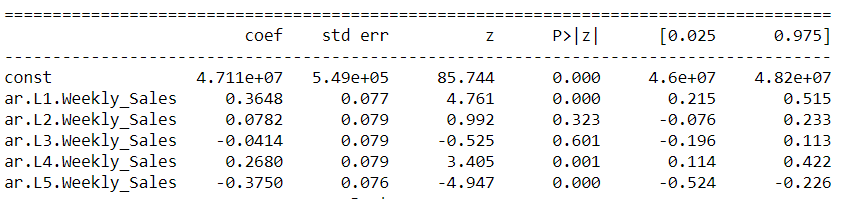
The next part of the process was to construct both an AR and MA model for the process. To decide what order to use for each model, we iterated through the orders 1 to 10 for the AR model and the orders 1 to 8 for the MA model in order to find the order with the lowest information criteria (Akaike's Information Criterion and Bayesian Information Criterion) as the lower this value the better the model fits the data.

|  |  |
| --- | --- |
|  |  |

Fig x. Chart of Information Criteria against the Model Order for both model types

From a plot of the AIC/BIC vs the model order it was apparent that an order of 5 was the best fit for each model- this matches with our earlier autocorrelation chart which highlighted 5 as a significant variable.

Creating an AR model of order 5 produced a model with the following coefficients:



This means the final model for an AR(5) model would look like so:

Confidence Intervals

Or:

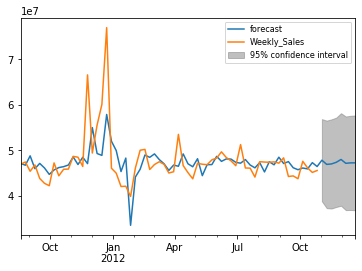
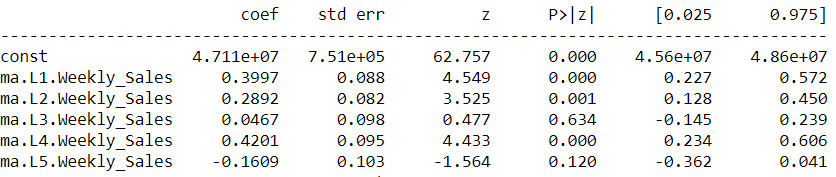


Fig x. AR(5) model

The Autoregressive model seems to roughly follow the actual data trend, only with a slight lag and a much smaller amount of variance compared to the actual dataset. However, compared to an MA model there is no smoothing out of consecutive peaks and troughs, and it is capable of predicting much further into the future than an MA model, with a confidence interval of ± 1\*107 Weekly Sales units.

Creating an MA model of order 5 produced a model with the following coefficients:



This means the final model for an MA(5) model would look like so:

Or:

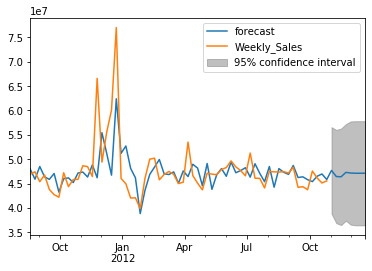
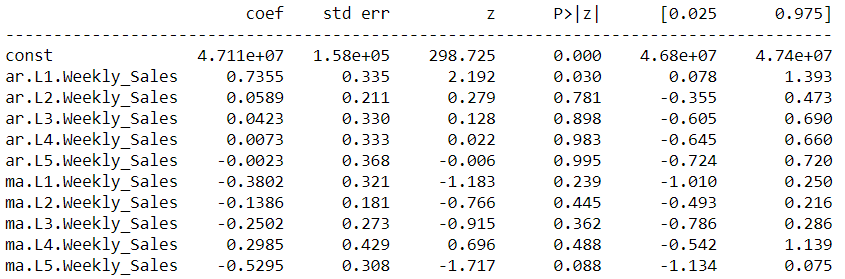


Fig x. MA(5) model

Compared to the AR model this comes much closer to hit the magnitude of the peaks and troughs in the actual data. However, when a peak and trough of similar magnitude are too close to one another they get averaged out and replaced with a smoother line. Additionally, a Moving Average model is only capable of predicting values two intervals ahead of the final datapoint, before it decays into a straight line of the overall weekly sales average.

Finally by combining the two models to create an ARMA(5,5) model, this allows us to remove some of the lag from the AR model whilst also reducing the smoothing present in the MA model, at the expense of less predictive power than the AR model and smaller peaks and troughs than the MA model. It gives us a model with the following coefficients:



This means the final model for an ARMA(5,5) model would look like so:

Or:

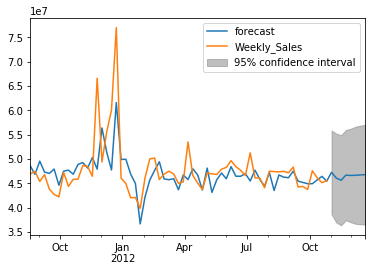


Fig x. MA(5,5) model

This gives the following 10 predictions for the last 5 weeks of real data (based on the preceding 138 weeks) and 5 weeks afterwards (all sales values rounded to 3 S.F.):

|  |  |  |  |
| --- | --- | --- | --- |
| Week Start | Actual Sales | Predicted Sales | % Difference |
| 2012-09-28 | 43734899 (4.37\*107) | 4.49\*107 | 2.6 |
| 2012-10-05 | 47566639 (4.76\*107) | 4.49\*107 | -0.60 |
| 2012-10-12 | 46128514 (4.61\*107) | 4.57\*107 | -0.88 |
| 2012-10-19 | 45122411 (4.51\*107) | 4.64\*107 | 2.8 |
| 2012-10-26 | 45544116 (4.55\*107) | 4.54\*107 | -0.22 |
| 2012-11-02 | - | 4.73\*107 | - |
| 2012-11-09 | - | 4.61\*107 | - |
| 2012-11-16 | - | 4.56\*107 | - |
| 2012-11-23 | - | 4.67\*107 | - |
| 2012-11-30 | - | 4.66\*107 | - |