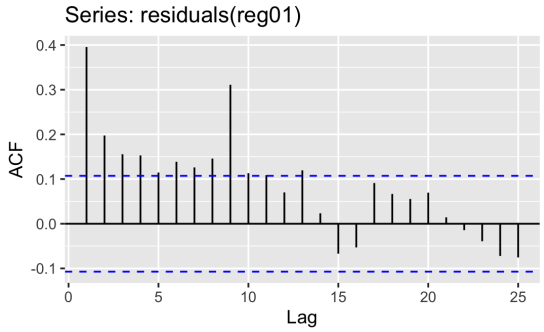
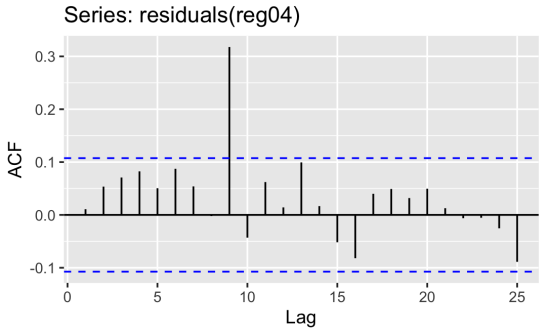
Bill Henderson
CSCI-E-116
Spring 2025

Assignment 05

The models chosen for analysis were all looking to predict the sales dependent variable and multiple variations on the four explanatory variables. Some generated variables for more information were made as well, these were day of the week as an integer (dow), day of the month as an integer (Dom) and month as an integer for more explanatory variables. There was also a stepwise reduction regression done, but its components matched reg02, so

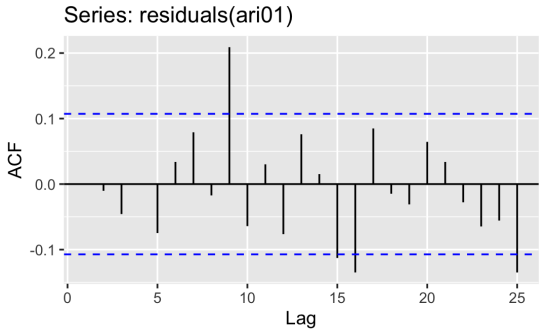
First looking at residuals, the worst of the linear regressions was reg01 which included the branded and non branded searches, all of the generated time variables and the TV advertising variable. I also limited the training set to dates after 10 October 2016 since the TV advertising variable was zero for dates before this. Below is the graphic for the Autocorrelation Function and we can see almost all lags are positive and some significant seasonality in the chart.





The best residuals were on reg04, which had both branded and non-branded searches, the day of the week, month, and a 1 day lagged version of the sales results. Theres still some seasonality in the metrics as well as a significant spike around the 9th lag and most values being positive, but nearly all lags are within the significance lines. Reg03 and 05 were very similar, 03 added the TV Advertisement variable and 05 added a 7 day lagged TV advertising parameter in thoughts that maybe the TV advertising had a delayed response. Since 04 was the simplest of the 3, it seemed best.

The ARIMA model (right) did very well with the autocorrelation function comparatively. There is some increasing as lags increase, but there are lags both positive and negative, and the vast majority are within the significance boundaries.



metrics	reg01	reg02	reg03	reg04	reg05	reg06	reg07	step
r.sq	0.760	0.759	0.793	0.791	0.789	0.785	0.707	0.759
adj.r.sq	0.751	0.752	0.786	0.785	0.782	0.777	0.699	0.752
std err	326.089	325.741	303.207	303.738	306.370	309.291	359.258	325.741

I also gathered the R^2, adjusted R^2 and Standard Error stats for all the linear regressions. Again regressions 03 through 05 have the strongest response with high R^2 values for both

regular and adjusted metrics as well as the lowest standard errors. The ARIMA model summary doesn't show an R^2 value, but it does have an RSME which was around 303, very close to our top linear regressions.

Finally we had our predictions and their measure for RMSE. The ARIMA model performed the best against all the other linear regressions. The next closest linear regression in terms of RMSE was reg01 which was essentially our baseline model, including all explanatory variables, none of which had any lag components.

metrics	reg01	reg02	reg03	reg04	reg05	reg06	reg07	reg.arima	reg.step
rmse	399.514	395.444	432.963	414.776	502.4	741.529	611.836	339.58	395.444