Walmart Sales- Analysis and Forecasting

Sandip Palit, Vyas Swaminathan

Abstract

- **Objective 1:** Determine the factors which have an effect on sales

    1. Is there a relationship between temperature and sales?
    2. Is there a relationship between CPI and sales?
    3. Is there a relationship between fuel price and sales?
    4. Is there a relationship between unemployment and sales

- **Objective 2:** Make forecasts for projected sales for that store and department

Variables Analyzed

- Weekly Sales
- Temperature
- Fuel Price
- CPI
- Unemployment Rate

Data Preparation

- ‘Train.csv’ - historical sales data for 45 Walmart stores from different regions
- ‘Features.csv’ - historical data of various features relevant to sales
- Analysis was done on a merged dataset having the following columns:
    1. Store#, Dept#, Date, Weekly Sales, Temperature, Fuel Price, CPI, Unemployment Rate.
- Store # 4 and Department # 12 were randomly chosen for the purpose of this analysis.
- Weekly Sales, Temperature, Fuel Price, CPI, Unemployment Rate were converted into Time Series.
- Data was split into in-sample (90%) and out of sample (10%) datasets

Exploratory Data Analysis

Examined each Time Series for Stationarity

- Analysis of raw time series data, ACF , PACF plots and ADF tests all show that none of the time series satisfy the condition of stationarity needed for further analysis

Differenced non-stationary time-series until stationarity is achieved

- Differenced once to make stationary : Weekly Sales, Fuel Price, Temperature, Unemployment
- Differenced twice to make stationary : CPI

Analyzed Cross Correlations between variables

- Fuel Price and CPI do not have statistically significant cross-correlations with weekly sales
- Unemployment and Sales have very small correlations of interest with weekly sales
- Temperature and Sales have cross-correlation in first 2-3 lags

Analyzed co-integrations between variables

- Temperature and sales appear to be cointegrated. The ideal next step would have been to test these variables for Granger causality and fit a VECM model. We have made a note of this detail and proceeded to fit a standard VAR model

Estimation

1. Model1: Sales + Temperature

   - VARSelect recommended a VAR model of order 5 based on lowest AIC

   - Since there wasn't a significant difference between VAR(1) and VAR(5) in terms of AIC, we chose the VAR(1) model in the interest of parsimony.

2. Model2: Sales + Temperature + Unemployment

   - Since there wasn't a significant difference between VAR(1) and VAR(6) in terms of AIC, we chose the VAR(1) model in the interest of parsimony.

3. Model3: Sales + Temperature + Unemployment + Fuel Price

   - VARSelect recommended VAR (1).

4. Model4: Sales + Temperature + Unemployment + Fuel Price + CPI

   - VARSelect recommended VAR (1).

Comparison between Various Models

Model	Time Series Included	Adjusted R2	Out of Sample RMSE
Model 1	Sales + Temperature	0.18	1055.7
Model 2	Sales + Temperature + Unemployment	0.20	1058.5
Model 3	Sales + Temperature + Unemployment + Fuel Price	0.19	1059.6
Model 4	Sales + Temperature + Unemployment + Fuel Price + CPI	0.19	1059.2

Model Diagnostics - Residuals

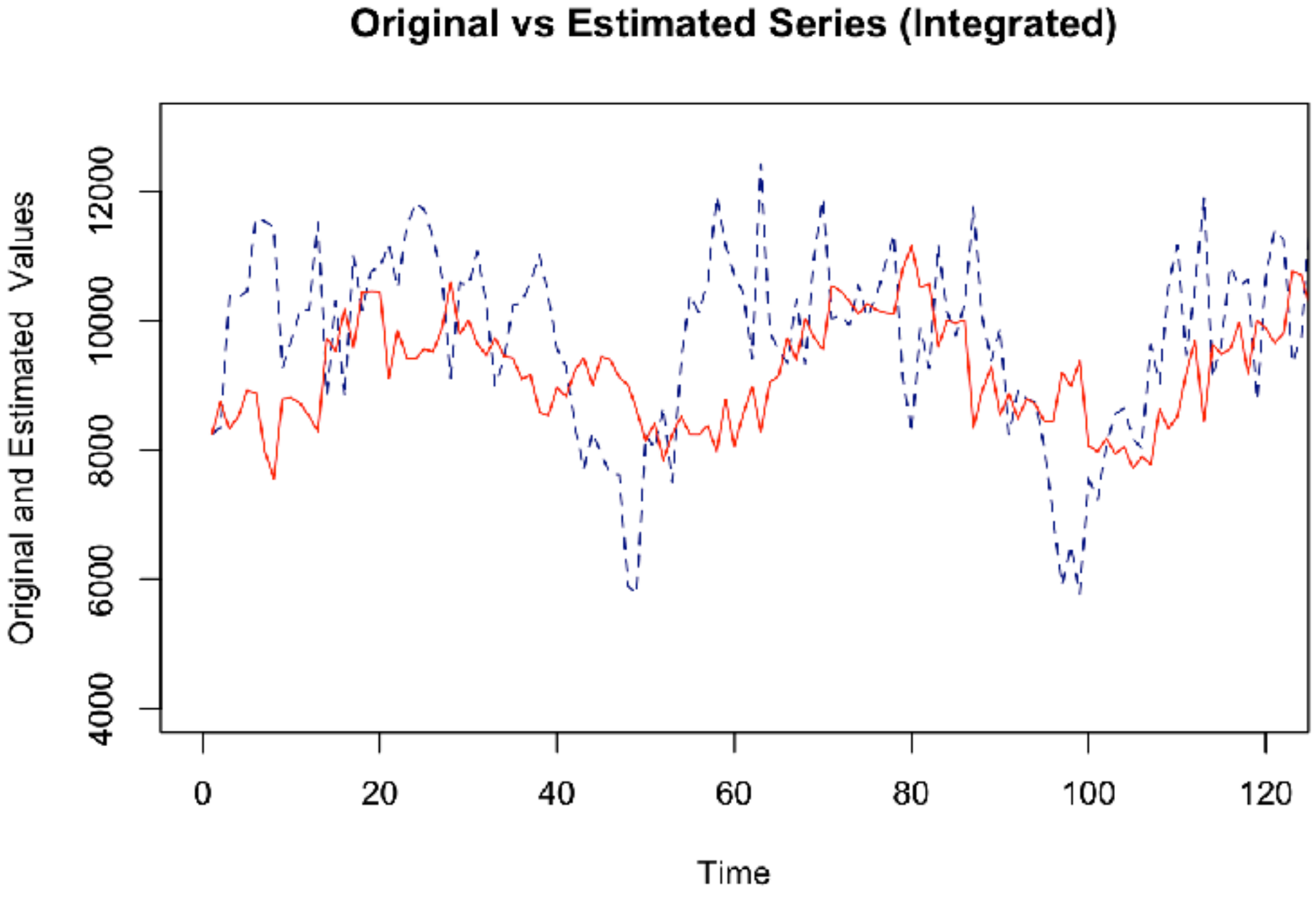
Independence Test

Based on the Ljung-Box test, the residuals of the model can be considered independent and identically distributed.

Normality Test

The histogram and Q-Q plot shows that the residuals are approximately normally distributed.

Model Diagnostics - In Sample Fit



Hypothesis Testing

- There is a strong relationship between sales and first lag of temperature. However , the 2 time series are co-integrated

- No statistically significant relationships between CPI , Fuel price and sales could be detected

- There appears to be a relationship between sales and first lag of unemployment

Forecasting