Book Retailer Sales of Over 2 Decades

By: Emilio Avalos, Jaime Fastino, Nicole Heidi Romangsuriat

Agenda

01.
Introduction to Problem

02.
Time Series Plot & GG Seasonal Plot

03.
Models Investigated + Analysis

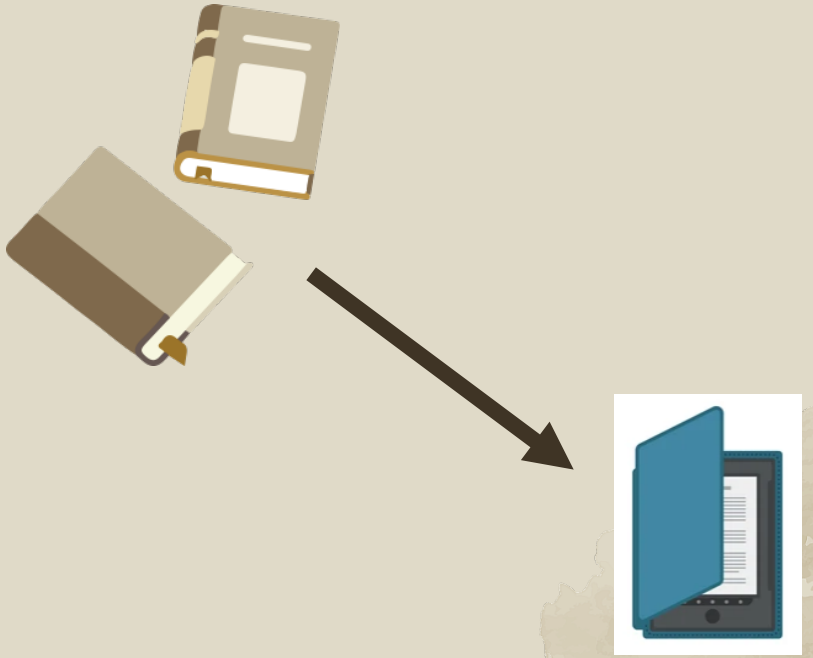
04.
Accuracies & Recommendation

Introduction to Problem

Bookstore Retail Sales

- Recent discussion about e-book platforms -> Interest in investigating effects on bookstore sales

- Original Hypothesis:
  The emergence of bookstore alternatives such as e-book platforms (eg. Kindle and Wattpad) and digital comic platforms (eg. Webtoon) have caused brick-and-mortar bookstore sales to dwindle



Considered Factors Influencing Book Retail Industry Sales

01. Education Rates

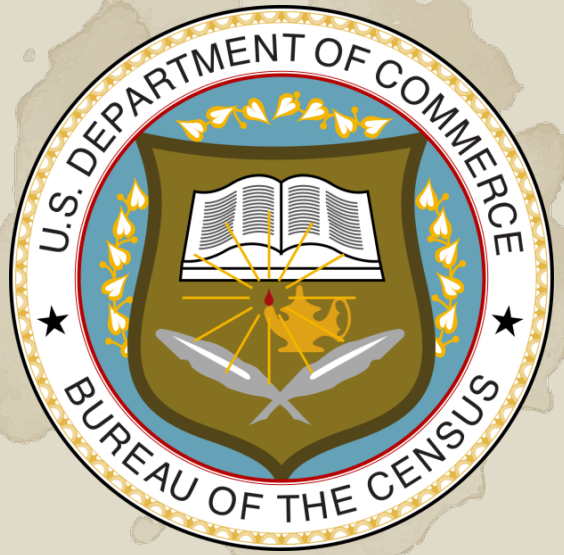
02. Literacy Rates

03. E-Book Industry

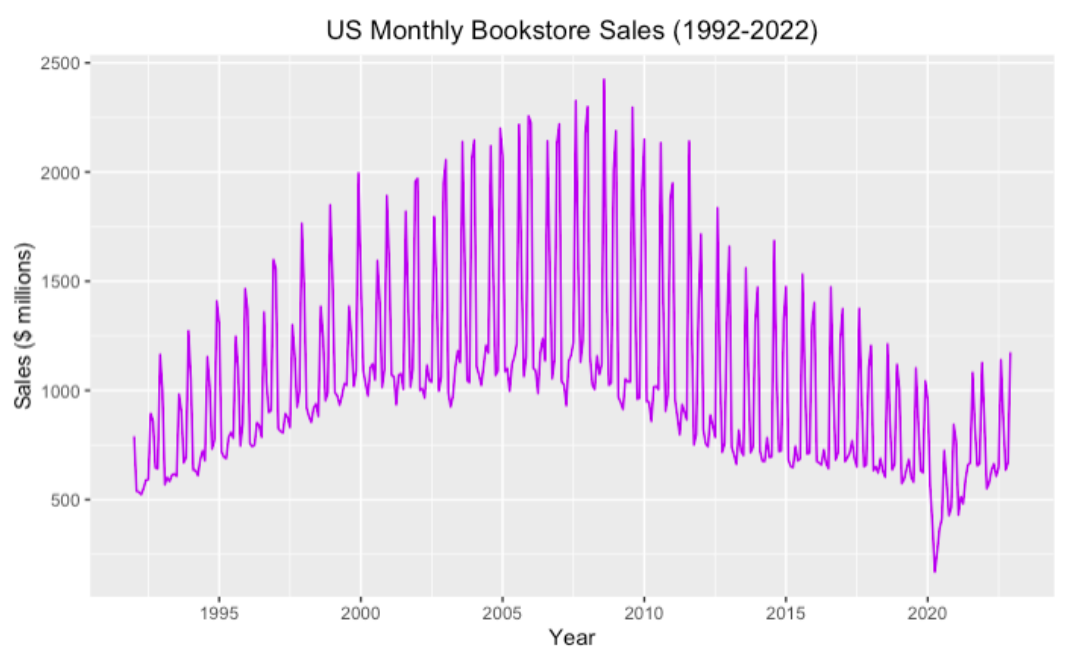
04. “Book Thrifting”

Dataset

- U.S. Census Bureau
- “Retail Food and Services Sales 1992-2022” dataset
- Compiled monthly book retailer sales into a dataset
- 372 data values



Time Series Plot



Qualitative Analysis

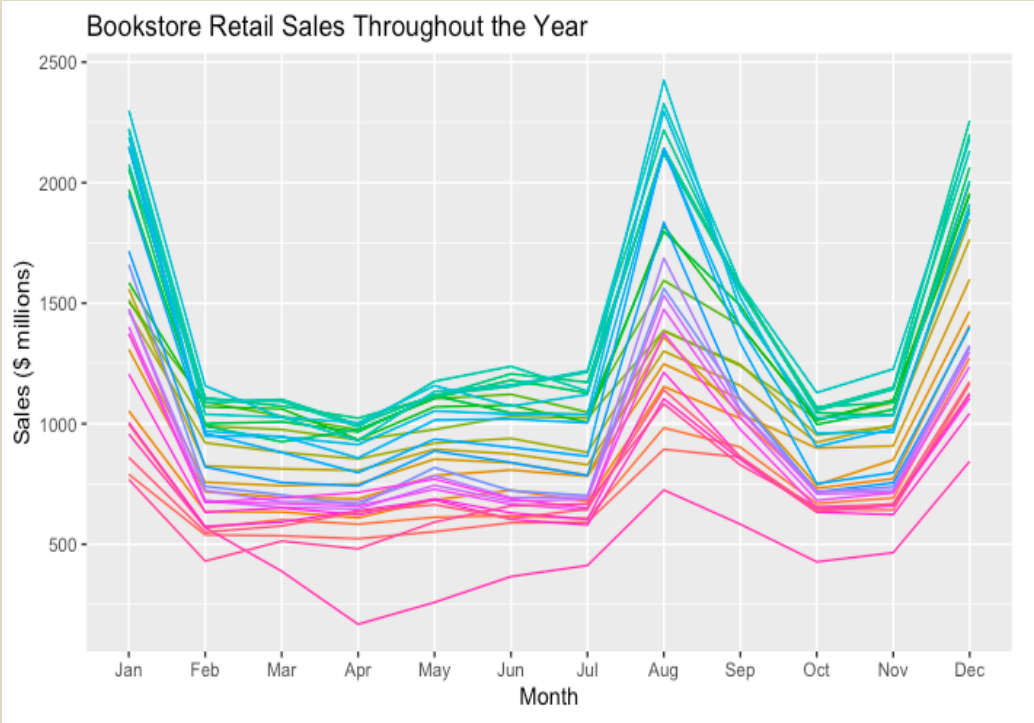
- Non-constant trend
  - ↑↓↑
- Non-constant annual seasonality!
- Outlier: 2020, COVID

Annual Seasonality

**January: back-to-school (spring)

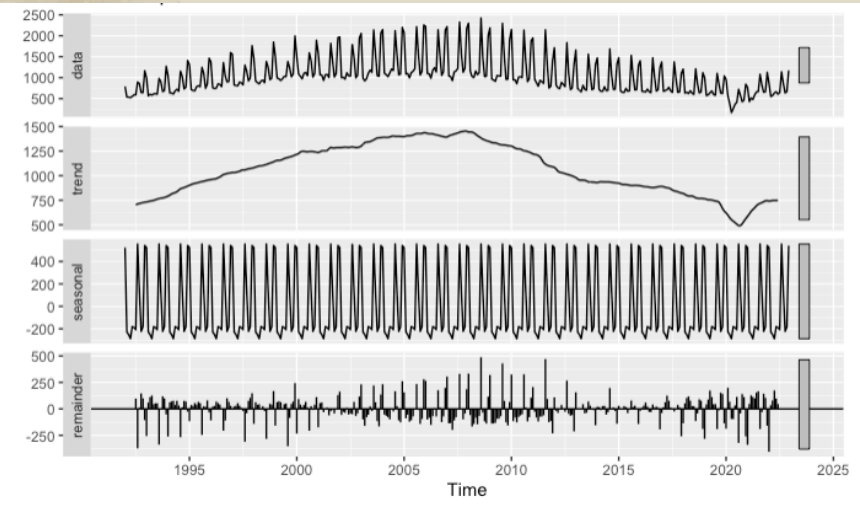
**August: back-to-school (fall)

**December: Christmas shopping



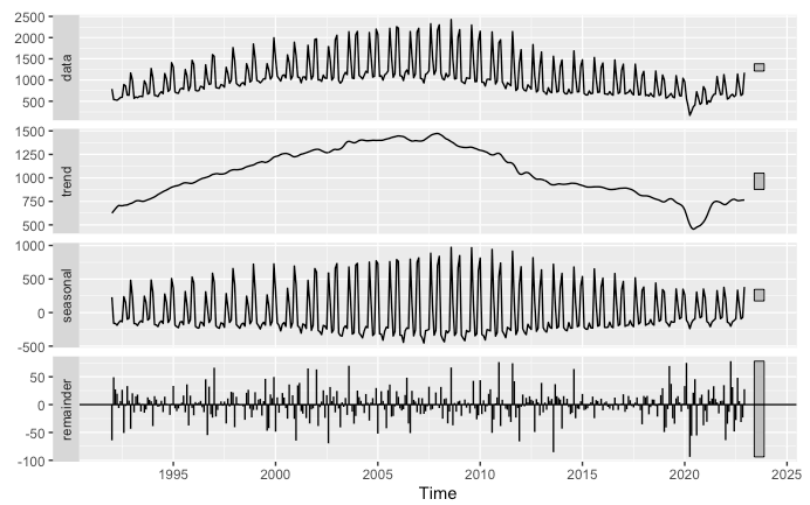
Models Investigated + Analysis

Decomposition



Classical

Wave-pattern in remainder



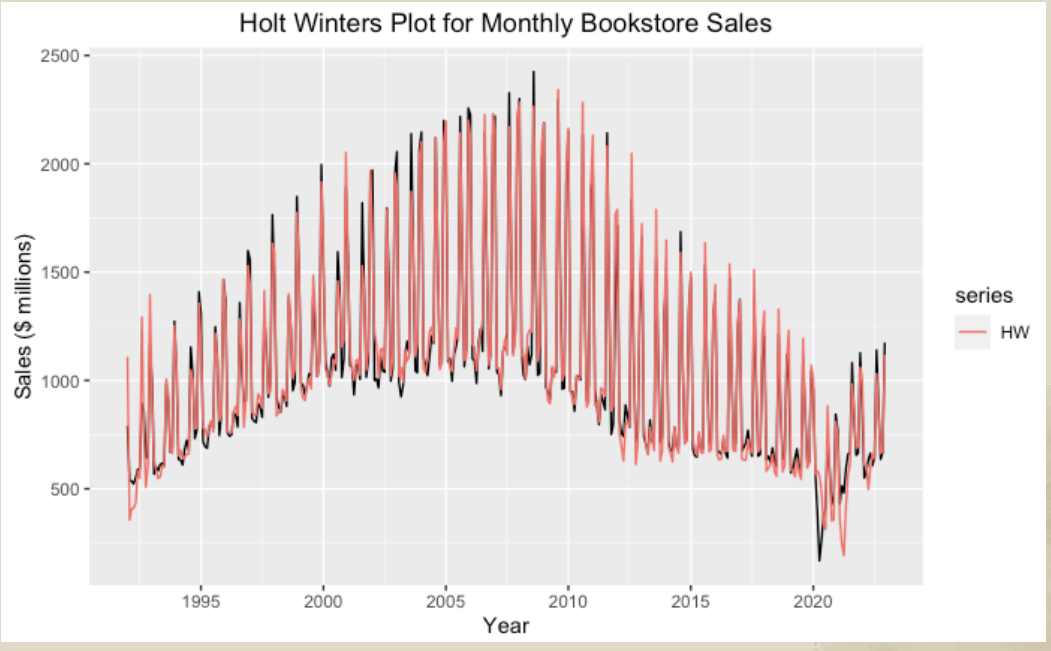
STL

Models non-constant seasonality better

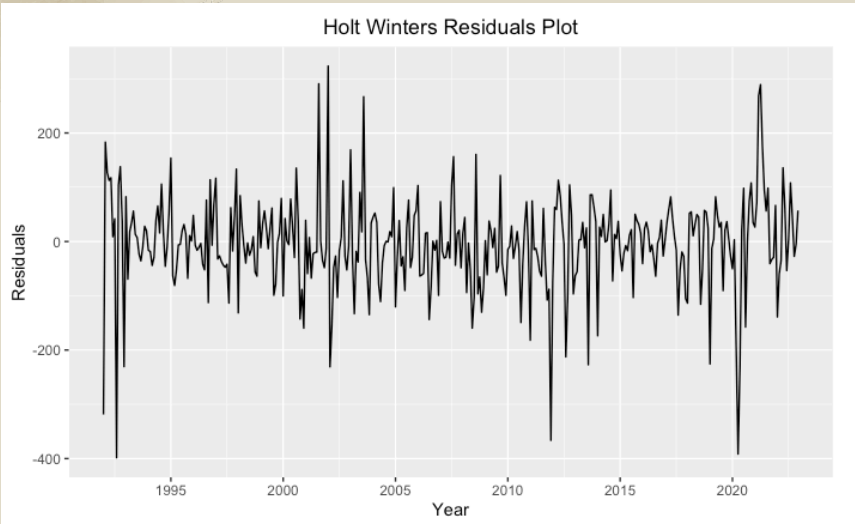
Holt Winters

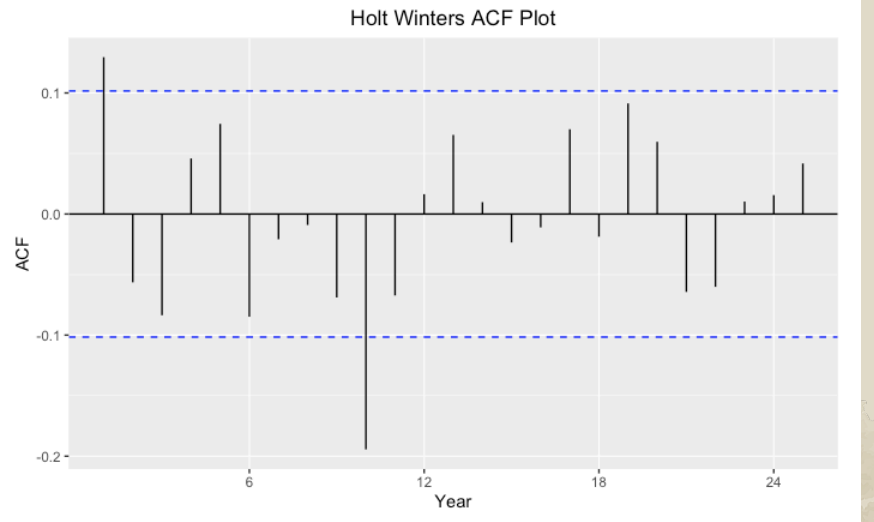
```
Smoothing parameters:
  alpha = 0.2527
  beta  = 0.0032
  gamma = 0.7417
```

- Some smoothing of level
- Little smoothing of trend
- Significant smoothing of seasonality
- Good model fit!



Holt Winters Residuals





- White noise

- 2 autocorrelations
- Lag 1, Lag 10

SARIMA Family of Models

1) Follows auto.arima recommendation of ARIMA(1,1,1)(0,1,1)[12]

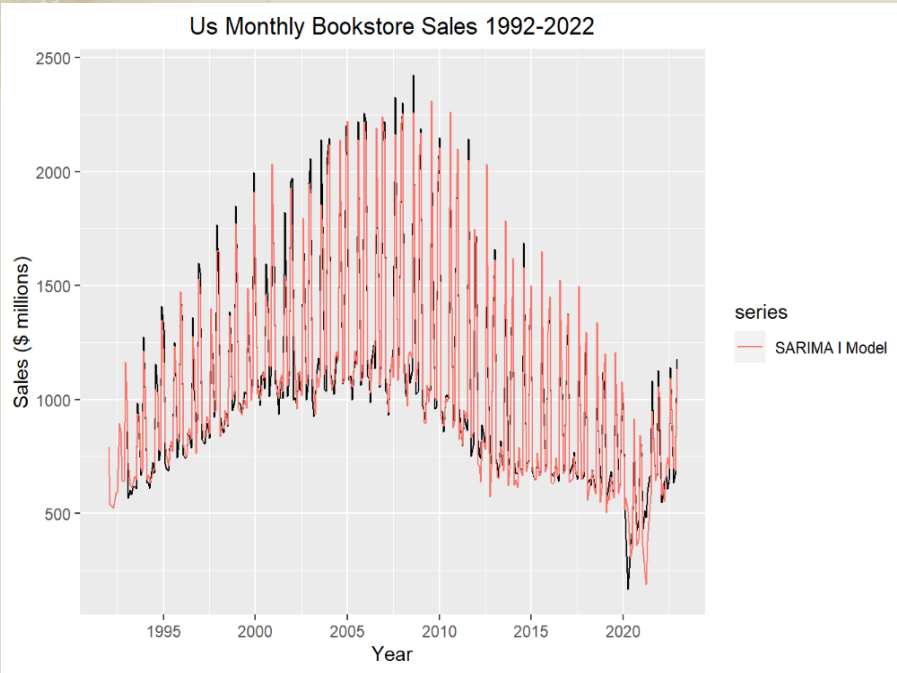
- All terms statistically significant given the 95% confidence interval
- Residuals fairly stationary and minimal autocorrelation
- RMSE of 80.78

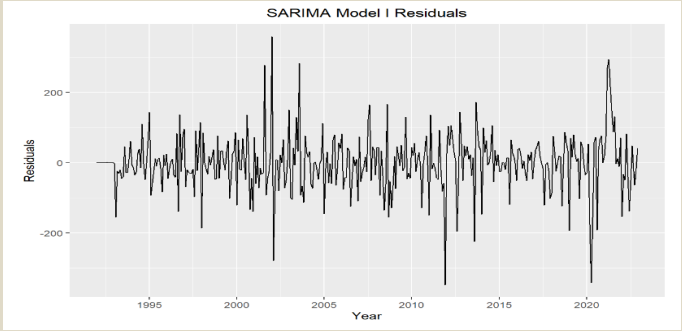
↑↑↑ **Chosen Model**

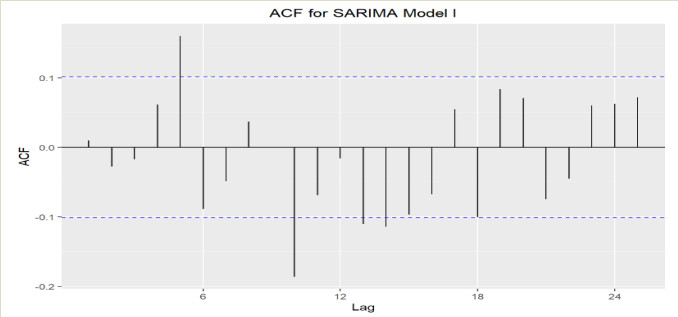
2) Increased seasonal autoregressive term with ARIMA(1,1,1)(1,1,1)[12]

- Extra seasonal autoregressive term was insignificant
- Residuals and autocorrelation similar to the first SARIMA model
- RMSE of 80.58, seasonal AR insignificant

SARIMA FAMILY



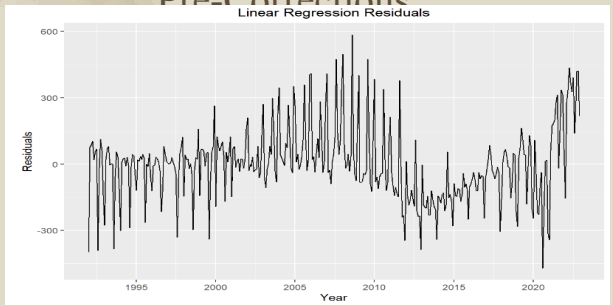


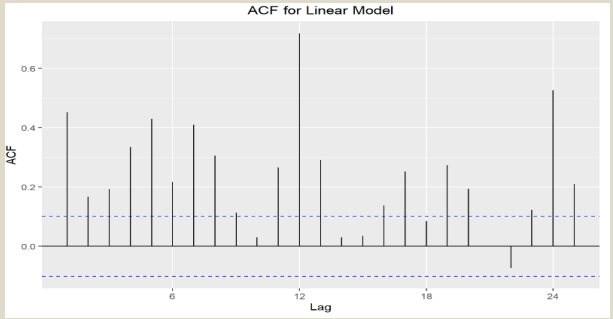


Linear Regression

$$BookSales = 1,555.38 - 1,501.30trend - 4,358.98trend^2 - 734.46S_2 - 755.69S_3 - 791.19S_4 - 685.81S_5 - 691.71S_6 - 708.89S_7 + 45.20S_8 - 369.18S_9 - 720.77S_{10} - 681.04S_{11} + 38.01S_{12}$$

Pre-Corrections



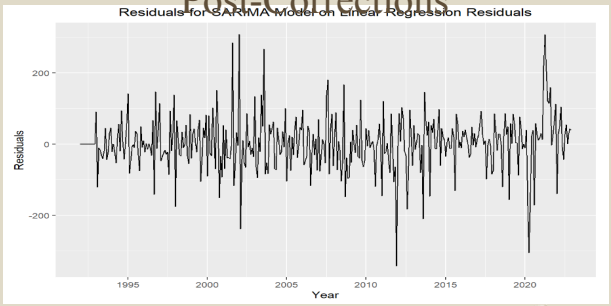


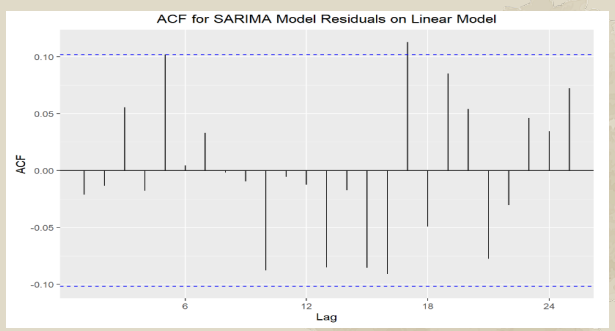
- All variables significant except for August & December

ARIMA (3,0,3)(0,1,1)[12] on Residuals

- Improved RMSE

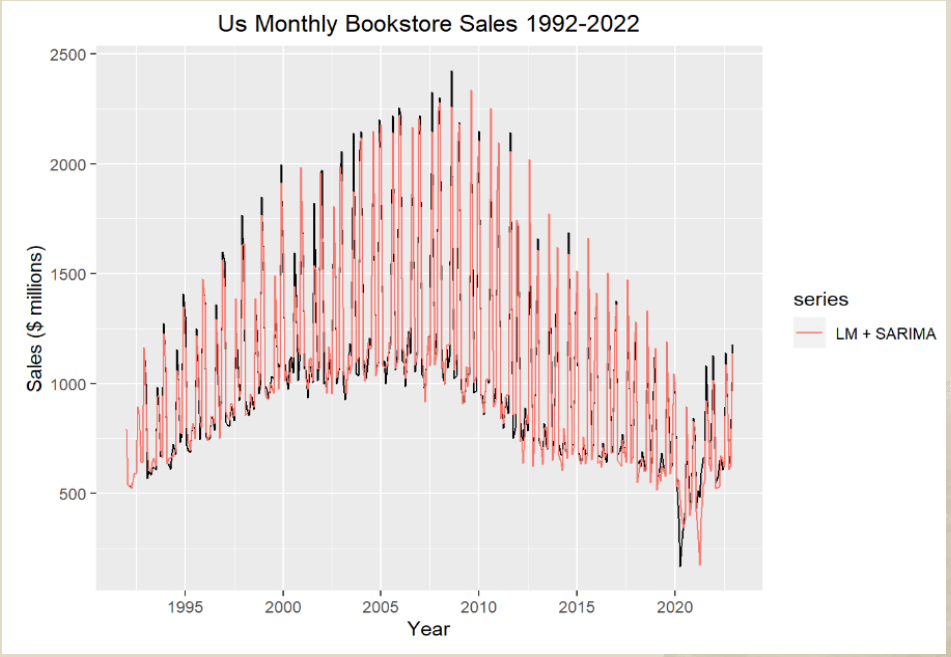
Post-Corrections





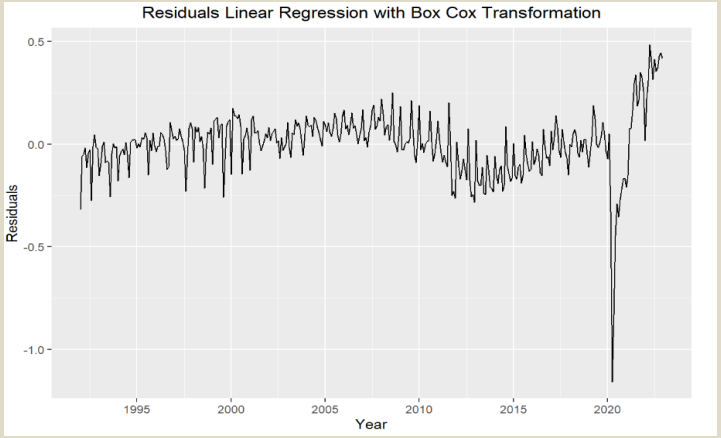
Linear Regression with SARIMA

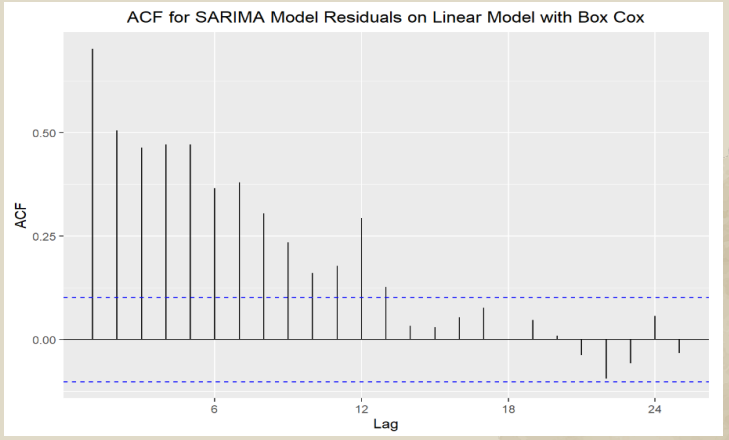
- All variables significant; except for August and December dummy variables
- Improved RMSE from 165.88 to 76.65, an improvement of 53.79%
- Multiple R-Squared: 85.43%



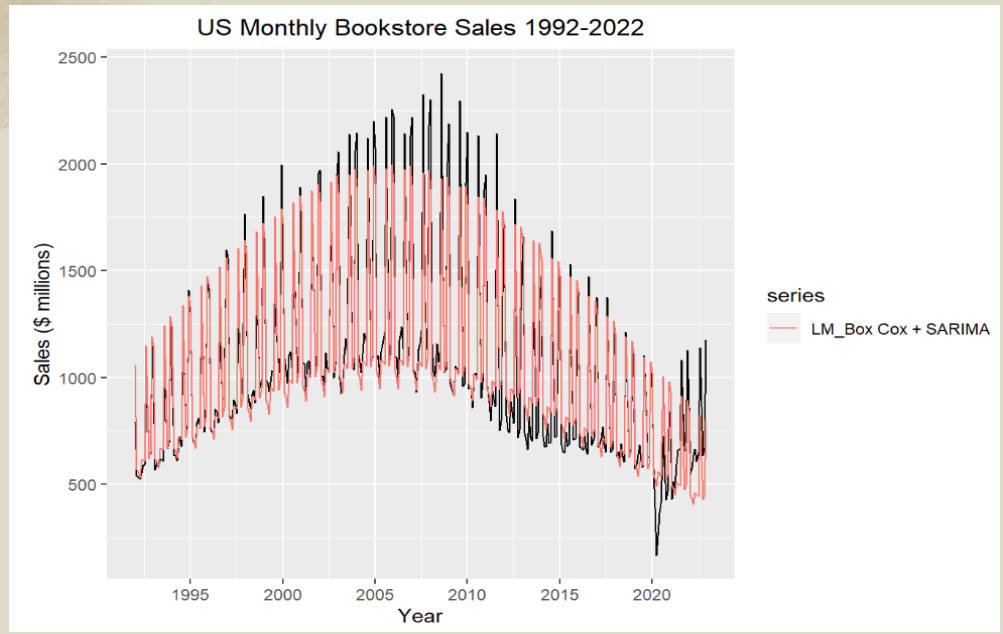
Linear Regression with Box Cox Transformation

- All variables significant; except for August and December dummy variables
- Box Cox transformation did not improve the linear model that much – will need to conduct a ARIMA(1,0,2)(1,0,2)[12] on the residuals





Linear Regression with Box Cox & SARIMA



- After applying SARIMA model on residuals:
  - RMSE of 127.89
  - Multiple R-Squared: 87.35%

Accuracies & Recommendation

Comparison of RMSE Values

Model	S.Naive	Holt-Winters	ARIMA	SARIMA	Linear Regression		
					Normal	+ SARIMA on Residuals	+ Box Cox & SARIMA on Residuals
RMSE Value	437.41	87.71564	80.78441	80.58239	165.8814	76.65521	127.8861
	7th	4th	3rd	2nd	6th	1st	5th

Recommended Model

- Linear Regression with SARIMA on Residuals
- Lowest RMSE value -> most accurate
- Best Model Fit
- Best for in-sample predictions / forecasting

Thank You!

Questions are welcome.