BUDT 730

Data, Models and Decisions

Lecture 16

Time Series (1)

Prof. Sujin Kim

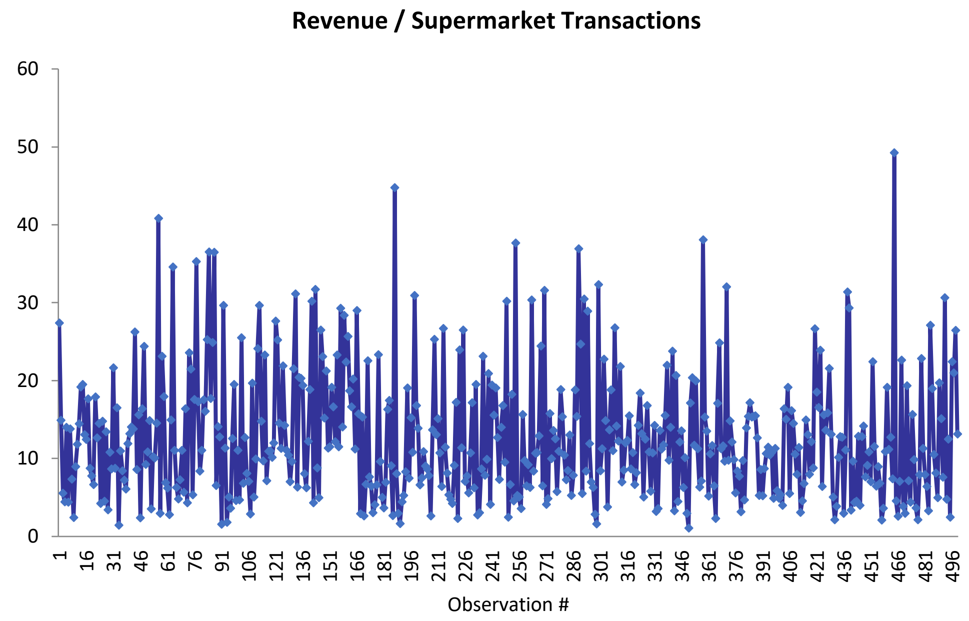
Agenda

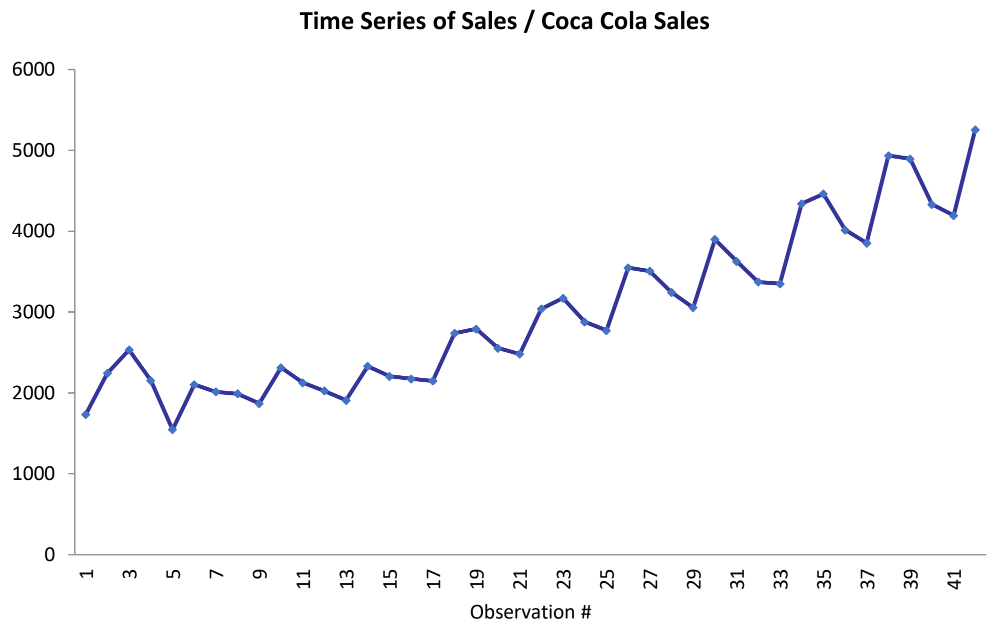
- Time series building blocks (Components)
- Regression-based methods
- Data file: Coca Cola Data.xlsx

Cross-Sectional vs. Time Series Data

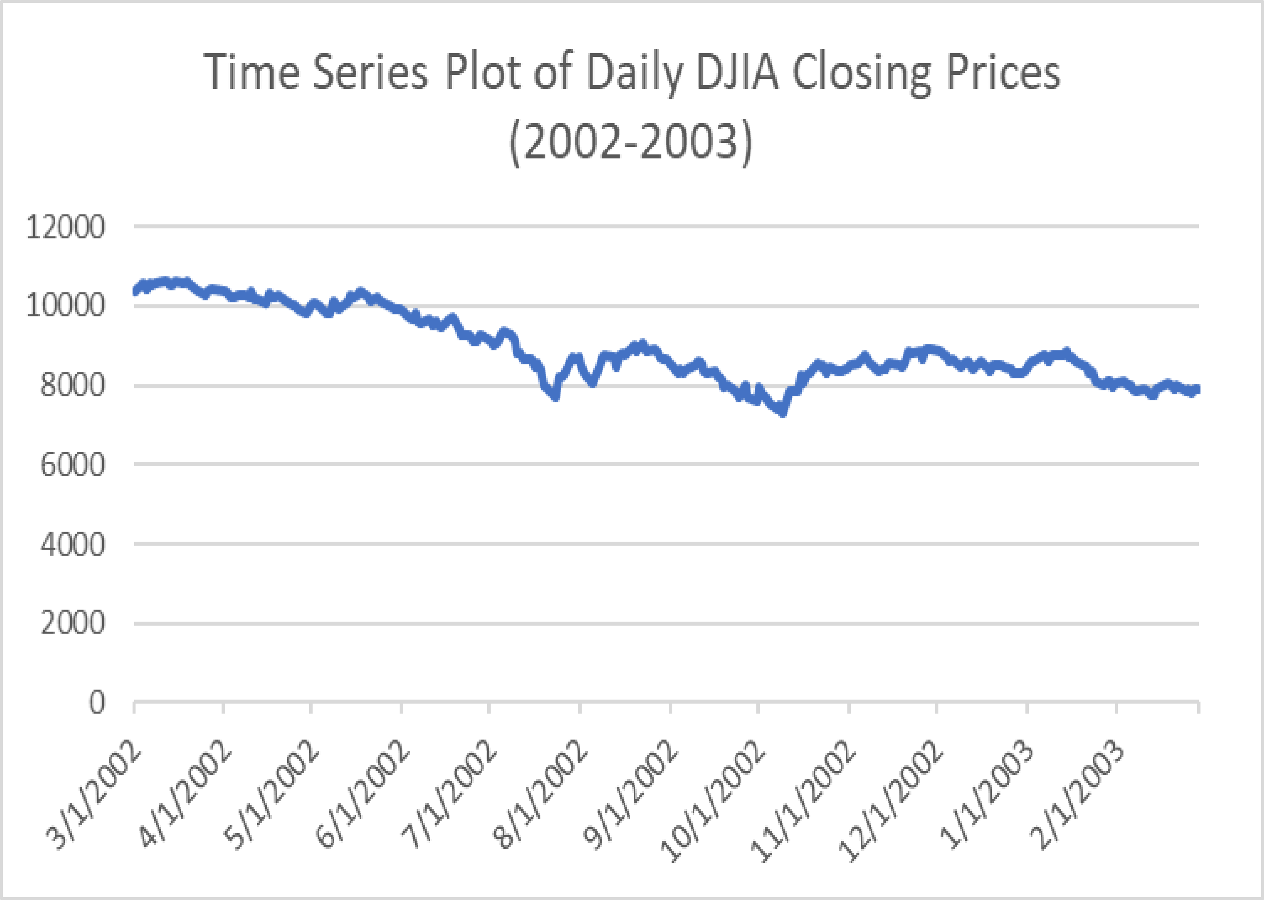
A set of data collected at the same pointe of time

A series of data points indexed in time order





Example: Dow Jones Industrial Average



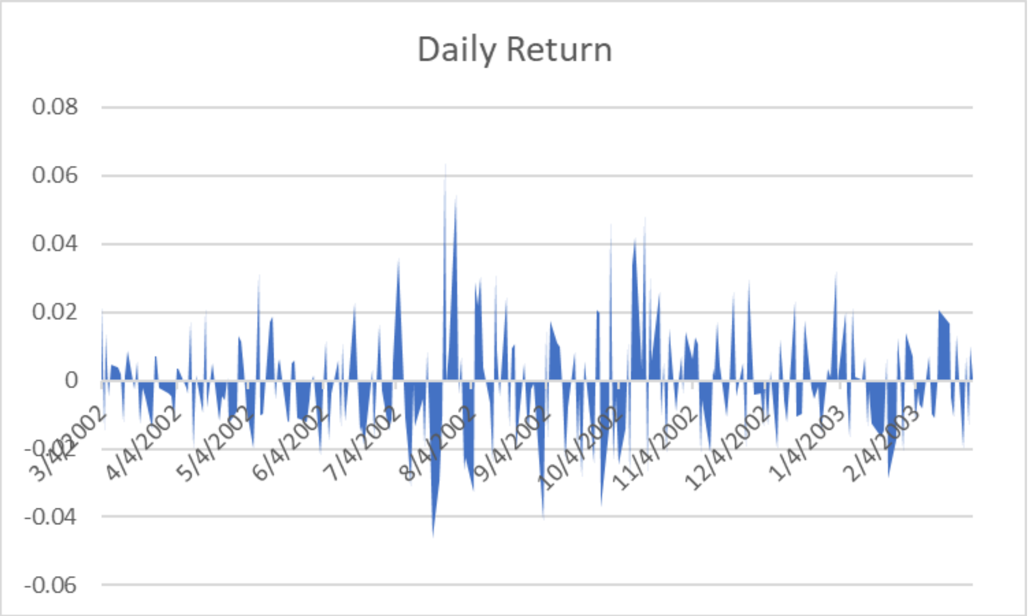
Is this a times series?

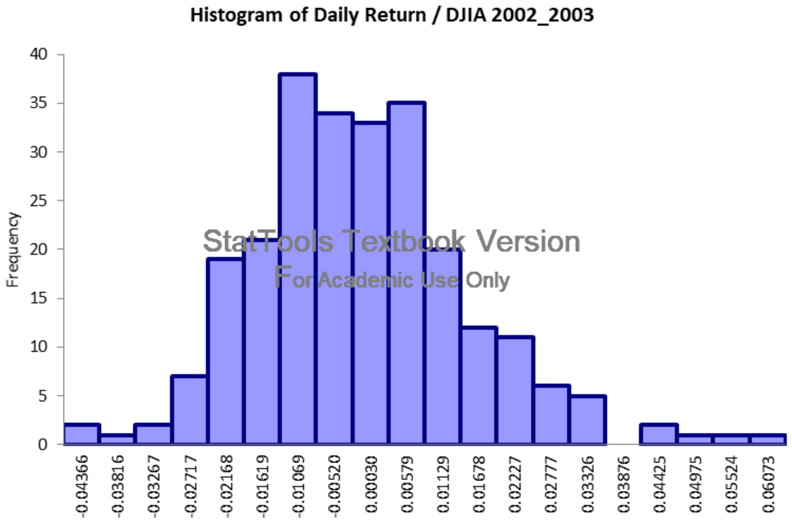
Example: Dow Jones Industrial Average

- Daily returns:

$$DR_{Today} = \frac{DJIA_{Today} - DJIA_{Yesterday}}{DJIA_{Yesterday}}$$

Is this a times series?

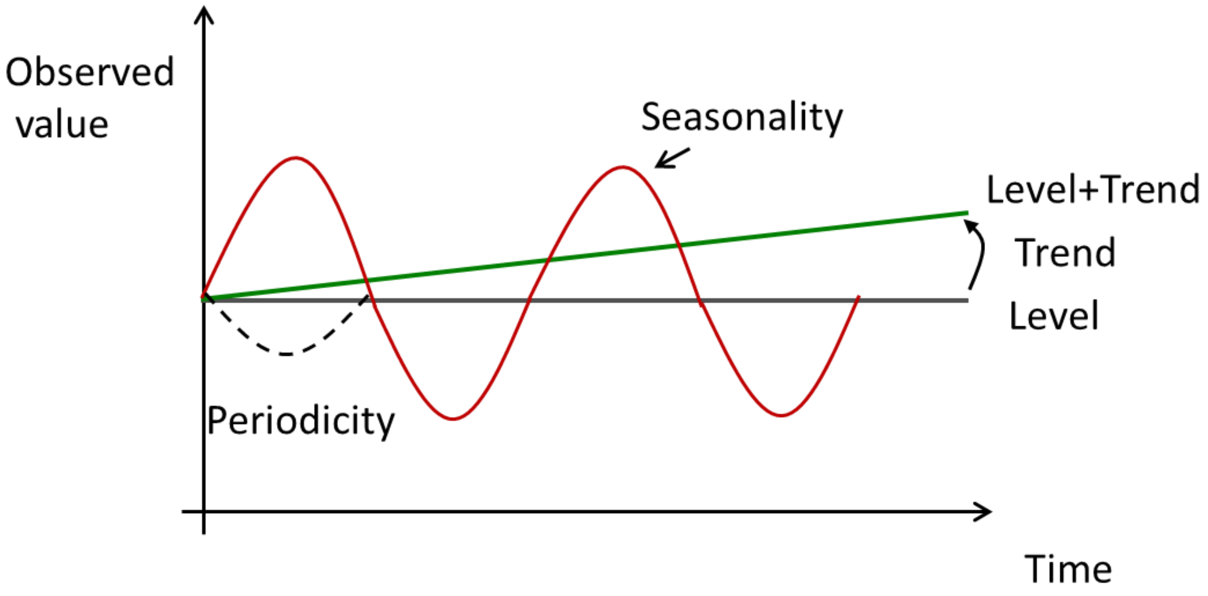




We consider the daily returns as a cross sectional data and use a Normal probability model.

Building Blocks of a Time Series

1. Level (always present)
2. Trend: steady increase/decrease over time.
3. Seasonality (Periodicity): pattern that repeats itself every season
4. Random/noise (always present)
   - Error component
   - Amount that is unexplained
   - Forecast error
     = observed value – forecast



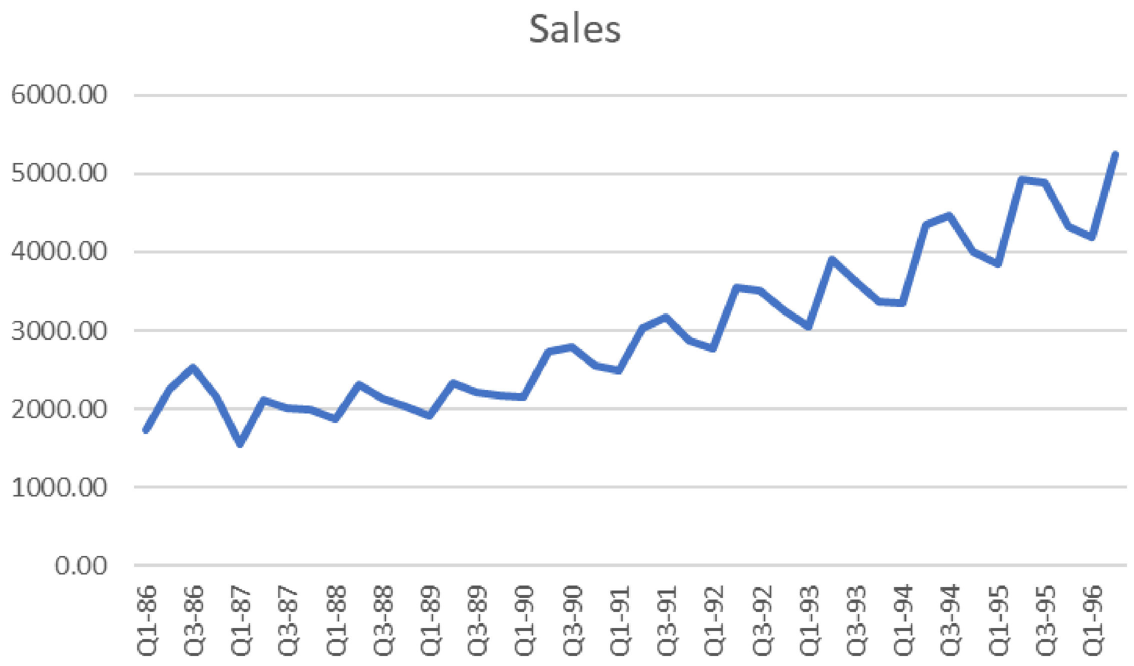
A Motivating Example – Coca Cola Sales

- Data contains quarterly sales for Coca Cola (in $M) from Q1-1986 to Q2-1996
- Our goals are the following:
  - Explore several time series models for **forecasting sales for 4 subsequent quarters:**
    - Training Data: Q1-1986 to Q2-1995
  - We will **validate our model against the last 4 quarters**:
    - Validation Data: Q3-1995 to Q2-1996

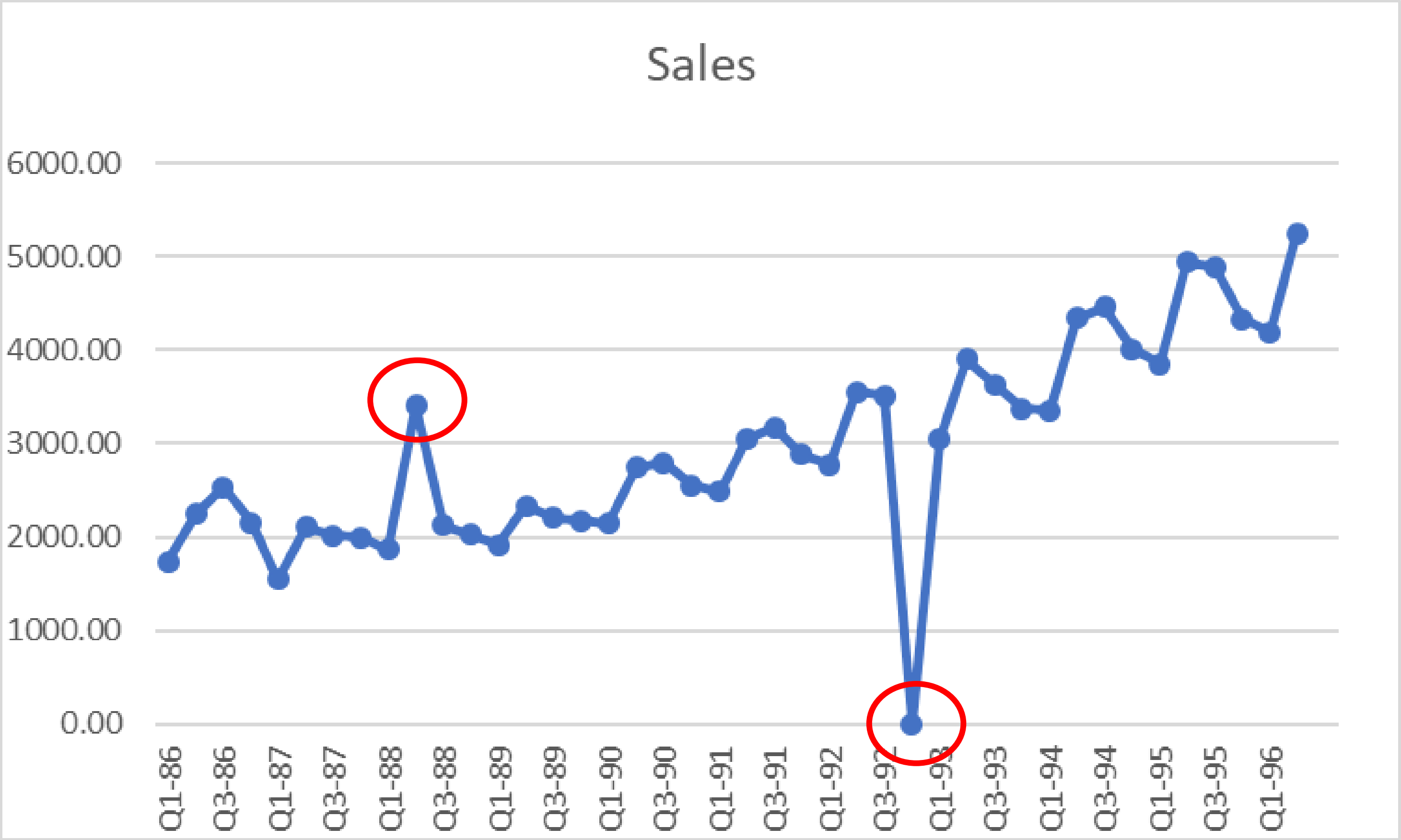
First Step: Time Series Plot

From the time series plot that the series exhibits:

- Level
- Trend
- Seasonality
- Noise

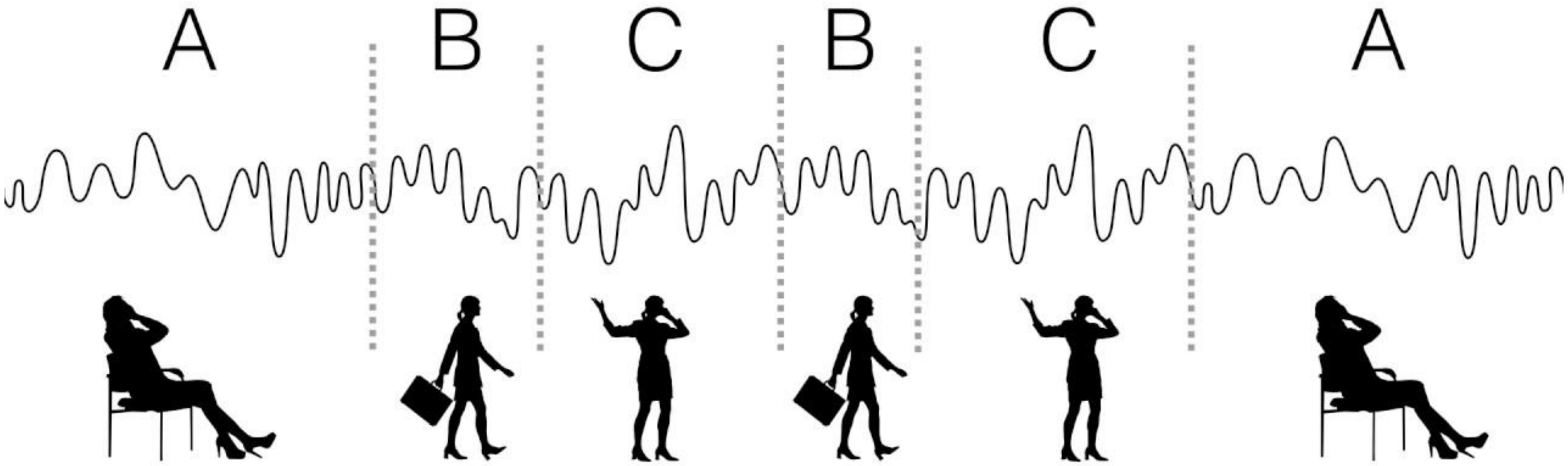


First Step: Time Series Plot



Check if there are any outliers or missing values

Regression Based Methods



Regression Model

- M = number of seasons
- Building Blocks of Time Series
  - Level: Intercept ($\boldsymbol{a}$)
  - Trend: Time variable (in days/months/quarters/years, $\boldsymbol{t}$)
  - Seasonality: Dummies for each season period ($\boldsymbol{S_i}$)
  - Noise: Residual = Forecast Error ($\boldsymbol{\varepsilon}$)

$$\underbrace{\boldsymbol{Y_t}}_{\text{Observed value}} = \boldsymbol{a} \underbrace{+ \boldsymbol{b_t} * \boldsymbol{t} + \boldsymbol{b_1} * \boldsymbol{S_1} + \boldsymbol{b_2} * \boldsymbol{S_2} + \ldots . \boldsymbol{b_{M-1}} * \boldsymbol{S_{M-1}}}_{\text{Forecast}} + \boldsymbol{\varepsilon}$$

Add up all four components: **Additive (Linear) Model**

Process the Data

- We need variables to capture each component of the time series:
  - Level – Intercept
  - Trend – Add running number for time (**t**)
  - Seasonality – Add dummies for each season
    - For quarterly seasons, we have four dummy variables: Q_1, Q_2, Q_3 , Q_4
    - Use 3 variables (e.g., Q_1, Q_2, and Q_3) whose coefficients indicate how much each quarter differs (on average) from the reference quarter (Q_4)
- Partition:
  - Training data: Q1-1986 to Q2-1995
  - Validation data: Q3-1995 to Q2-1996 (last four quarters)

Linear Regression Model

- Open ‘Coca Cola Processed Data.xlsx’
- Build a regression model using ‘Training Data’
- Dependent variable: Sales
- Independent variables: time and seasonality
- Compute RMES and MAPE

Example: Coca Cola Sales

Steps to process data:

1. Add the number of quarters since start, T
2. Add a categorical variable ‘QuarterIndex’
3. Create dummy variables for each quarter using ‘QuarterIndex’.
   (For other regression models, also create T^2 and Log(Sales))
4. Partition data: Save a training data (‘train’) and a validation data (‘test’)

R packages and functions

- Package ‘Metrics’ : to compute RMSE and MAPE.
- Functions:
  - length(): set the length of vectors
  - lines(): A generic function taking coordinates given in various ways and joining the corresponding points with line segments.

R Script

```r
# load Metrics to compute RMSE and MAPE
library(Metrics)

# time series plot
plot(Coca_Cola_Processed$T,Coca_Cola_Processed$Sales, type="l")
```

```r
# Splitting data
ndata<-length(Coca_Cola_Processed$T) # length of data
nTrain <- ndata - 4 #length of training data
train<-Coca_Cola_Processed[1:nTrain,]
test<-Coca_Cola_Processed[nTrain+1:4,]
```

```r
# Trend model
train.lm.trend<- lm(Sales ~ T,data=train)
summary(train.lm.trend)
observed<-test$Sales
predicted<-predict(train.lm.trend,test)
```

```r
# plot data and forecasts
plot(Coca_Cola_Processed$T,Coca_Cola_Processed$Sales, type="l")
# plot fitted values in the training period
lines(train.lm.trend$fitted, lwd = 2)
lines(c(nTrain+1:4),predicted, lwd = 2, col="blue")


# compute rmse and mape
rmse.lm.trend<-rmse(observed,predicted)
mape.lm.trend<-mape(observed,predicted)*100
print(c(rmse.lm.trend,mape.lm.trend))
```

Trend Model

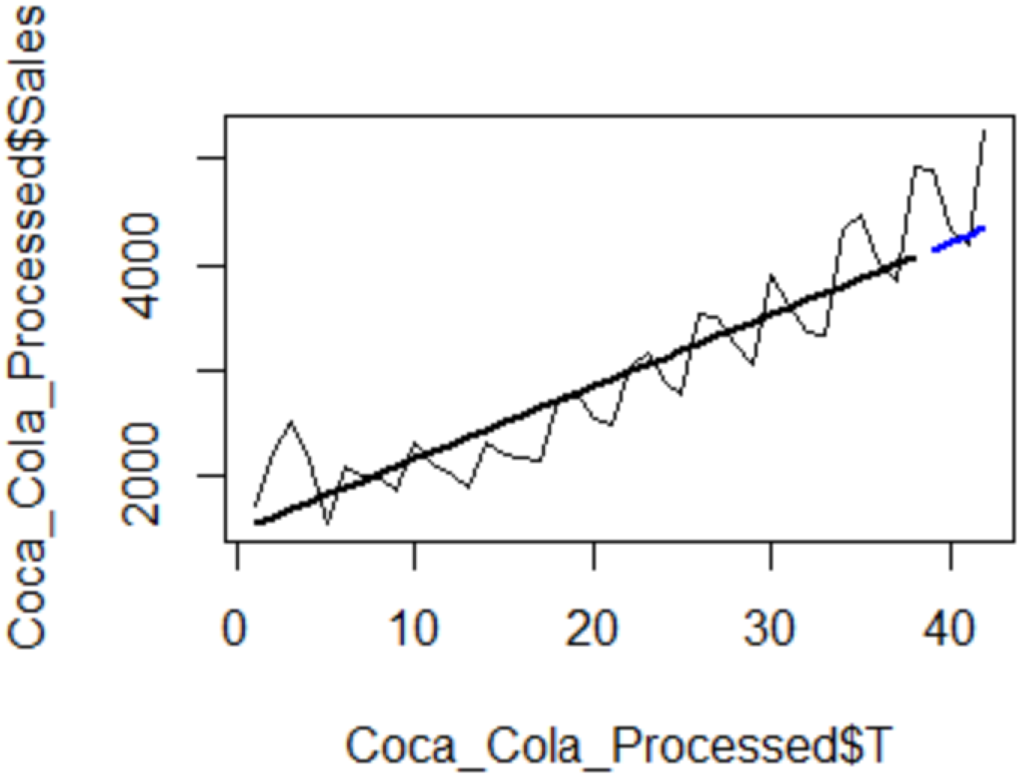
```
Call:
lm(formula = Sales ~ T, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-499.65 -292.80  -17.43  178.54  858.60 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1490.736    122.247   12.19 2.41e-14 ***
T             68.070      5.464   12.46 1.29e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 369.4 on 36 degrees of freedom
Multiple R-squared:  0.8117,	Adjusted R-squared:  0.8065 
F-statistic: 155.2 on 1 and 36 DF,  p-value: 1.291e-14
```

Linear Model: Results



Linear Model with rend	
RMSE	591
MAPE	9.34%

Linear Model: Trend + Seasonality

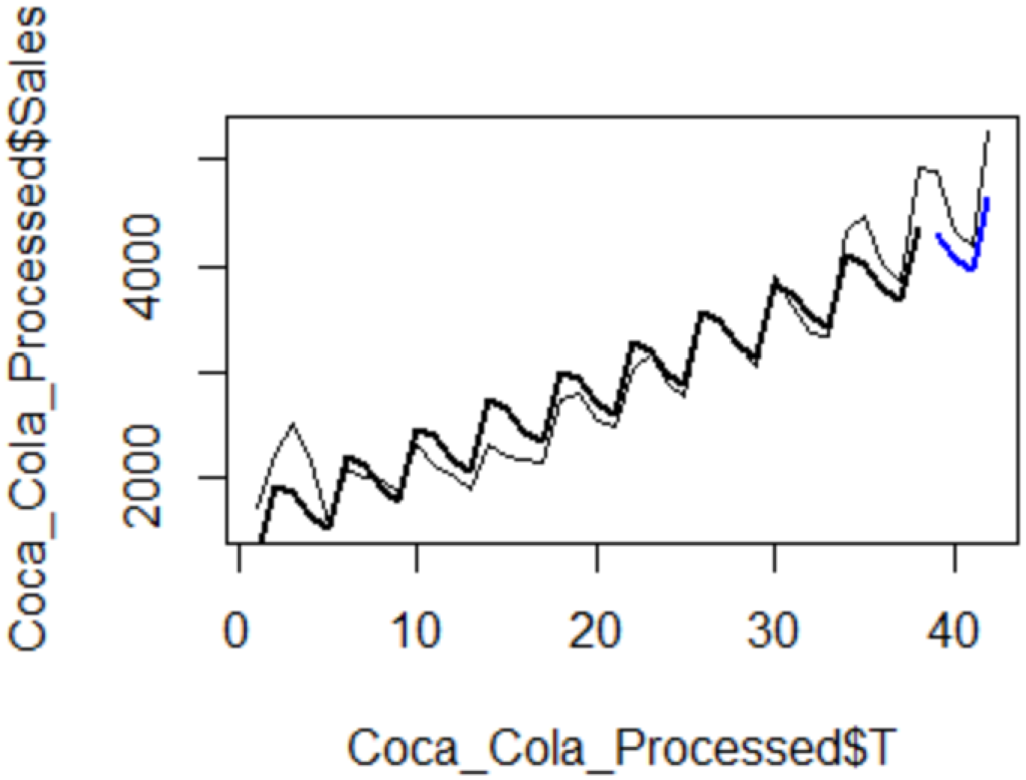
```
Call:
lm(formula = Sales ~ T + factor(QuarterIndex), data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-461.12 -154.91  -95.69   86.69  678.44 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1186.416    120.281   9.864 2.28e-11 ***
T                       67.692      4.208  16.086  < 2e-16 ***
factor(QuarterIndex)2  609.753    127.152   4.795 3.37e-05 ***
factor(QuarterIndex)3  465.879    130.564   3.568  0.00112 ** 
factor(QuarterIndex)4  172.684    130.632   1.322  0.19529    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 284.2 on 33 degrees of freedom
Multiple R-squared:  0.8978,	Adjusted R-squared:  0.8855 
F-statistic: 72.51 on 4 and 33 DF,  p-value: 7.111e-16
```

Linear Model: Trend + Seasonality



Linear Model	
RMSE	465
MAPE	8.92%

Interpretation

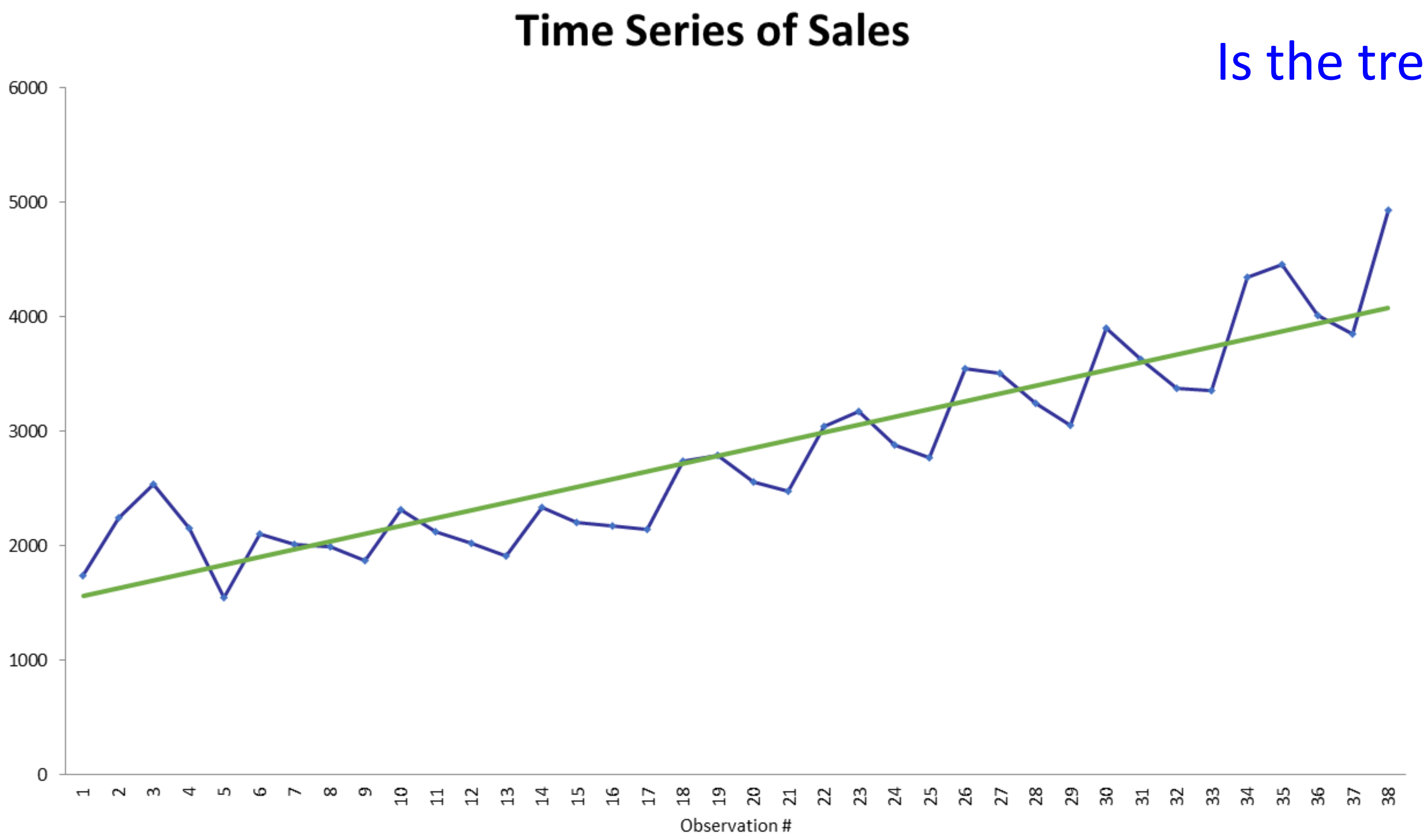
$$\boldsymbol{Y_t = a + b_t * t + b_1 * S_1 + b_2 * S_2 + b_3 * S_3 + \varepsilon}$$

- Coefficient b_t represents the expected change in Y_t from one period to the next
  - If $b_t > 0$, the trend is upward
  - If $b_t < 0$, the trend is downward
- The *a* term is less important, it technically represents the expected value of the series at t = 0
- Coefficient on each dummy is the impact of that seasonality period

Interpreting the Model Coefficients

- Interpretation of b_t= 67.69
  - Seasonally adjusted sales increase by an average of $67.69M per quarter
- Interpretation of b_4 = 172.68
  - After adjusting for trend, sales in Q_4 are higher than sales in $Q_1$by an average of $172.68M

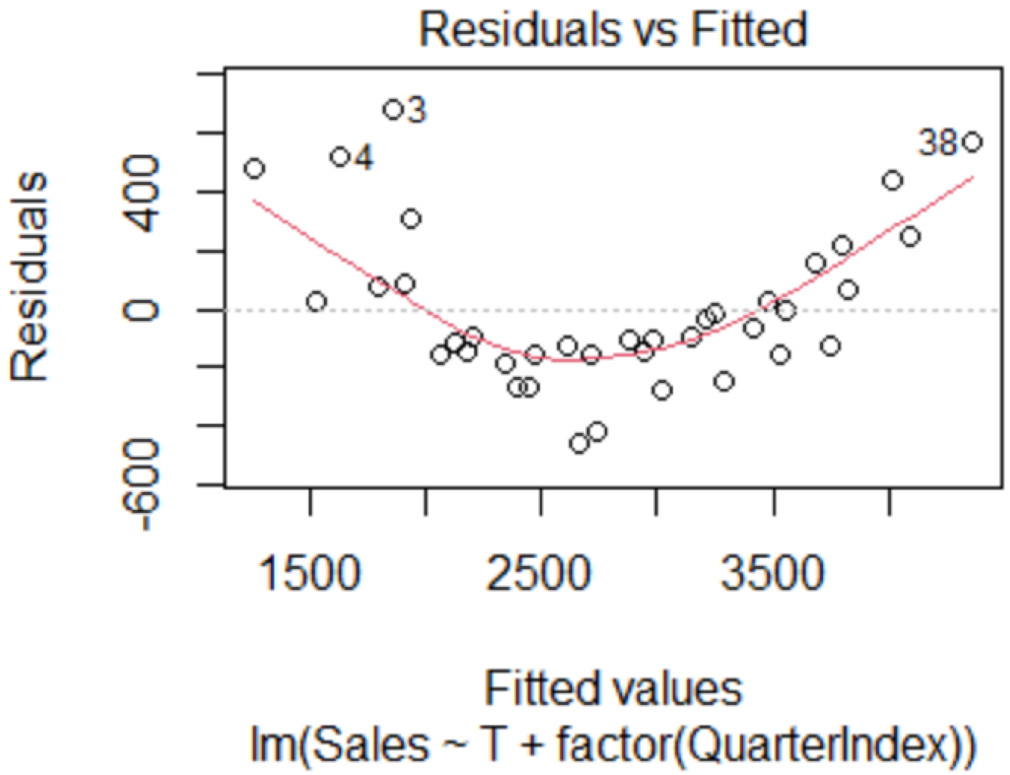
Nonlinear Trend



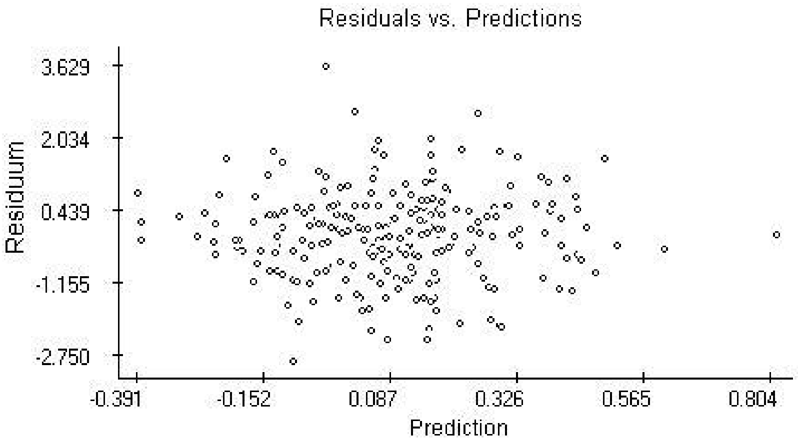
Is the trend really linear?

Scatterplot of Residuals

Exhibit nonlinear trend

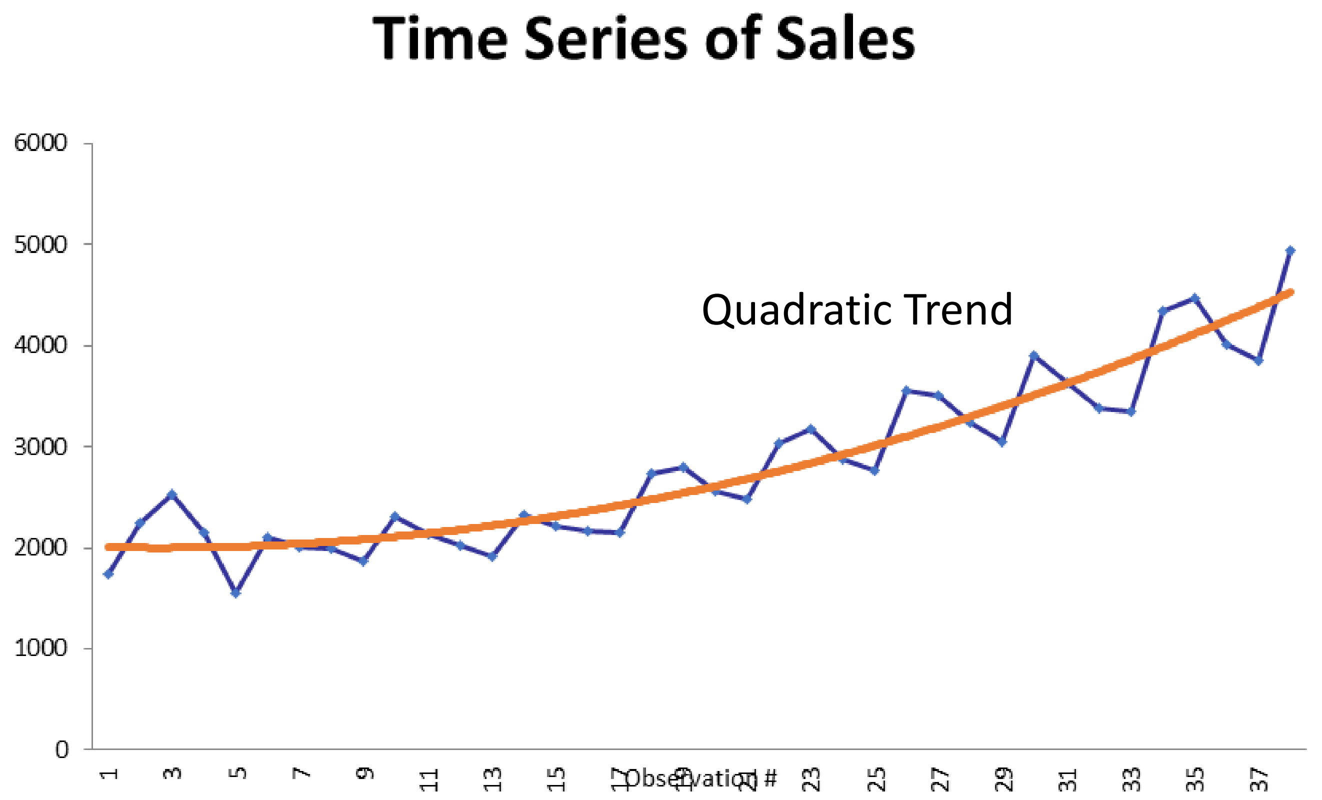


Residual plot without trend



Nonlinear Trend: Quadratic Trend

$$Y_t = a + b_t * t + b_{t^2} * t^2 + b_1 * S_1 + b_2 * S_2 + b_3 * S_3 + \varepsilon$$



Nonlinear Trend: Quadratic Trend

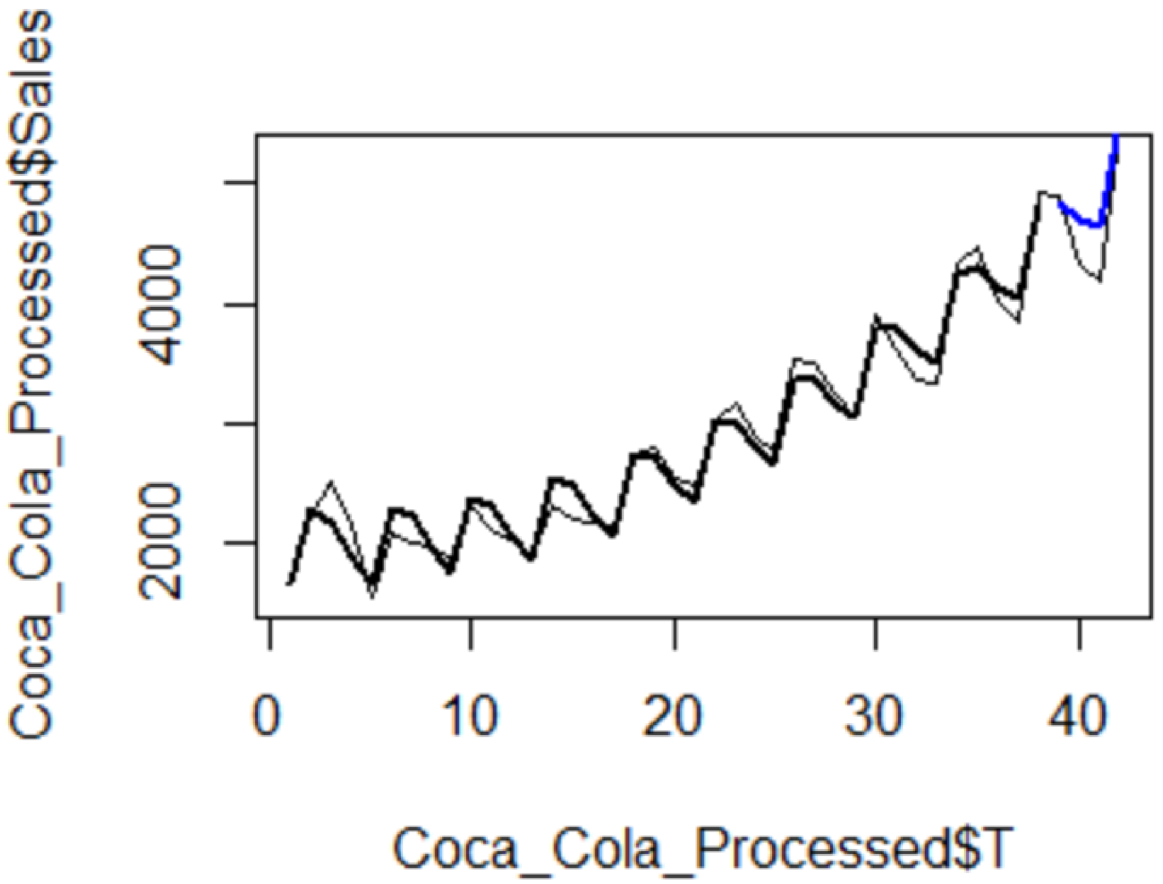
```
Call:
lm(formula = Sales ~ T + Tsqrd + factor(QuarterIndex), data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-283.43 -132.23   33.95  113.50  339.45 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          1694.8878    92.1843  18.386  < 2e-16 ***
T                     -12.2698     9.9389  -1.235  0.22600    
Tsqrd                   2.0503     0.2472   8.292 1.79e-09 ***
factor(QuarterIndex)2 609.7534    72.7656   8.380 1.42e-09 ***
factor(QuarterIndex)3 517.8197    74.9807   6.906 8.11e-08 ***
factor(QuarterIndex)4 224.6248    75.0193   2.994  0.00527 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 162.6 on 32 degrees of freedom
Multiple R-squared:  0.9676,    Adjusted R-squared:  0.9625 
F-statistic: 190.9 on 5 and 32 DF,  p-value: < 2.2e-16
```

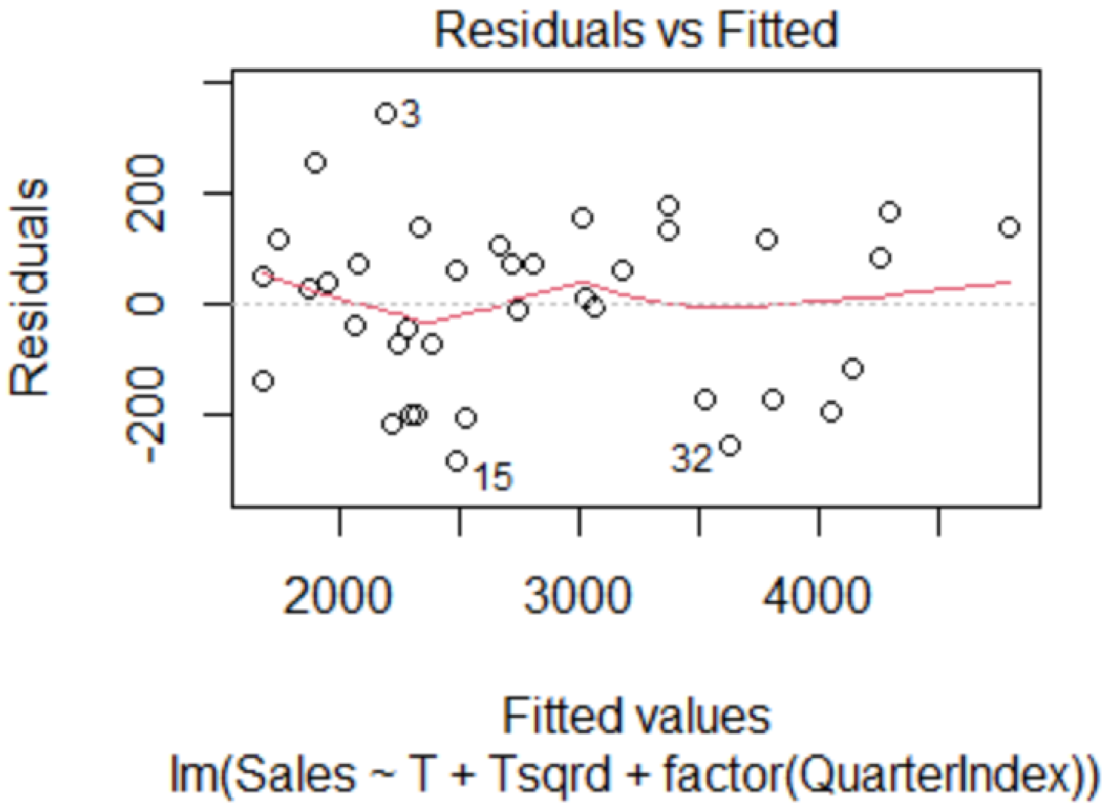
Nonlinear Trend: Quadratic Trend



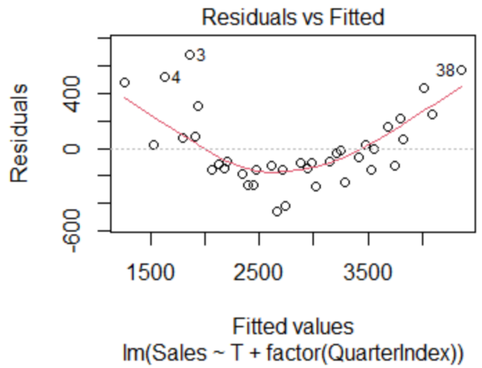
Linear Model	
RMSE	464.98
MAPE	8.92%

Quadratic Model	
RMSE	301.74
MAPE	5.76%

Scatterplot of Residuals: Quadratic Model



Residual plot of the linear model



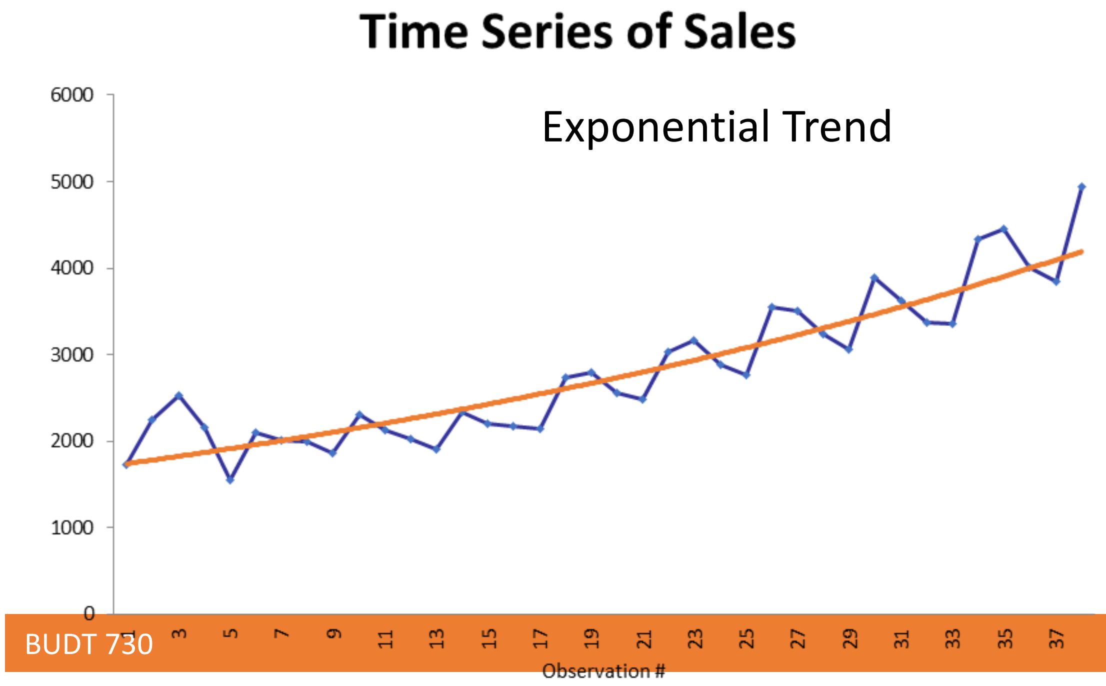
Nonlinear Trend: Exponential Trend

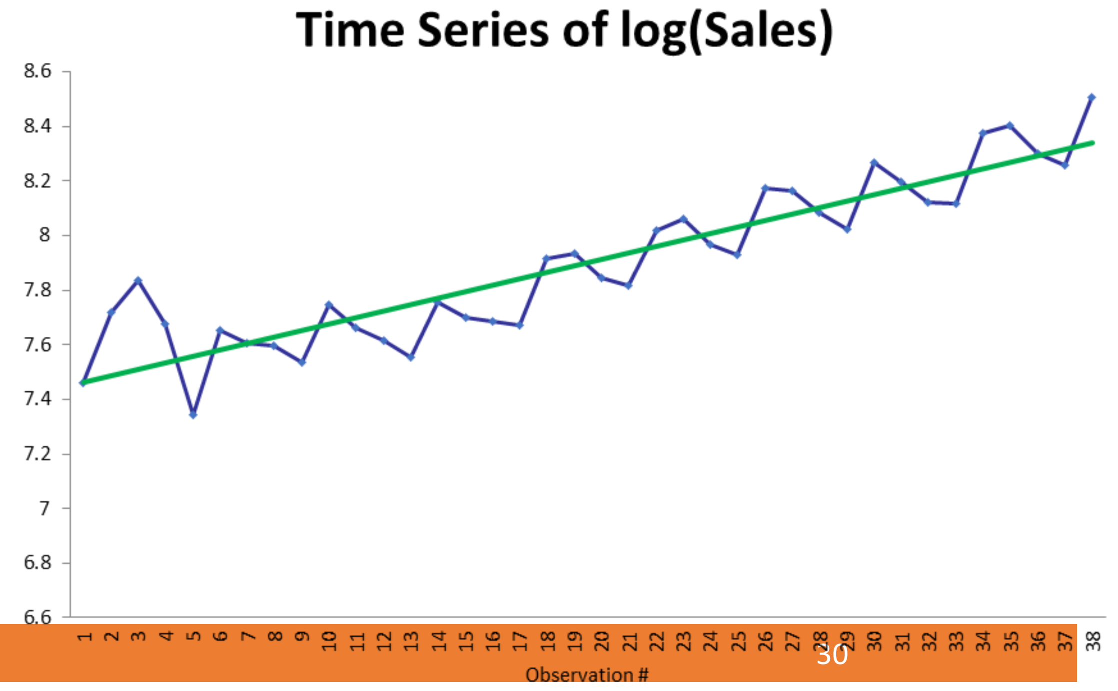
Multiplicative Model :Multiply all four components

$$Y_t = Exp(a)\ Exp(b_t\ t) Exp(b_1 S_1) Exp(\ b_2\ S_2) \ldots Exp(b_3 S_3) Exp(\varepsilon)$$

Use log transform of Y_t

$$\Rightarrow\ Log(Y_t)\ = a\ + b_t\ * t\ +\ b_1 * \ S_1\ +\ b_2\ *\ S_2 +\ \ \ldots . b_{M-1}\ *\ S_{M-1} +\ \varepsilon$$





Nonlinear Trend: Exponential Trend

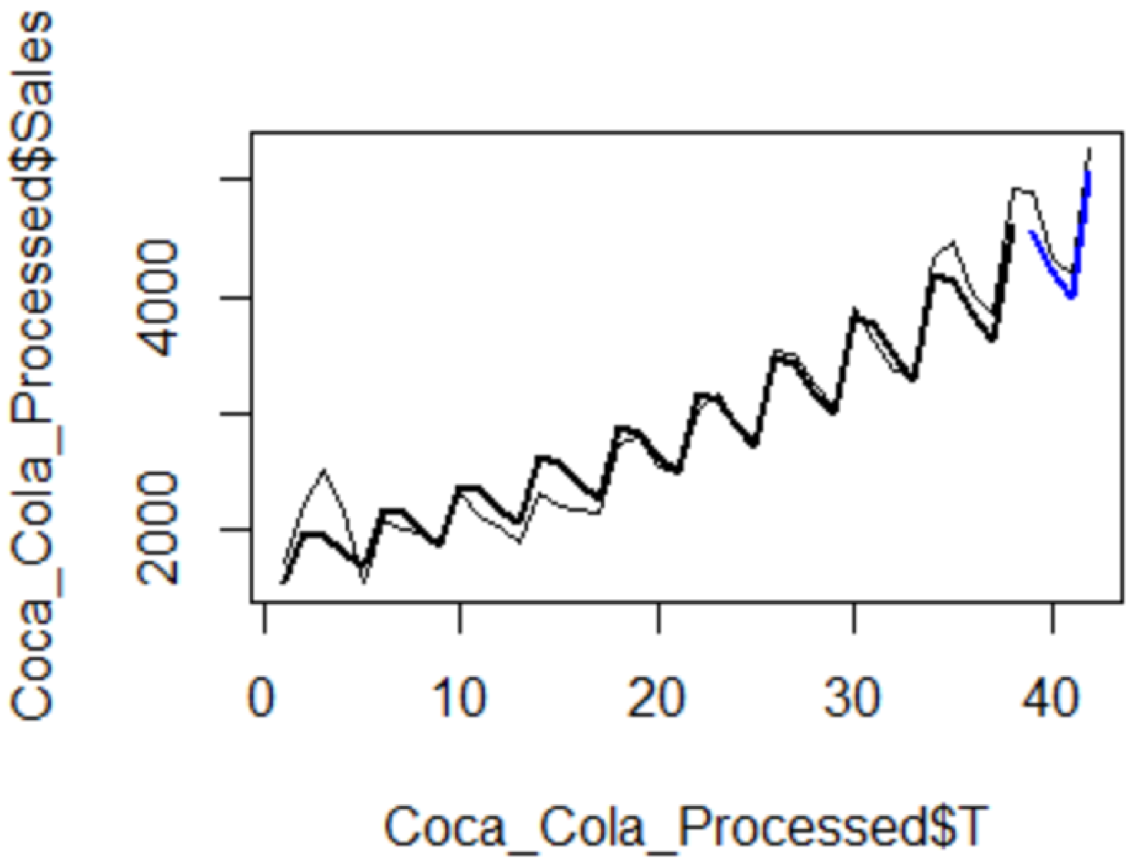
```
Call:
lm(formula = log(Sales) ~ T + factor(QuarterIndex), data = train)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.158334 -0.047680 -0.001473  0.021617  0.263198 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           7.322220   0.036379 201.274  < 2e-16 ***
T                     0.023603   0.001273  18.545  < 2e-16 ***
factor(QuarterIndex)2 0.218458   0.038457   5.681 2.47e-06 ***
factor(QuarterIndex)3 0.181250   0.039490   4.590 6.14e-05 ***
factor(QuarterIndex)4 0.082226   0.039510   2.081   0.0453 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08595 on 33 degrees of freedom
Multiple R-squared:  0.9216,    Adjusted R-squared:  0.9121 
F-statistic: 96.92 on 4 and 33 DF,  p-value: < 2.2e-16
```

Nonlinear Trend: Exponential Trend



Linear Model	
RMSE	464.98
MAPE	8.92%

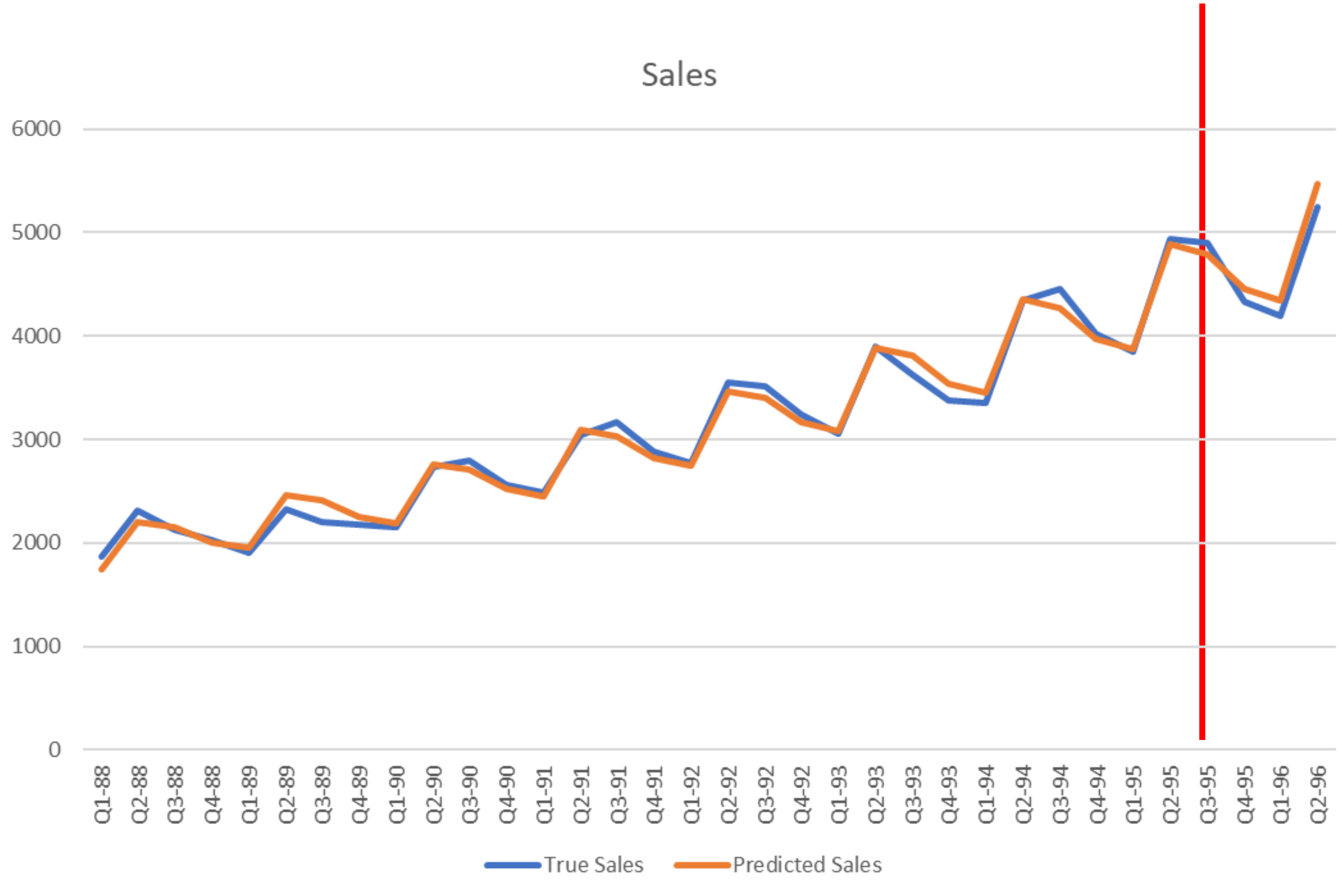
Quadratic Model	
RMSE	301.74
MAPE	5.76%

Exponential Model	
RMSE	225.52
MAPE	4.47%

Training Data

- Would more training data result in better predictions?
  - Is it necessary to use the full data (data from Q1-86)?
  - How much data would be appropriate?

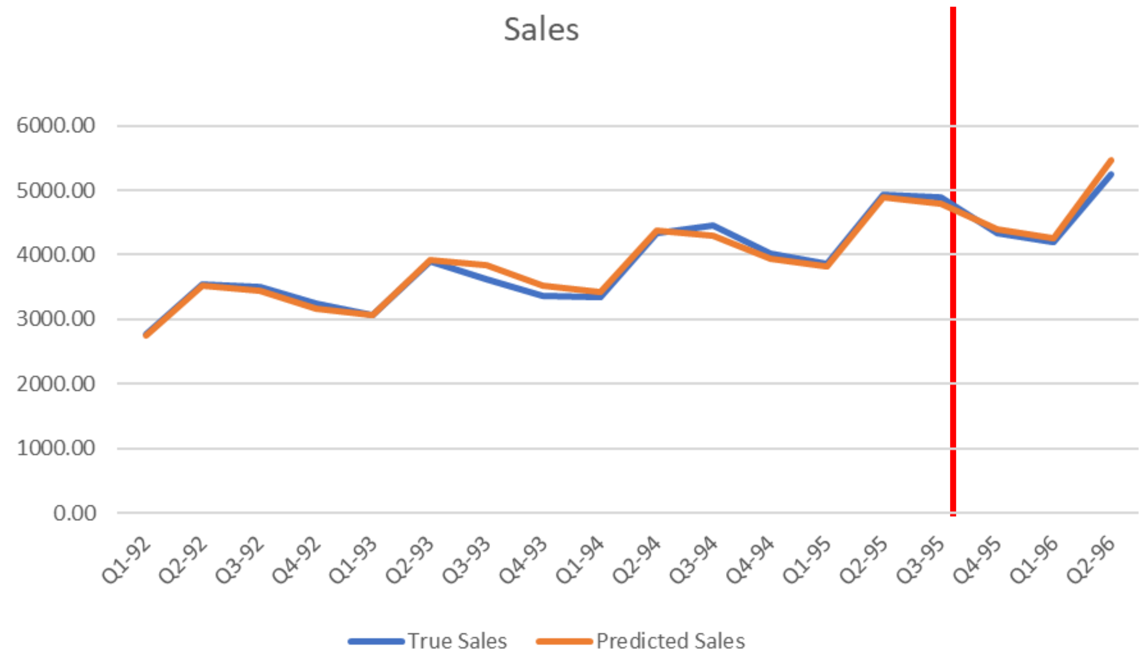
Exponential Model with Training Data from Q1-88 to Q2-95



Exponential Model with full data	
RMSE	225.52
MAPE	4.47%

Exponential Model with 88-95	
RMSE	154.68
MAPE	3.16%

Exponential Model with Training Data from Q1-92 to Q2-95



Exponential Model with full data	
RMSE	225.52
MAPE	4.47%

Exponential Model with 88-95	
RMSE	154.68
MAPE	3.16%

Exponential Model with 92-95	
RMSE	127.56
MAPE	2.32%

Is this model really better than the previous one?

- It might overfit the data.

Training Data

- Additional data does not always improve the predictions because the data might be outdated
- Need to make sure that the training data is sufficient enough to avoid the overfitting issue.
- How to detect overfitting? – challenging task!
  - In cross sectional data, we can use a cross validation – construct multiple pairs of training and validation data
  - It does not work in times series – time series data is ordered and should not be shuffled randomly!
  - Then, what to do – estimate the range of errors and the model error should not be too much below a certain level
  - Use some domain knowledge

Pros and Cons of Regression

- Interpretable
  - Each coefficient corresponds to one component
  - Easy to generate forecasts into the future
- Flexible
  - Easily incorporate external factors (other than time and seasonal factors) into models

$$Y_t = a + b_T * T + b_1 * S_1 + b_2 * S_2 + \dots . b_{M-1} * S_{M-1} + c_1 E_1 + \cdots + c_N E_N + \varepsilon$$

    - Temperature, precipitation, repackaging, introducing a new product
- Inflexible
  - **Stationarity assumption:** Assumes that mean, trend, and seasonality are all constant over time
  - **Static model**: Doesn't allow for changes over time

Next …

- Exponential Smoothing Model