Regression-Based Forecasting

Midterm Score Summary

- 49 questions ; 60 points

	Minimum	25th Percentile	Mean	75th Percentile	Maximum
Out of 60	28	38	42	45	50
Out of 100	47	63	70	75	83

Grading policy

- Grade assignment is based on relative performance
- If the highest aggregate score in the class is 80% (say), they will be treated as the top of the class to receive the highest grade.
- Top x% will get A, the second top y % will get A- and so on ……
- This grading process is equivalent to curving
- The grade you receive is based on **relative** performance not absolute
- I do **not** enforce any absolute percentages as cutoffs for final grades

Course Assessment

Type	**Weight**
Homework's (four)	20%
Midterm Quiz 1	20%
Midterm Quiz 2	20%
Project (Report + Presentation)	30 + 10%
	100%

Final Project (40%)

- Specify a business problem
- Identify a relevant dataset
- Business context could be in any area or function
- Assessment
    - Report (30%) + Presentation (10%)
- Presentation
    - 15-minute presentation on one of the classes of last week
    - **Presentation date(s) i**n the syllabus file

Final Report

- Formal report
  - Introduction, Problem description, Approach (Regression / Classification)
  - Data Analysis, Results, Inference
  - Conclusions, recommendations
- Regression: *k*-NN as Regression, Linear Regression & Regression Tree
- Classification: *k*-NN as classification, Logistic Regression & Classification Tree
- Assess the performance & recommend the best predictive model
- 8-10 pages including any tables and graphs (excluding code)
- Two or Three key insights from the entire analysis
- Submit the code with comments at end of the report

Public datasets for final project

kaggle

- https://www.kaggle.com/
- Online community of data scientists and machine learners
- Owned by Google Inc.
- Register yourself, and you can download datasets for free
- As of June 2017, Kaggle passed over 1,000,000 registered users
- Variety of datasets
- Your imagination only limits possibilities

Final Project presentation

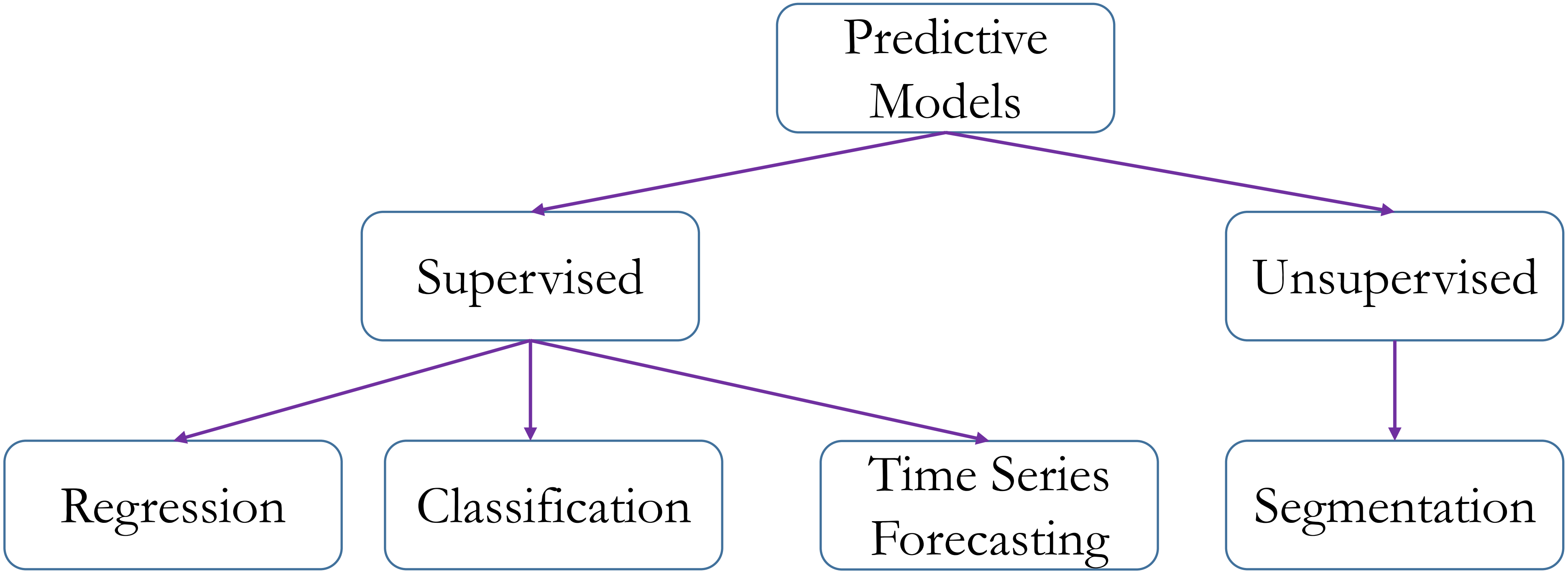
- Presentation (10%)
  - 15-minute presentation followed by a 10-minute Q&A
  - **May 31st (Tue) & Jun 02nd (Thu)**
  - Groups are randomly assigned to the 2 days
  - Groups should send the ppt file by 8 am on their presentation date
  - Each member of the group should **mention the contribution** of their work in the last slide of the presentation file
- **Everyone** must be present in the class on the presentation days
  - Zero scores for presentation assessment if absent

May 31st presentations

- ACB
- ATJ
- HJJ
- P

Jun 02nd presentations

- AJA
- DJK
- MRV
- TAP



Predictive Models

Supervised

Unsupervised

Regression

Classification

Time Series Forecasting

Segmentation

- ***k*-Nearest Neighbor**
- **Linear Regression**
- **Regression Trees**
- Neural Networks
- Ensembles
- ……

- ***k*-Nearest Neighbor**
- Naïve Bayes
- **Logistic Regression**
- **Classification Trees**
- Neural Networks
- Discriminant Analysis
- Ensembles
- ……

- **Regression-based**
- Smoothing methods
- ……

- **Clustering**
- ……

Time Series Forecasting

- Focus
  - Forecasting future values of a single time series
- Performed in nearly every organization that works with quantifiable data
- Applications:
  - Sales forecast in Retail stores
  - Forecast reserves, production, demand and prices in Energy companies
  - Forecast enrollment in educational institutions
  - Forecast tax receipts and spending in government
  - Inflation and Economic activity in World Bank, IMF

Previous topics applications

- Time was not considered in significance in the previous datasets
- Most of the datasets we studied in the previous topics are called cross-sectional data
- Here we study – time series data
- Today's technology has helped to record on very high frequent time scales
- An example from one of my research topic – Alibaba data

Time Series components

- Four components in time series
  - Level - Average level of the series
  - Trend – Change in series from one period to the next
  - Seasonality – Short-term cyclical behavior of the series
  - Noise – Random variation from other unknown causes
- Let's look at an example

Amtrak Ridership

- Monthly ridership

- January 1991 – March 2004

    - Period : January 1991 – March 2004

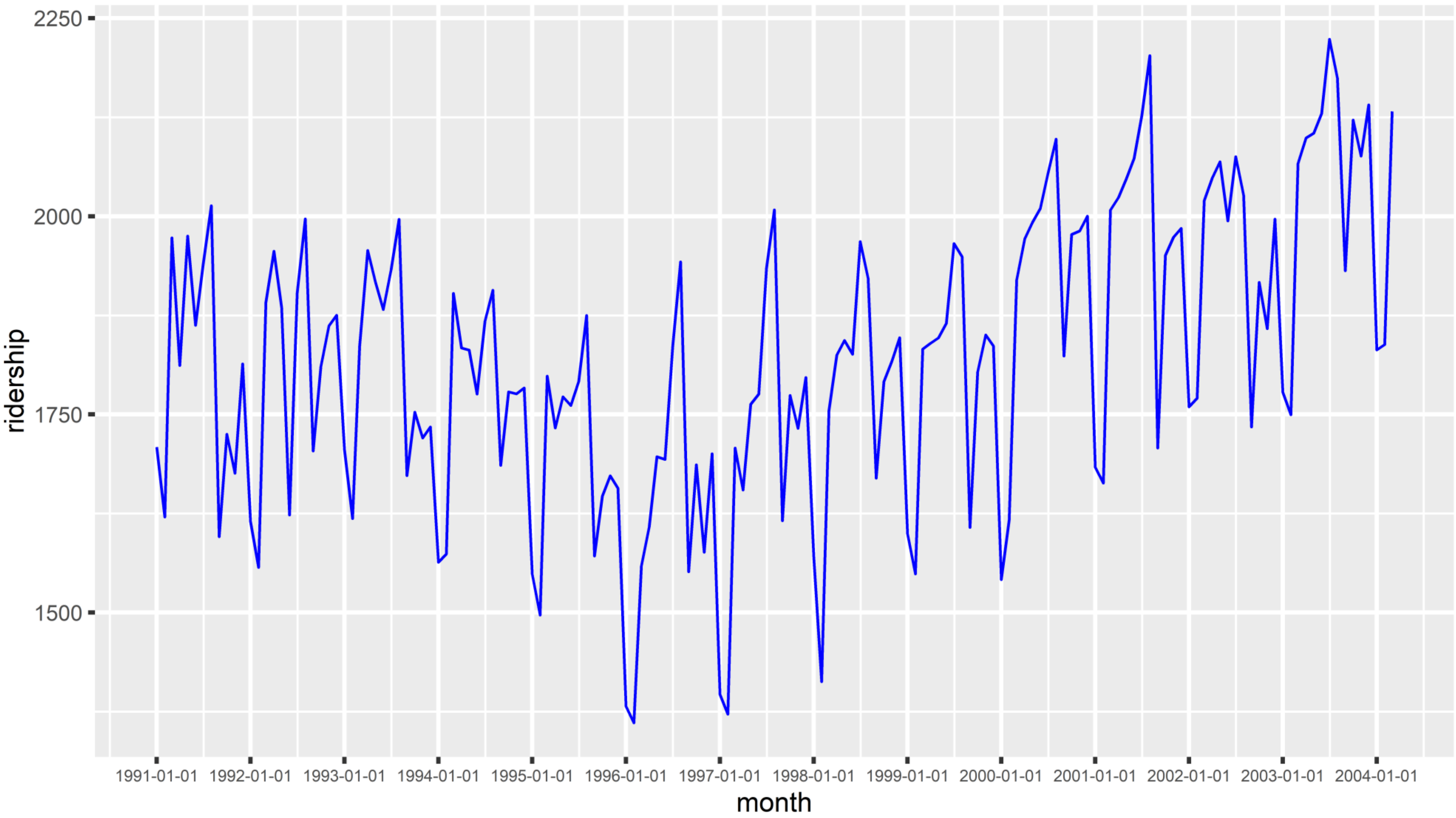
    - Ridership is in thousands

    - ~ 1,800,000 passengers per month

Today’s class mandatory steps

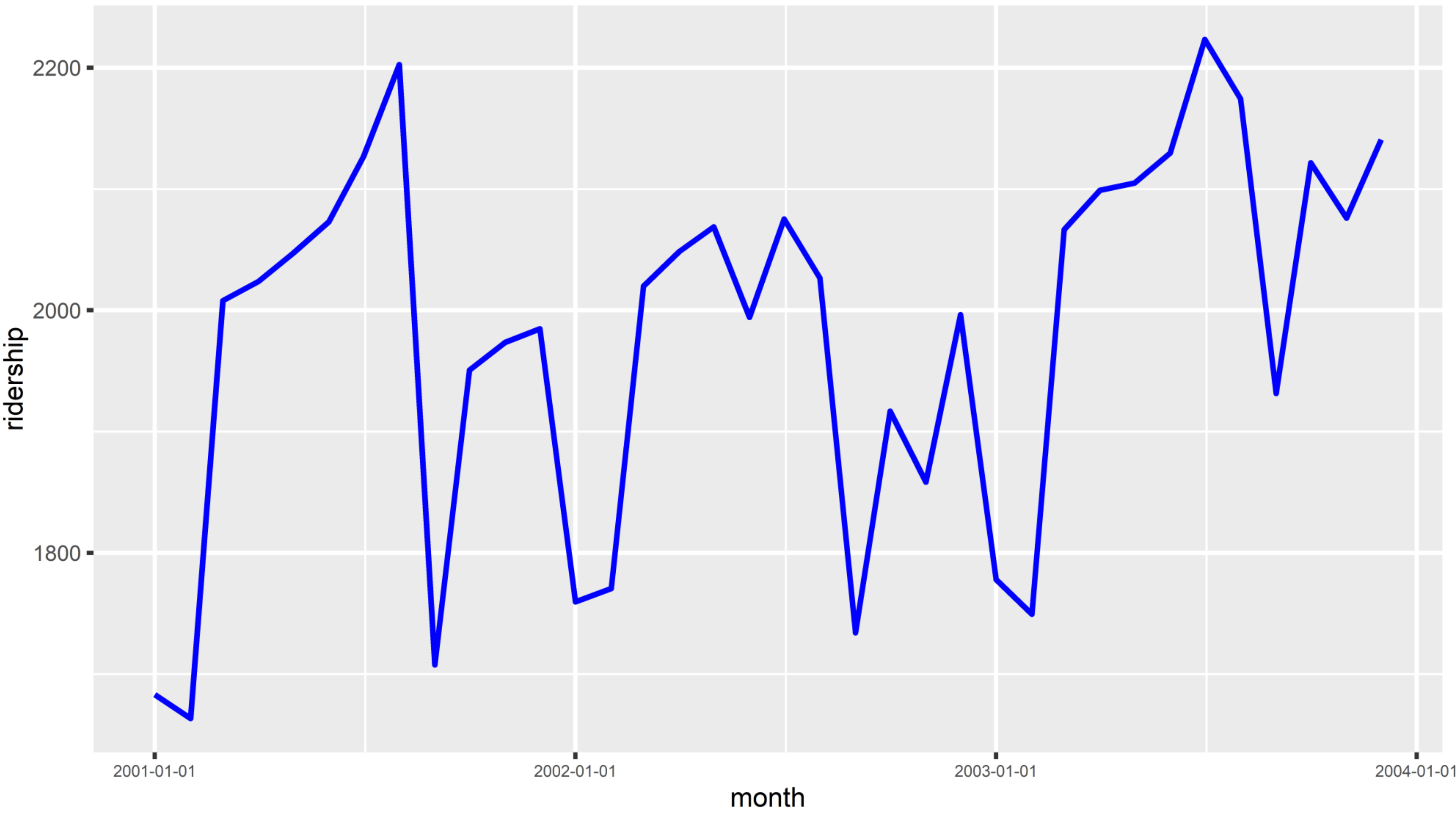
- Create a folder name “**m. regression_forecasting**” within the folder “**oba_455_555_ddpm_r/rproject**”
- Download “**regression_forecasting_code.R**”, and all **CSV** files from canvas
- Place all downloaded files in “**oba_455_555_ddpm_r /rproject/ m. regression_forecasting**”
- Open RStudio project
- Open “**regression_forecasting_code.R**” file within RStudio

Amtrak Ridership



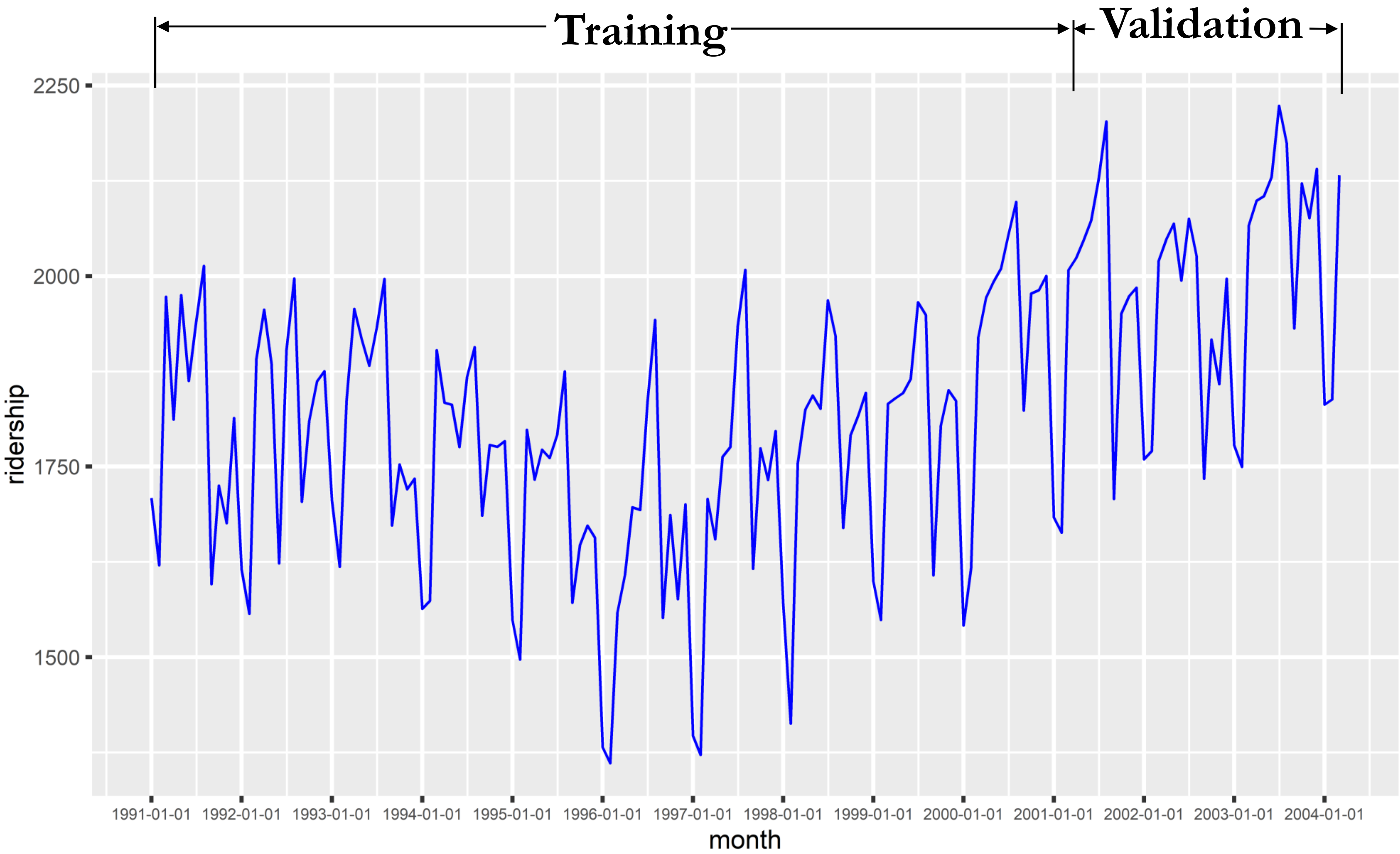
- Slight U-shaped trend, Annual seasonality
- Peak travel during July and August

Zoom from 2001 to 2003



© Pradeep Pendem

Data Partition

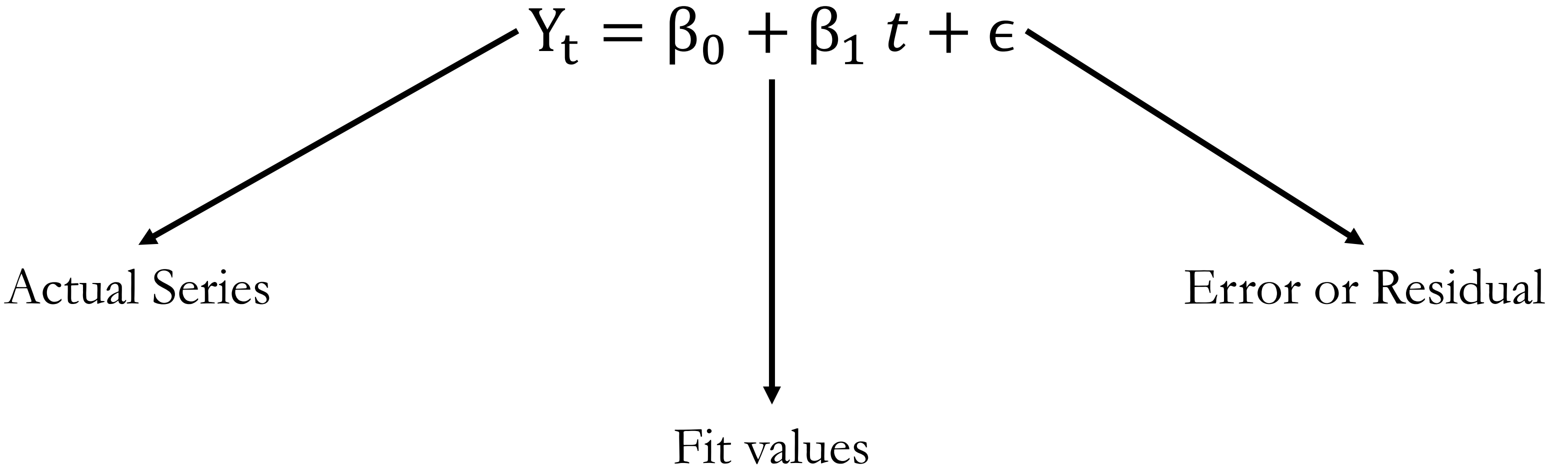


Trend models

- Commonly used trend models
  - ➢ Linear
  - ➢ Exponential
  - ➢ Polynomial

Linear trend

- The outcome variable Y is the time series
- Predictor X is the time index



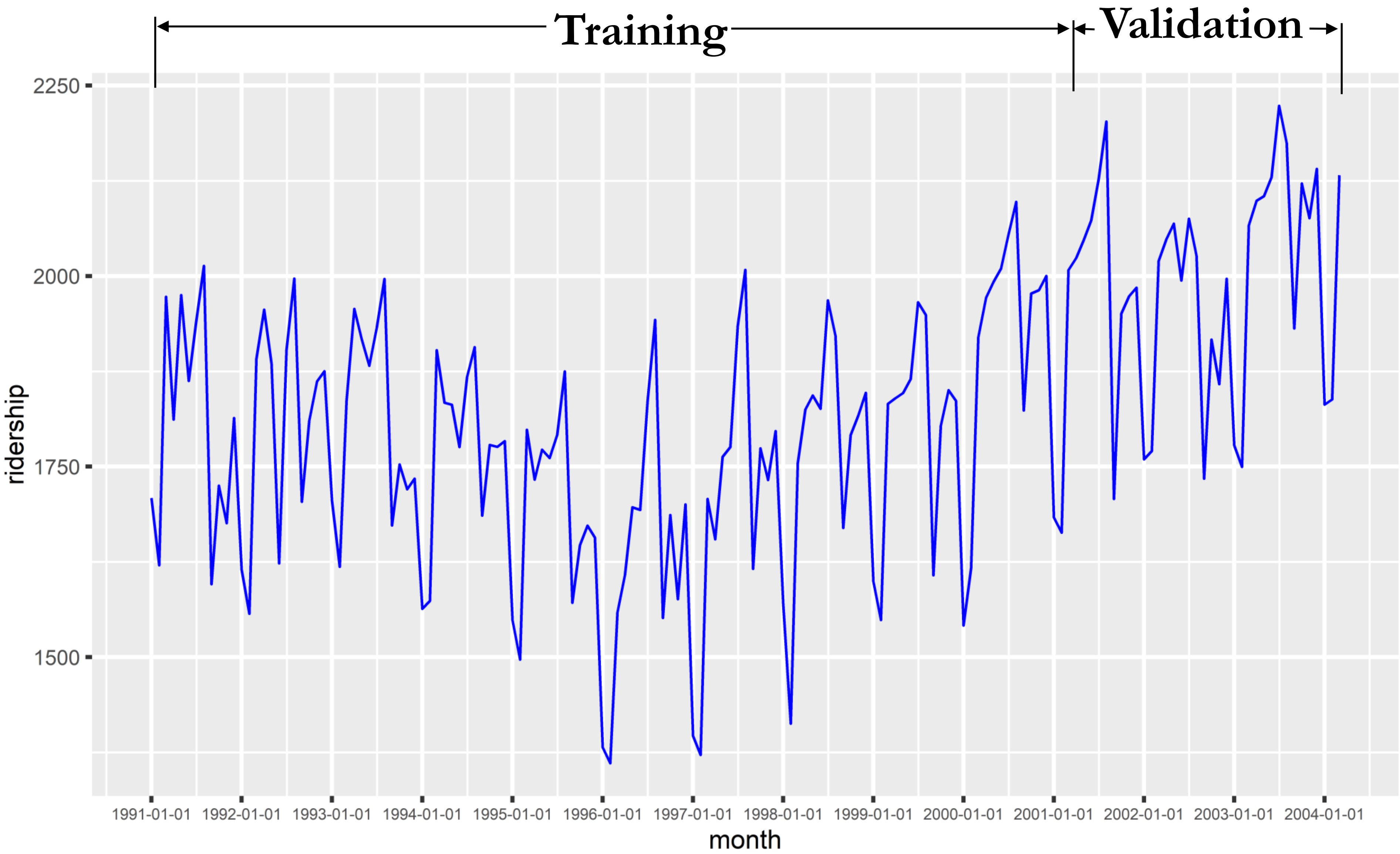
$$Y_t = \beta_0 + \beta_1\, t + \epsilon$$

Actual Series

Fit values

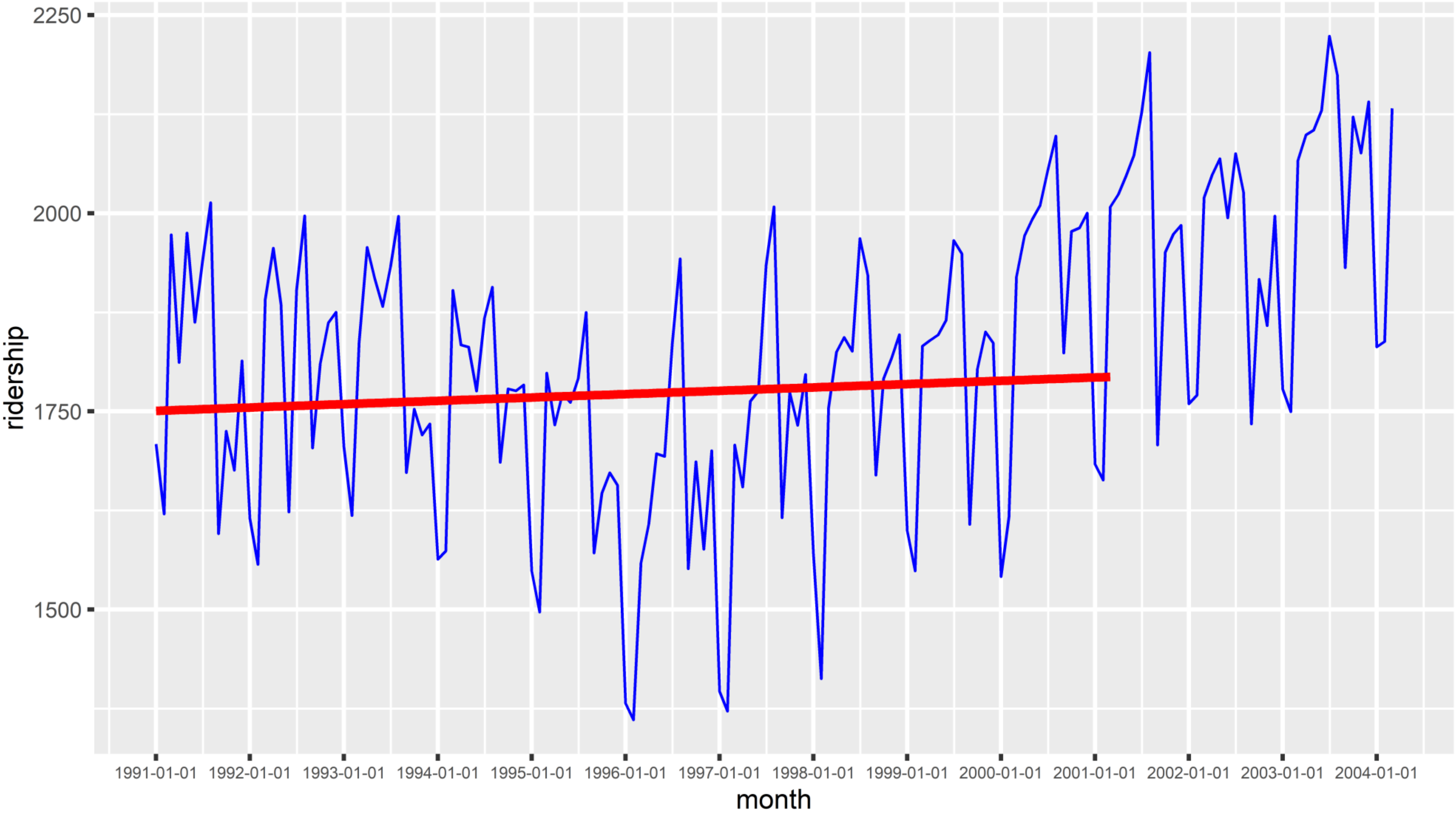
Error or Residual

Linear trend model on Training data

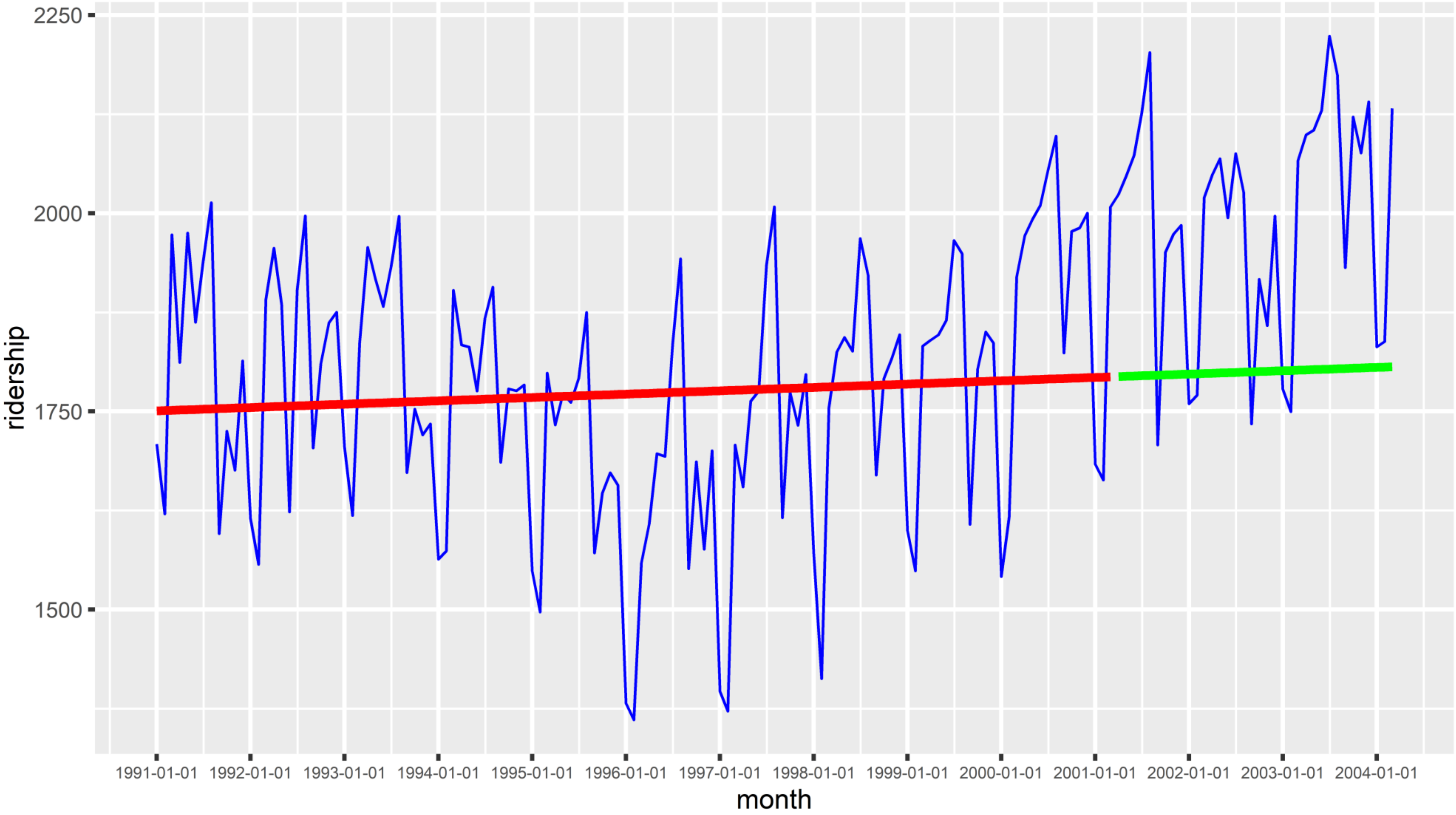


© Pradeep Pendem

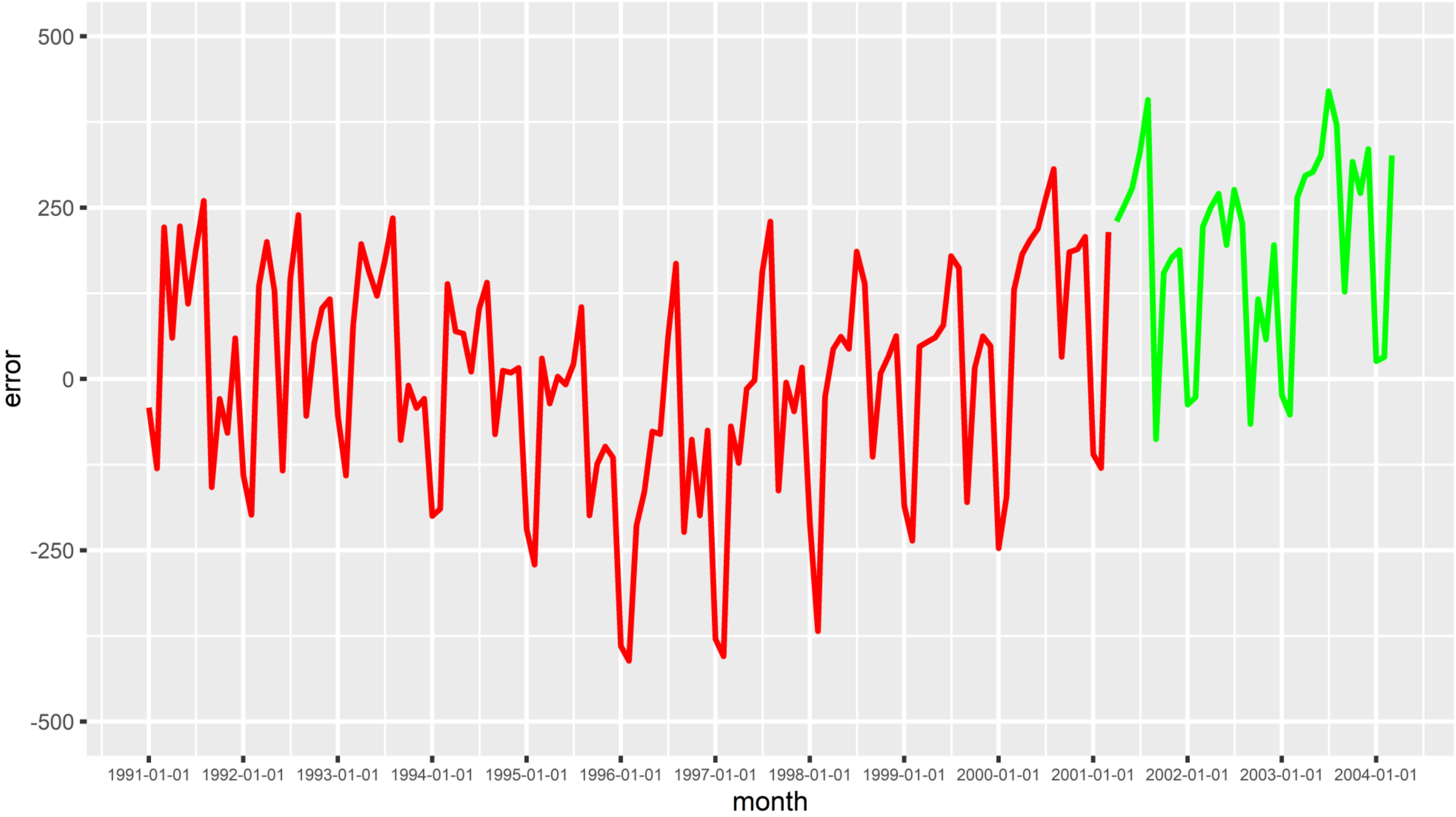
Linear trend – Fit



Linear trend – Fit and Prediction

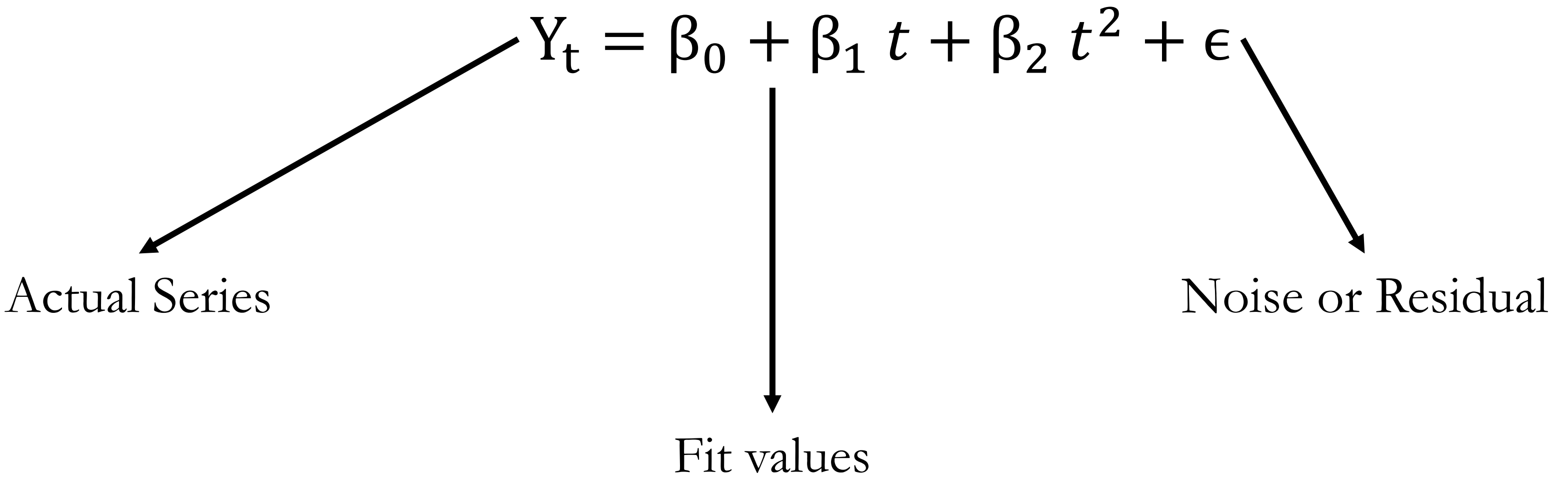


Linear trend – Error



© Pradeep Pendem

Polynomial trend



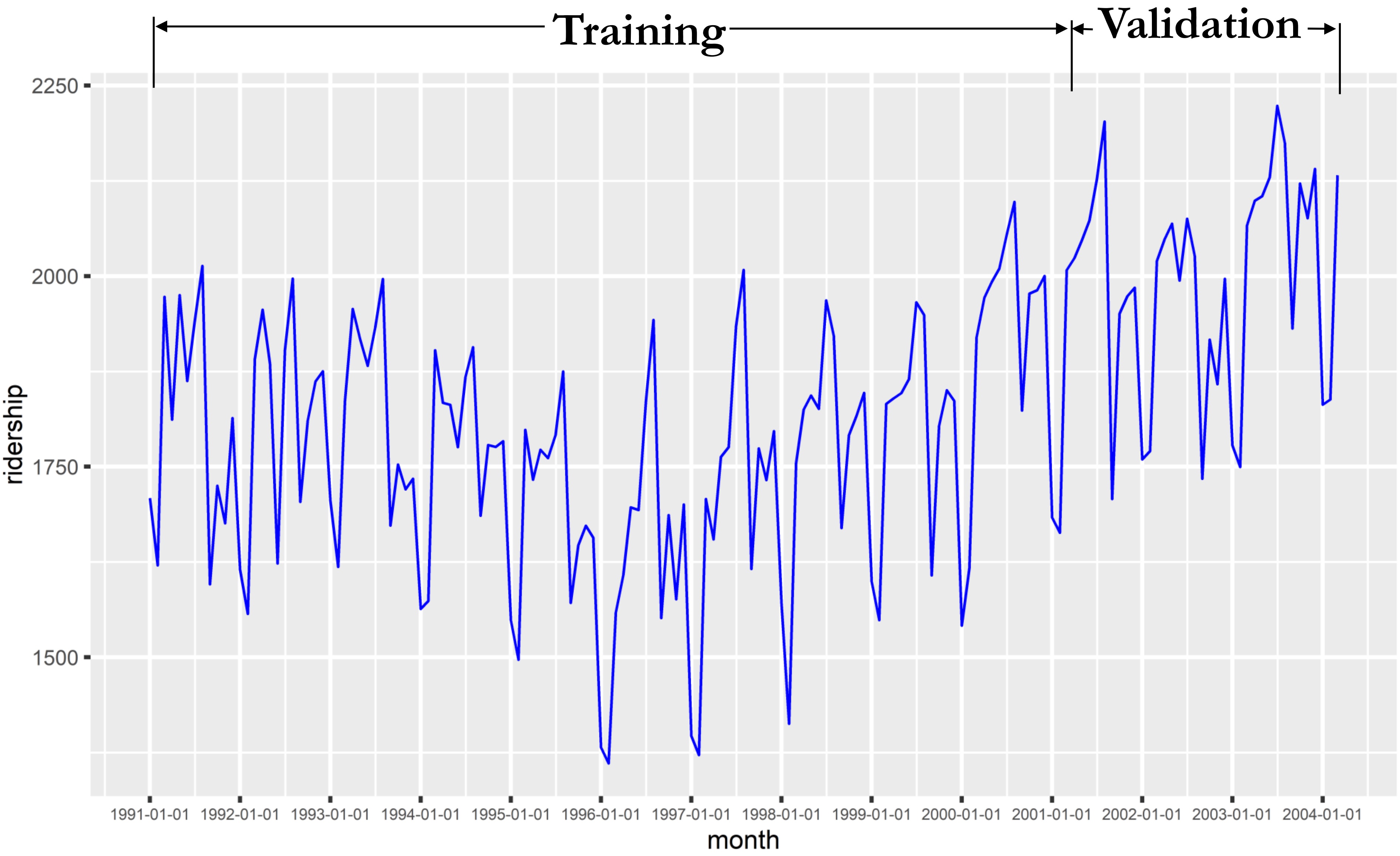
$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon$$

Actual Series

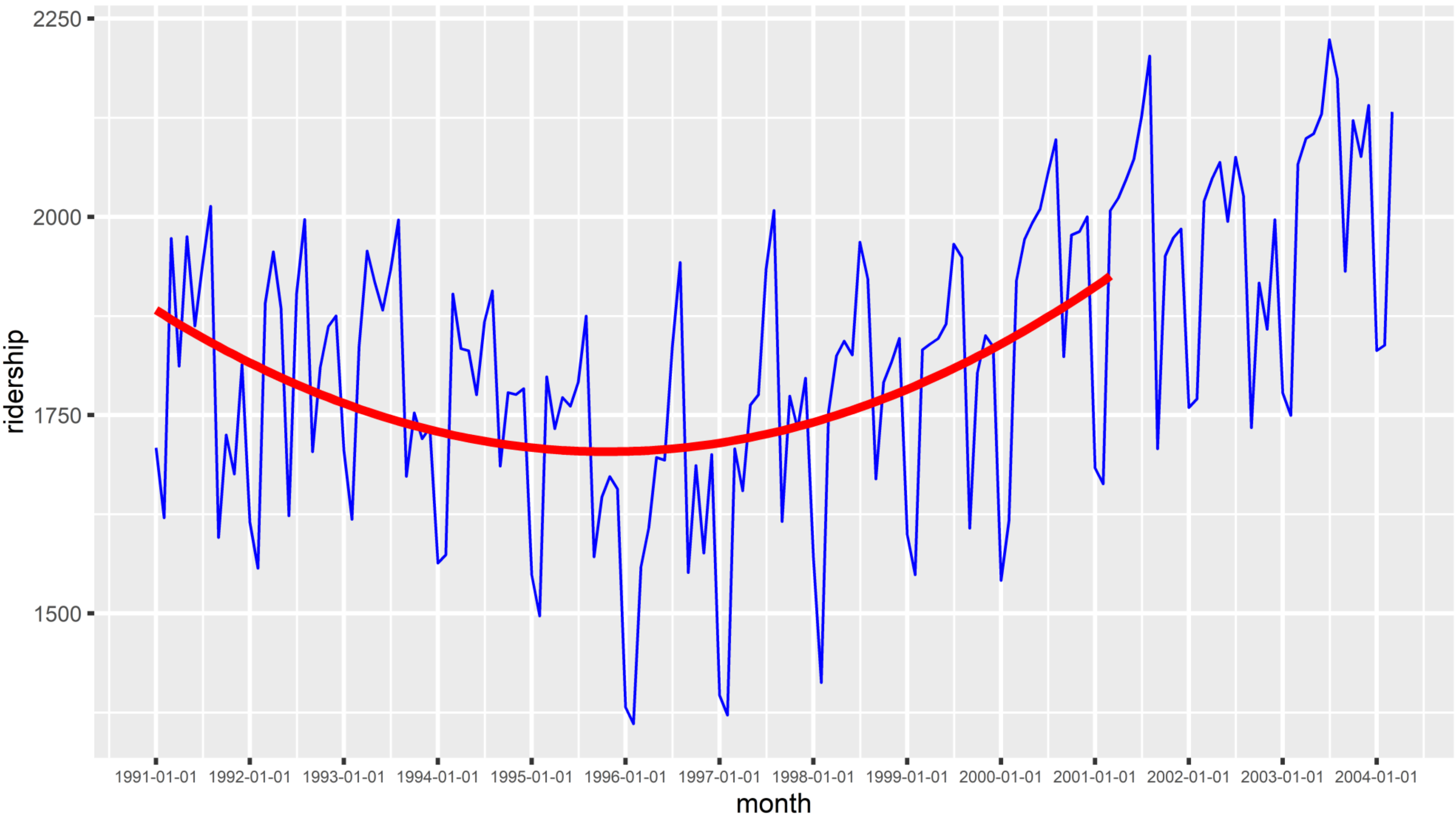
Fit values

Noise or Residual

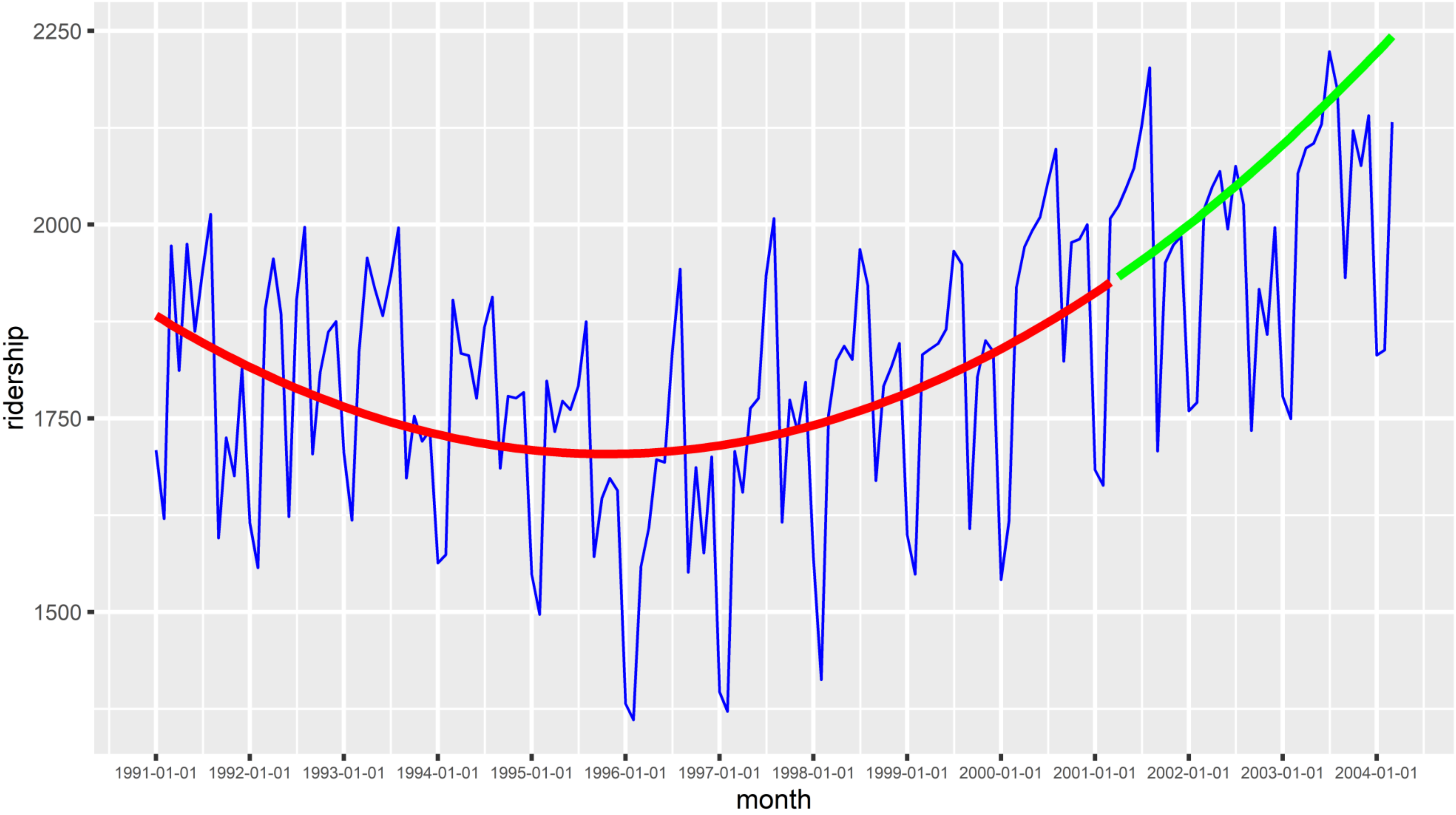
Polynomial trend model on Training data



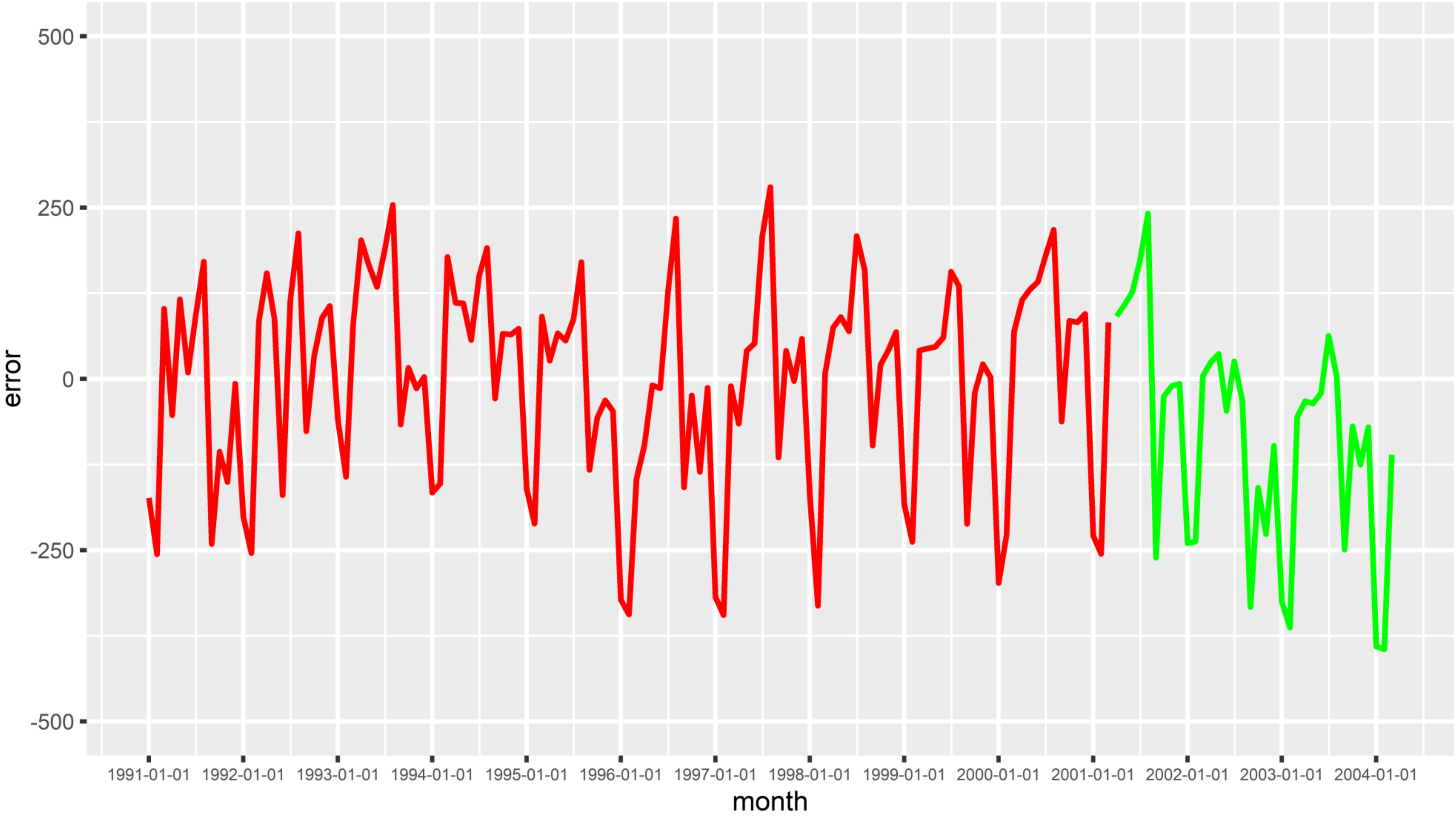
Polynomial trend – Fit



Polynomial trend – Fit and Prediction

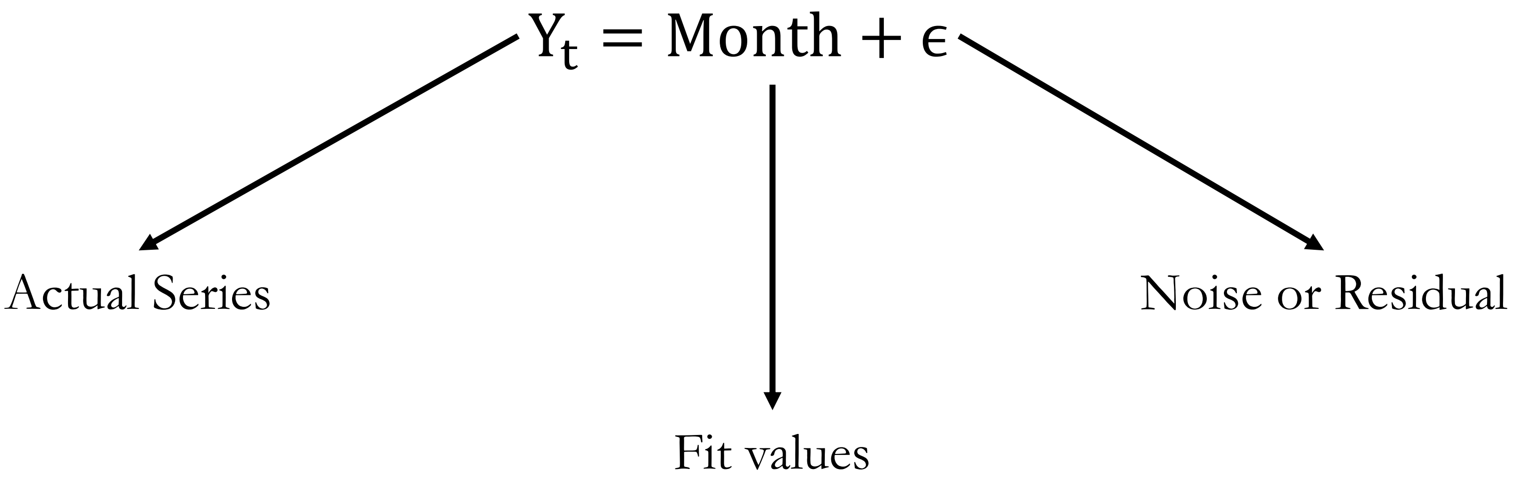


Polynomial trend – Error



© Pradeep Pendem

Seasonality



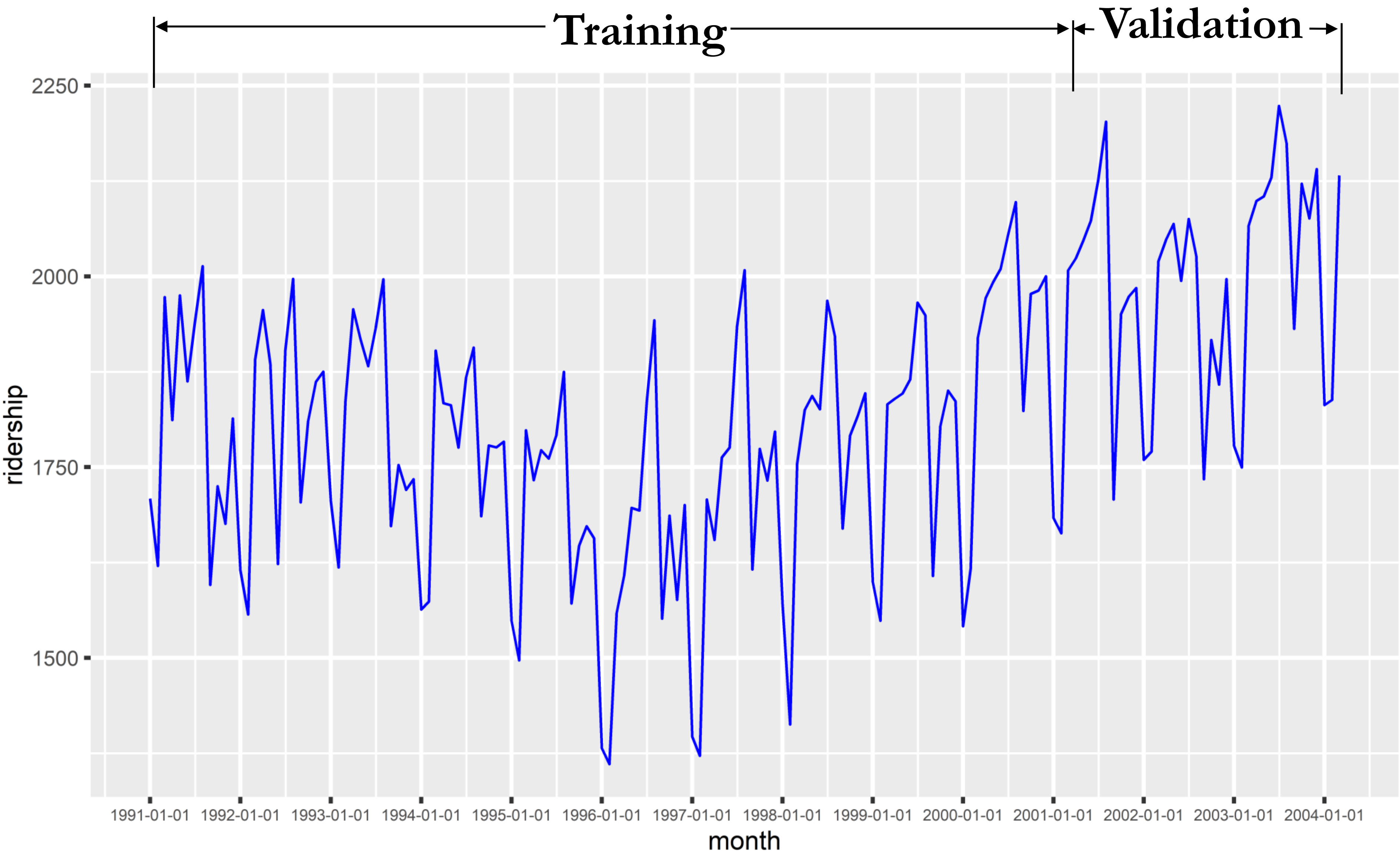
$$Y_t = \text{Month} + \epsilon$$

Actual Series

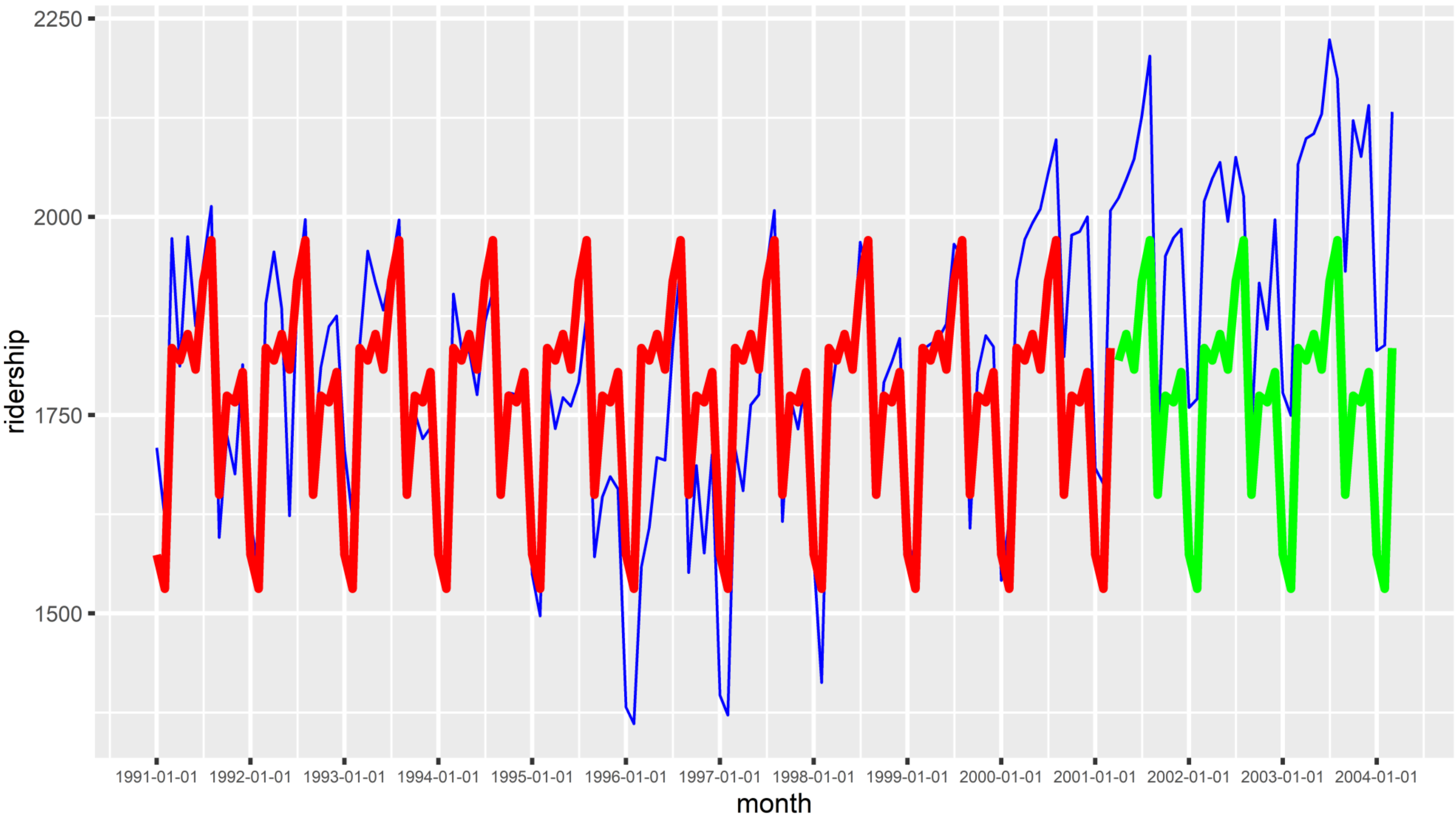
Fit values

Noise or Residual

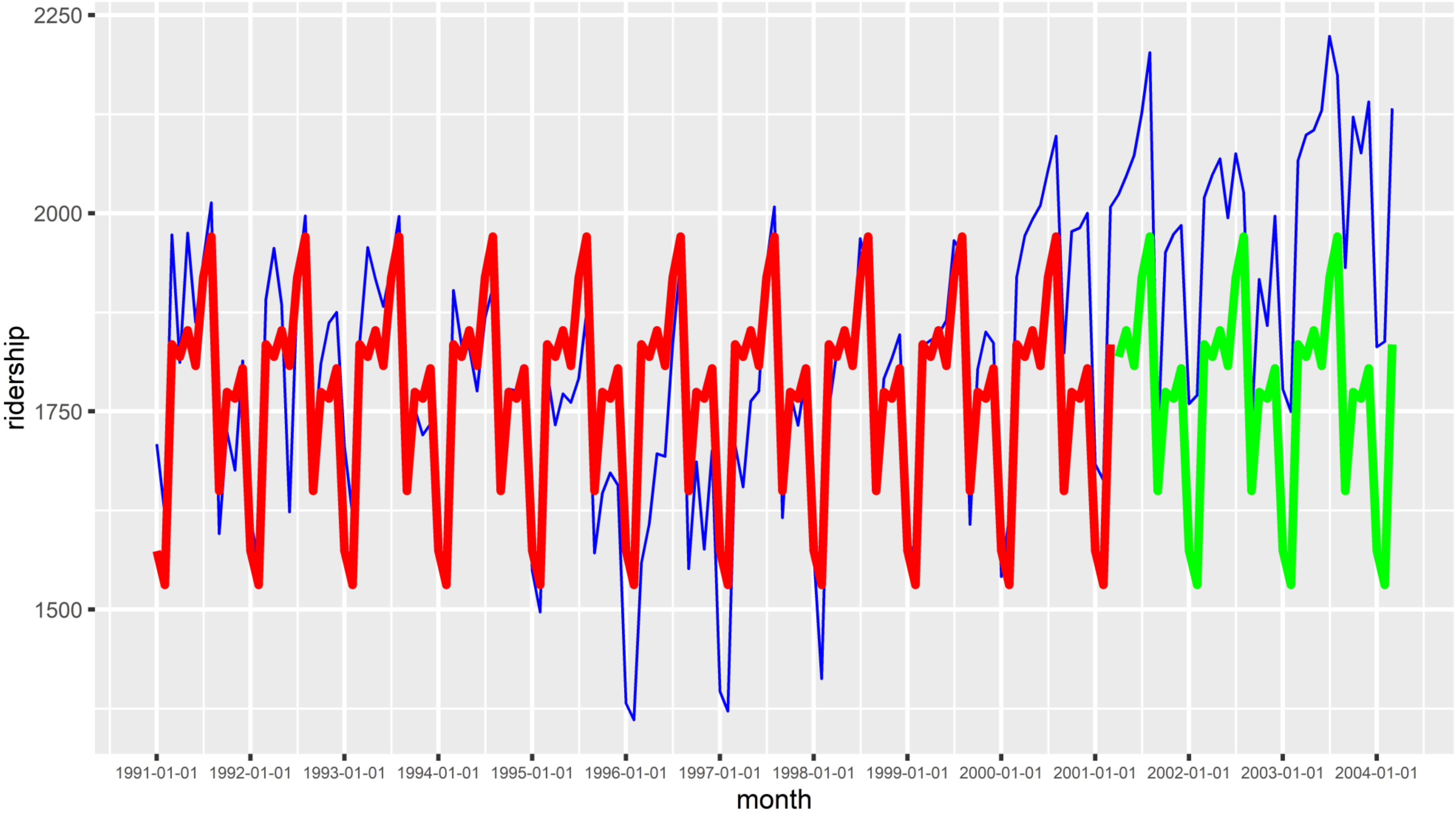
Seasonality model on Training data



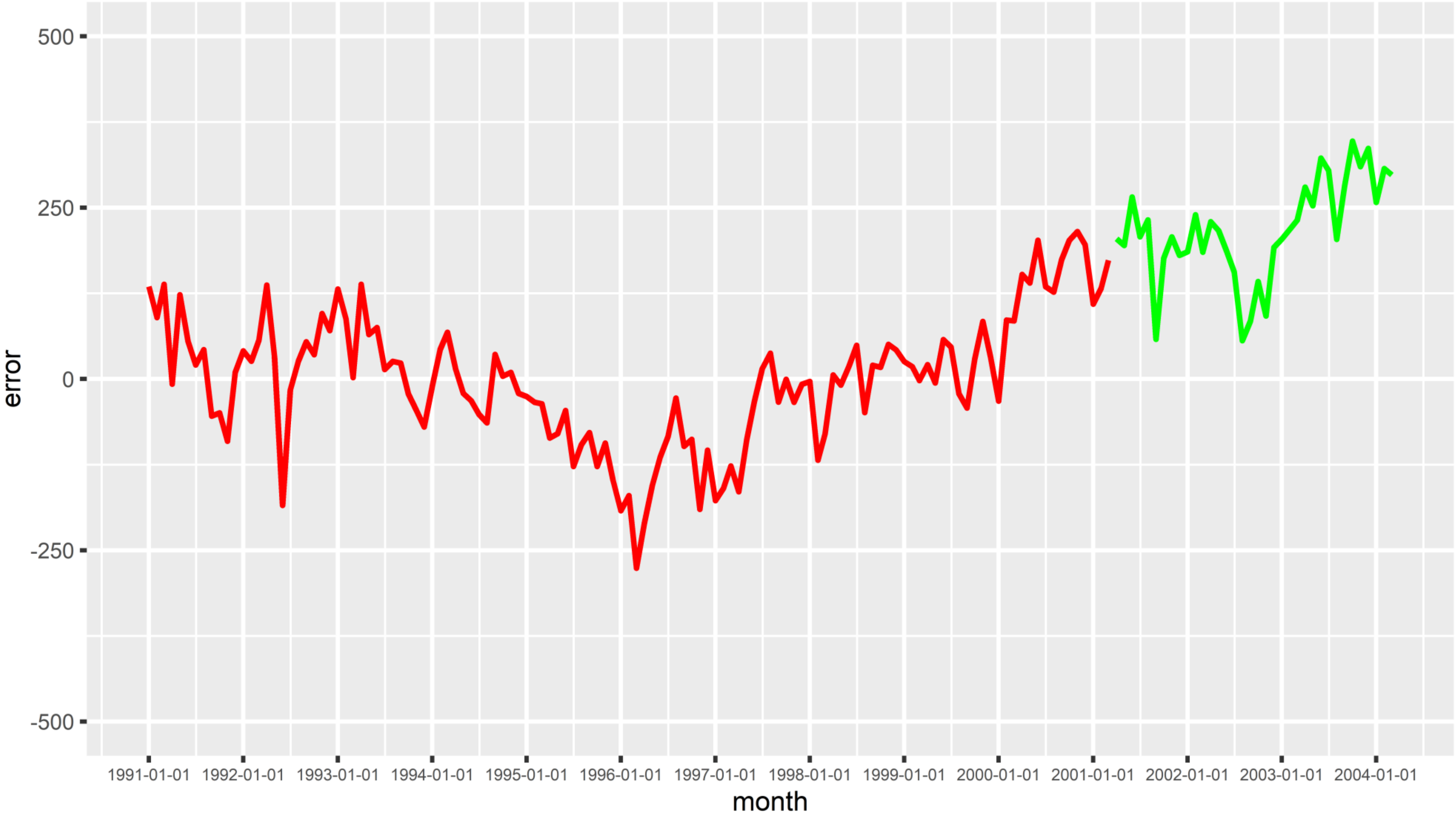
Seasonality– Fit



Seasonality– Fit and Prediction



Seasonality– Error



Polynomial trend and Seasonality

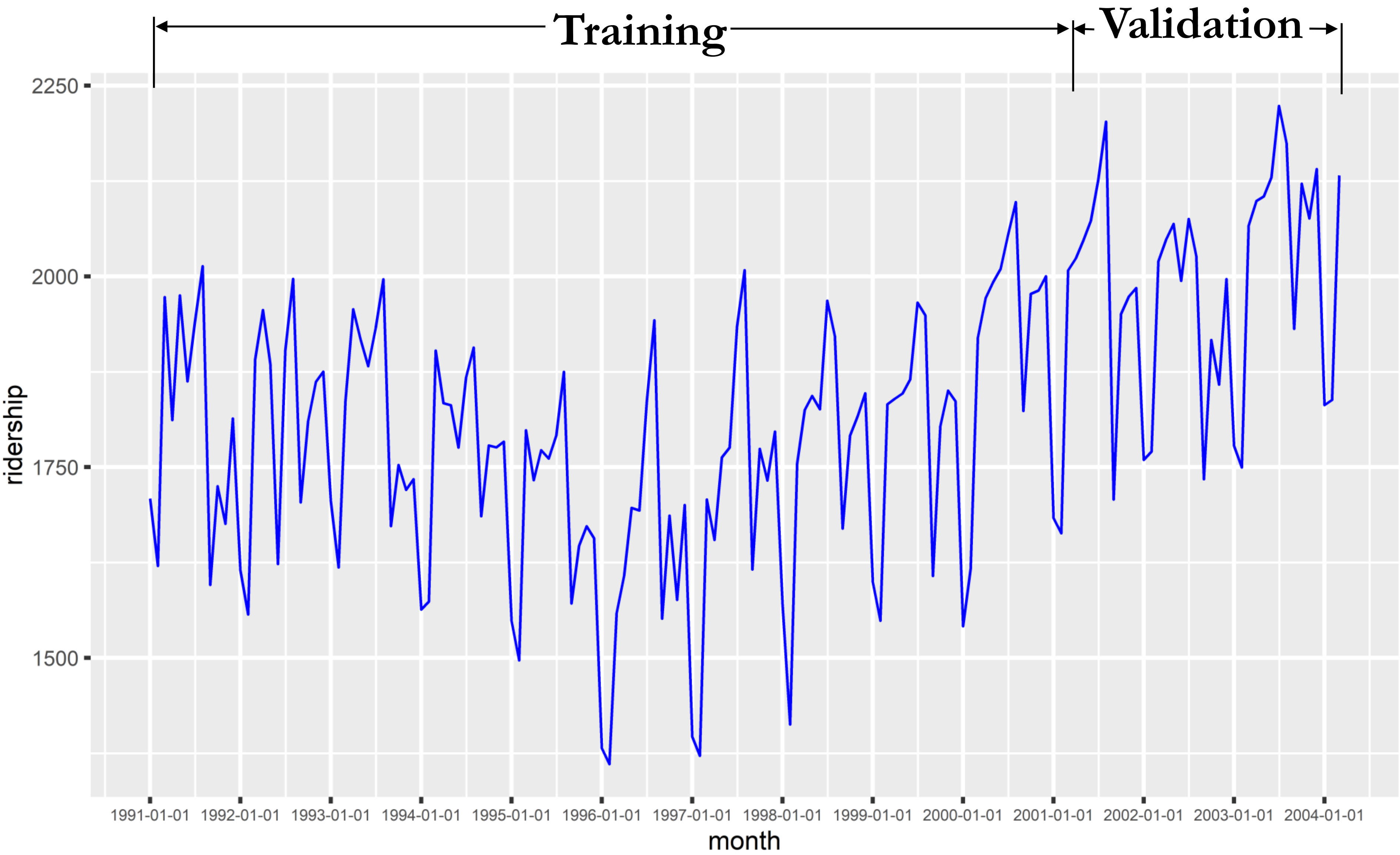
$$\mathrm{Y_t} = \beta_0 + \beta_1\, t + \beta_2\, t^2 + \mathrm{Month} + \epsilon$$

Actual Series

Noise or Residual

Fit values

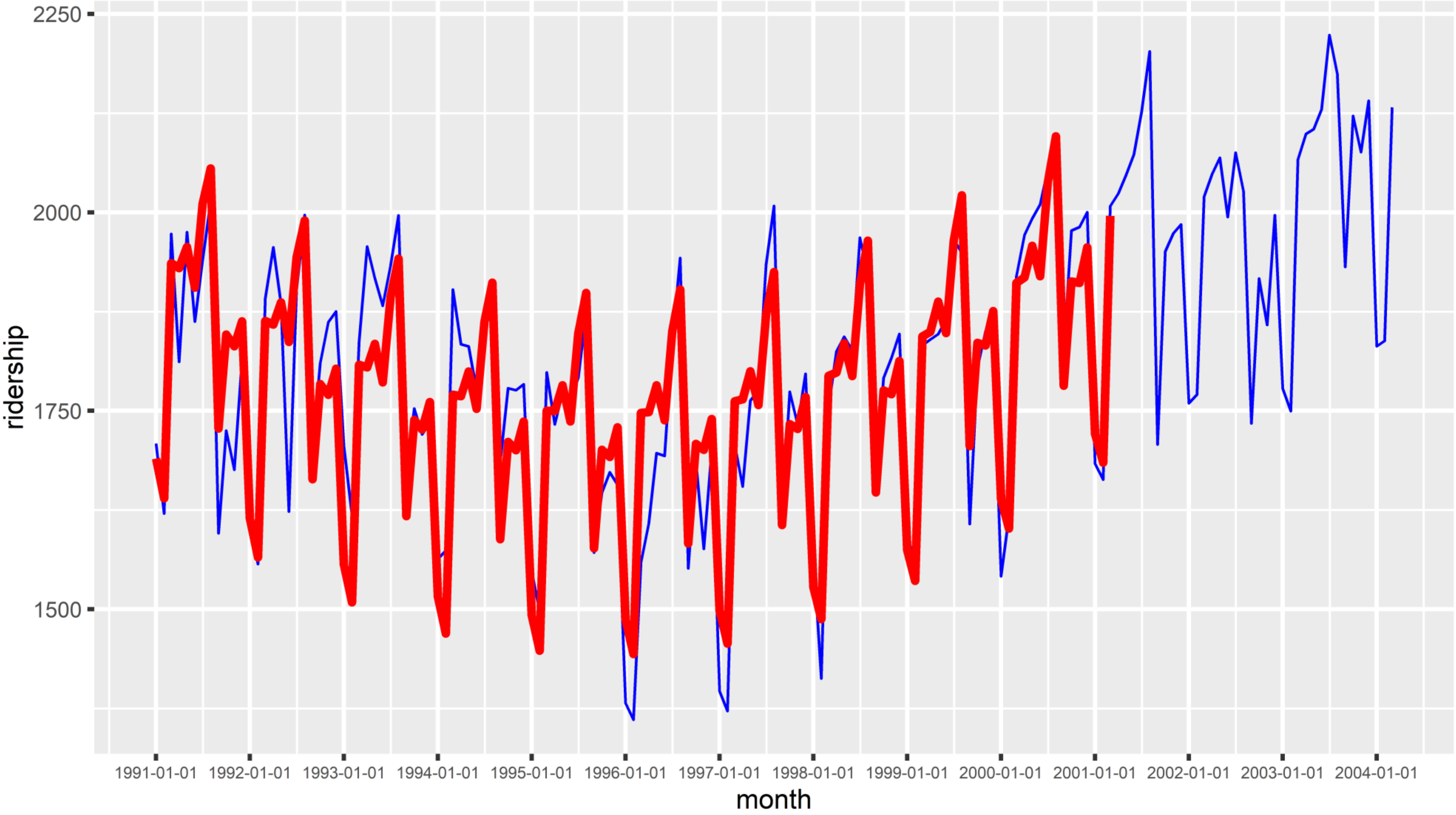
Polynomial + Seasonality model on Training data



Training

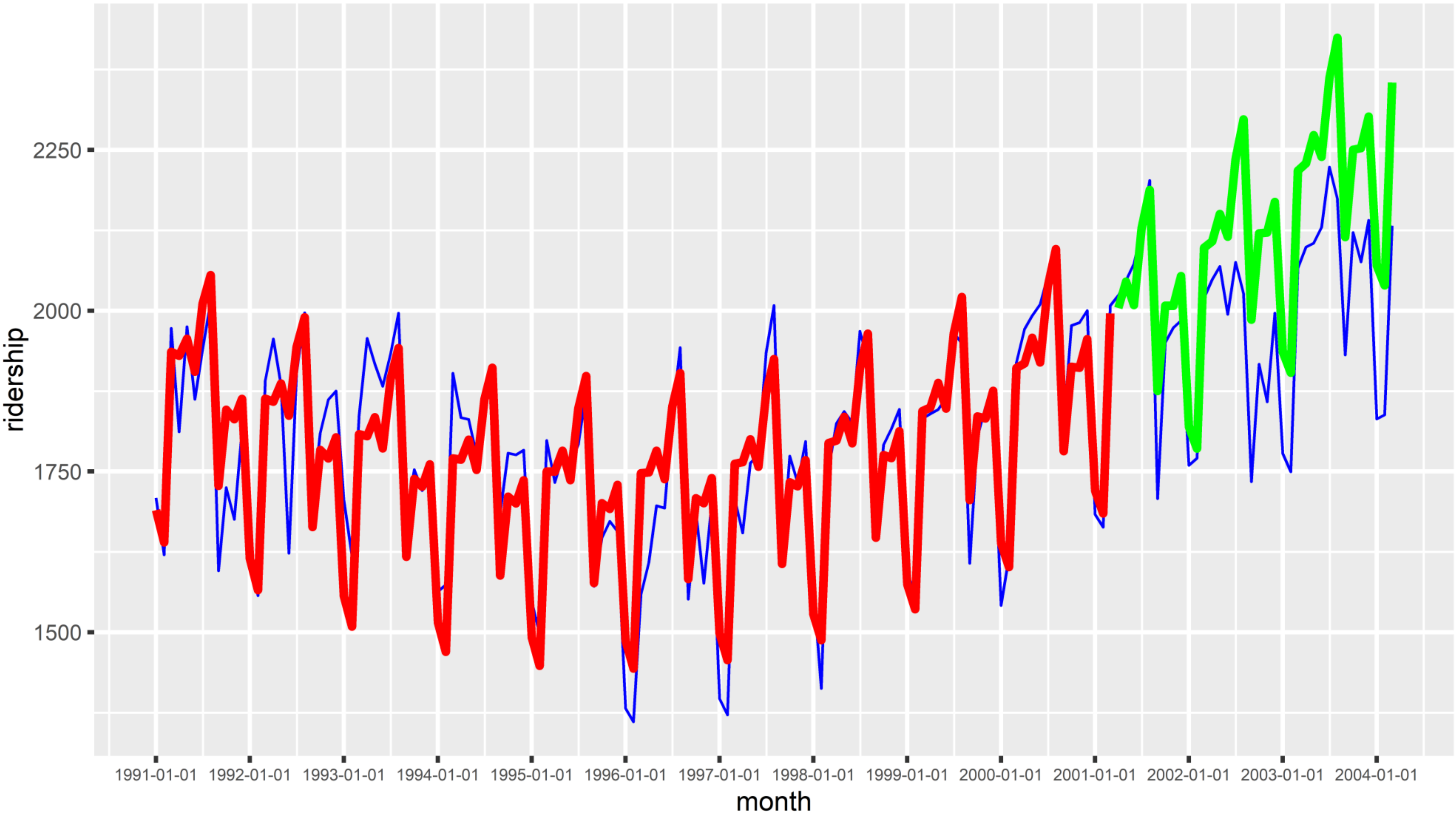
Validation

Polynomial + Seasonality– Fit

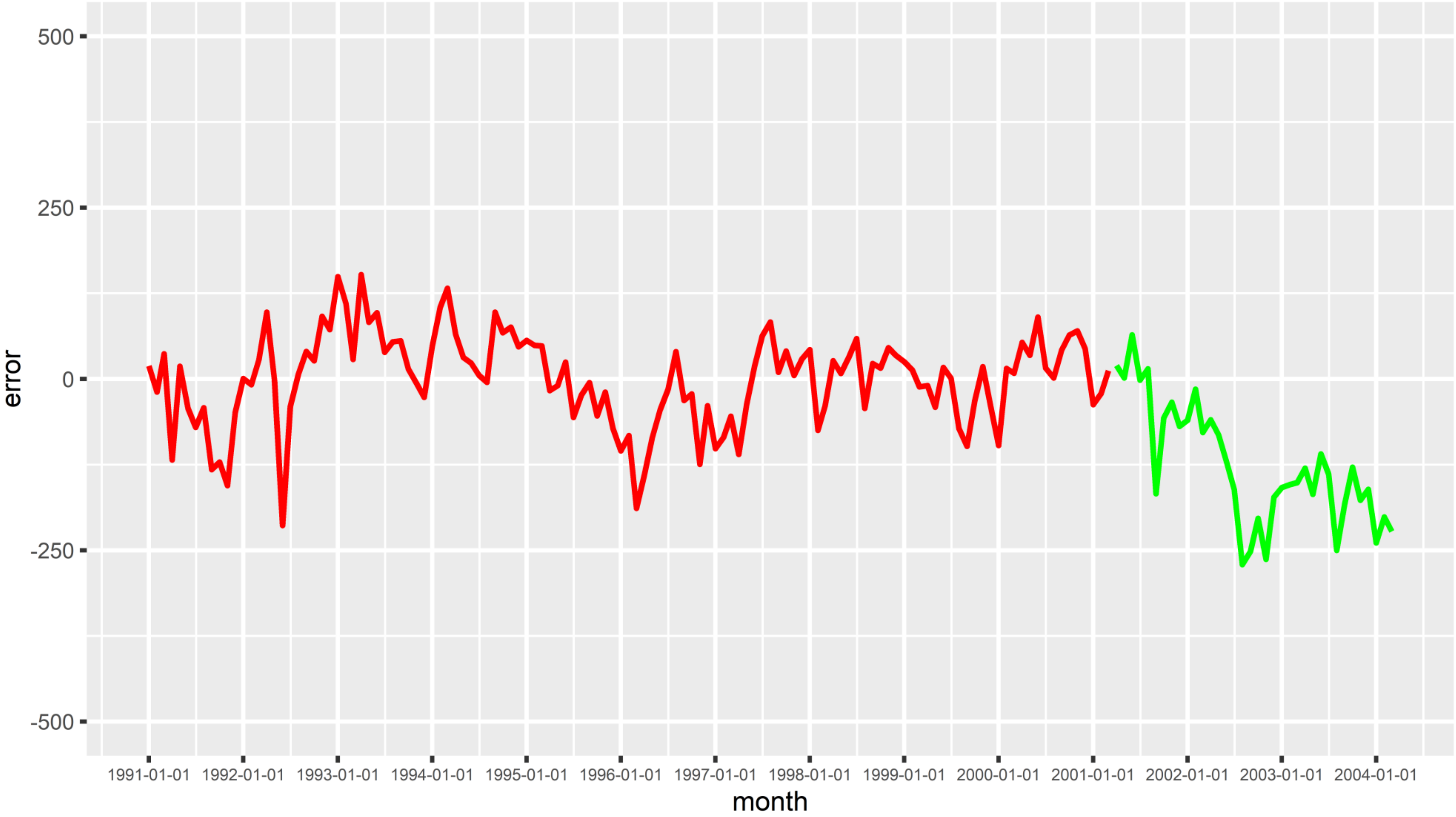


© Pradeep Pendem

Polynomial + Seasonality– Fit and Prediction



Polynomial + Seasonality – Error



RMSE – Training and Validation

Model	Train	Validation
Linear	158.92	239.48
Polynomial	146.97	179.84
Seasonality	96.34	229.65
Polynomial + Seasonality	66.76	153.25

MAPE – Training and Validation

Model	Train	Validation
Linear	7.53%	10.14%
Polynomial	7.01%	7.07%
Seasonality	4.32%	10.86%
Polynomial + Seasonality	3.01%	6.7%

Next class

- Classification Tree

Thank You