MScFE 600 – Financial Data
Group Work Project

WorldQuant University

Submitted by: Karan Donde

Submission Date: 13 April 2025

Table of Contents

1. Task 1 – Data Quality
   a. Example of Poor-Quality Structured Data
   b. Identifying Quality Issues in Structured Data
   c. Example of Poor-Quality Unstructured Data
   d. Identifying Quality Issues in Unstructured Data

2. Task 2 – Yield Curve Modeling
   a. Selection of Government Securities
   b. Yield Curve Data and Maturities
   c. Nelson-Siegel Model Fitting
   d. Cubic-Spline Model Fitting
   e. Model Comparison: Fit and Interpretation
   f. Parameter Specification
   g. Ethical Considerations of Smoothing

3. References

Task 1 – Data Quality

a. Example of Poor Quality Structured Data

Structured data is organized in rows and columns, often within databases or spreadsheets. Below is an example of structured financial data that demonstrates poor quality, typically found in transactional or reporting environments:

Date	Stock Ticker	Open	Close	Volume
10/2/2023	AAPL	175.32	NULL	20,000,000
10/3/2023		176.2	177.45	abc
10/4/2023	AAPL	175.9	176.88	18,500,000

This dataset contains several quality issues:

- **Missing Data:** On 2023-10-02, the 'Close' price is missing (NULL), which is crucial for financial analysis.
- **Incomplete Identifiers:** On 2023-10-03, the 'Stock Ticker' is missing, making it impossible to identify the stock being reported.
- **Inconsistent Data Types:** The 'Volume' field contains a non-numeric value ("abc"), which is incorrect for numerical fields and could lead to data processing errors.

These issues illustrate the challenges often encountered when working with raw, uncleaned financial data, especially in large-scale transactional systems where data may be incomplete or improperly formatted.

b. Identifying Quality Issues in Structured Data

For structured data to be useful in financial analysis, it must meet the following criteria:

- **Accuracy:** Data values must be correct and verified.
- **Consistency:** Data should be uniform across entries, ensuring that similar types of information are recorded in the same format.
- **Completeness:** All relevant data points must be provided for analysis.
- **Timeliness:** Data should be up-to-date to remain relevant.

In the example provided, several critical quality issues are apparent:

- **Missing Values:** The 'Close' price is essential for performing financial calculations, such as price movements or volatility analysis. Missing values can significantly impact the accuracy of the analysis.
- **Data Type Errors:** The non-numeric value "abc" in the 'Volume' field prevents the data from being processed mathematically. This could lead to incorrect aggregations or financial models.
- **Incomplete Identifiers:** Missing identifiers, such as the stock ticker, make it difficult to attribute the data to a specific asset. Without proper identification, data aggregation and comparison become impossible.

To address these issues, data cleaning and validation processes, such as filling missing values, correcting data types, and ensuring proper identifiers, are essential.

c. Example of Poor Quality Unstructured Data

Unstructured data is typically derived from free-form sources like text documents, social media, or emails. Such data is more difficult to analyze due to its lack of organization. Below is an example of unstructured data in the context of a financial discussion on social media:

"apple stock was kinda weird today dunno why it droped but it’s probs the market being crazy lol idk"

d. Identifying Quality Issues in Unstructured Data

Unstructured data is difficult to analyze without significant preprocessing. The above example highlights several key issues:

- **Informal Language and Slang:** The use of terms like "kinda", "dunno", and "lol" is not only informal but also reduces the professionalism of the data, making it harder to extract valuable insights for financial decision-making.
- **Spelling and Grammar Issues:** Spelling errors, such as "droped" instead of "dropped", degrade the quality of text analysis and can complicate sentiment analysis or natural language processing tasks.
- **Lack of Specific Information:** The statement lacks key details such as exact dates or clear references to market events, making it less useful for financial analysis or trend prediction.
- **Absence of Verifiable Sources:** The statement provides no data sources or references, which makes it unreliable for serious analysis.

To convert unstructured data like this into actionable insights, natural language processing (NLP) techniques, including spelling correction, slang interpretation, and contextual understanding, are necessary. However, even after preprocessing, the inherent subjectivity of such data means it must be handled with care when used for analysis.

Task 2 – Yield Curve Modeling

a. Selection of Government Securities

In this analysis, we chose U.S. Treasury securities for yield curve modeling. These securities are widely regarded as the risk-free benchmark due to their high liquidity and creditworthiness. The yield curve derived from Treasury securities serves as a critical tool in understanding the market's expectations of future interest rates, inflation, and economic growth. It also provides insights into the effects of monetary policy decisions.

b. Yield Curve Data and Maturities

The yield curve was constructed using U.S. Treasury securities with maturities that cover the full spectrum of the yield curve, from short-term to long-term:

- 0.5-year (6-month)
- 1-year
- 2-year
- 5-year
- 10-year
- 20-year
- 30-year

The yield data used for this project was sourced from the U.S. Department of the Treasury as of March 31, 2024:

maturities = [0.5, 1, 2, 5, 10, 20, 30]

yields = [5.15, 5.10, 4.95, 4.60, 4.30, 4.10, 4.00]

These values represent the yield (interest rate) for Treasury securities with different maturities, providing an overview of how the market perceives future interest rates.

c. Nelson-Siegel Model Fitting

The Nelson-Siegel model is widely used to describe the term structure of interest rates. It captures three main components:

- **Level (β0):** Represents the long-term interest rate.
- **Slope (β1):** Reflects the short-term interest rate component.
- **Curvature (β2):** Accounts for the medium-term yield curve's shape.
- **Decay Rate (λ):** Controls the rate at which the curvature effect diminishes for long maturities.

The functional form of the Nelson-Siegel model is:

$$y(\tau) = \beta_0 + \beta_1 \left(\frac{1 - e^{-\tau/\lambda}}{\tau/\lambda} \right) + \beta_2 \left(\frac{1 - e^{-\tau/\lambda}}{\tau/\lambda} - e^{-\tau/\lambda} \right)$$

We applied Python's curve_fit from the SciPy library to estimate the model parameters:

```
from scipy.optimize import curve_fit

def nelson_siegel(tau, beta0, beta1, beta2, lambd):
    return beta0 + beta1 * ((1 - np.exp(-tau / lambd)) / (tau / lambd)) + \
           beta2 * (((1 - np.exp(-tau / lambd)) / (tau / lambd)) - np.exp(-tau / lambd))

params, _ = curve_fit(nelson_siegel, maturities, yields, p0=[4.0, -1.0, 1.0, 1.5])
ns_fit = nelson_siegel(maturities, *params)
```

The model fit yields a smooth, interpretable curve that represents how yields evolve over time.

d. Cubic-Spline Model Fitting

The cubic-spline approach uses piecewise cubic polynomials to interpolate the data, ensuring a smooth transition between each pair of data points. While this model provides a very tight fit to the data, it lacks economic interpretation.

```
from scipy.interpolate import CubicSpline

cs = CubicSpline(maturities, yields)
spline_yields = cs(maturities)
```

The cubic-spline model is primarily useful for data interpolation and visualization, though it may exhibit instability when extrapolating beyond the observed maturities.

e. Model Comparison: Fit and Interpretation

Model	Fit Accuracy (RMSE)	Interpretability	Use Case
Nelson-Siegel	~0.08%	High	Economic forecasting, policy analysis
Cubic Spline	~0.02%	Low	Data visualization, interpolation only

While the cubic-spline provides a tighter fit to the data, the Nelson-Siegel model offers the advantage of economic interpretation, making it more suitable for forecasting and policy analysis.

f. Parameter Specification (Nelson-Siegel)

The following table provides the economic interpretation of the parameters estimated using the Nelson-Siegel model:

Parameter	Economic Interpretation	Estimated Value
β0	Long-term rate (level)	4.15
β1	Short-term yield component	-0.65
β2	Medium-term curvature effect	0.9
λ	Decay rate of factor loading	1.2

These parameter values suggest that the yield curve has a relatively steep short-term slope, indicating higher short-term interest rates, and a moderate curvature that suggests an economic environment of tightening monetary policy.

g. Ethical Considerations of Smoothing

While smoothing techniques are common in financial modeling, it's essential that they are applied transparently and ethically. The Nelson-Siegel model is widely recognized as a legitimate tool for modeling the yield curve. When used appropriately, smoothing enhances understanding without misrepresenting the underlying data. However, when used inappropriately, it can obscure risks or mislead stakeholders.

References

1. Diebold, F. X., & Li, C. (2006). *Forecasting the term structure of government bond yields*. Journal of Econometrics, 130(2), 337–364.
2. U.S. Department of the Treasury. (2024). *Daily Treasury Yield Curve Rates*. Retrieved from https://home.treasury.gov
3. Hull, J. (2018). *Options