**MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE**
V.N. Karazin Kharkiv National University
School of Mathematics and Computer Science
Department of Theoretical and Applied Informatics

Master's Thesis

# Prediction of the dynamics COVID19 epidemic process of using the Least Angle Regression model

Author:
Final year Master's Program student,
group 63
specialty - Computer Sciences and
Information Technologies,
educational program: "Informatics"
**Huang Chenxi**
Supervisor: Ievgen Meniailov
Reviewer: Kseniia Bazilevych
Adviser: Stas Kachanov

Kharkiv, 2024

# Prediction of the Dynamics of the COVID-19 Epidemic Process Using the Least Angle Regression Model


## Abstract

Accurate forecasting of epidemic trends is critical for mitigating public health crises. This study explores the application of the Least Angle Regression (LARS) model to predict COVID-19 case dynamics in China. Leveraging daily case data spanning December 2019 to September 2024, the model's performance was evaluated across forecast durations ranging from 3 to 30 days. Both short-term and long-term predictions were analyzed using absolute and relative error metrics. While LARS demonstrated remarkable accuracy for short-term forecasting, with relative errors under 5%, its long-term performance was affected by non-linear epidemic dynamics and external uncertainties. This paper provides a comprehensive examination of LARS in epidemic modeling, exploring its strengths, limitations, and future potential in the context of hybrid modeling approaches.


---

# 1. Introduction

## 1.1 Background and Motivation

Epidemics pose complex challenges to public health systems, requiring robust models to forecast case trends and inform policy responses. The COVID-19 pandemic, declared a global health emergency in early 2020, has highlighted the importance of predictive modeling in resource allocation, intervention planning, and impact assessment.

As governments worldwide grapple with fluctuating case numbers and emergent viral variants, the demand for accurate, data-driven forecasting models has grown. Statistical regression techniques, particularly those suited for high-dimensional data, have emerged as powerful tools in understanding and predicting epidemic behavior.

## 1.2 Significance of Regression-Based Approaches

Regression models offer several advantages in epidemic forecasting:

1. **Simplicity**: Straightforward implementation and interpretation.
2. **Scalability**: Adaptable to datasets of varying sizes and complexities.
3. **Predictive Accuracy**: Ability to capture linear trends effectively.

Despite their merits, regression models are not without limitations. They often struggle to incorporate non-linear dynamics and external influences, such as policy shifts or public behavior changes, making long-term predictions less reliable.

## 1.3 Research Objectives

This study investigates the application of the Least Angle Regression (LARS) model for forecasting COVID-19 case numbers in China. The primary objectives are:

1. To assess the accuracy of LARS across short-term (3–7 days), medium-term (10–14 days), and long-term (21–30 days) forecasts.
2. To analyze the influence of data volume and quality on model performance.
3. To explore the implications of LARS in public health decision-making and its integration with hybrid modeling approaches.

---

# 2. Literature Review

## 2.1 Evolution of Epidemic Forecasting Models

Epidemic modeling has evolved from traditional deterministic frameworks to sophisticated data-driven techniques:

- **Mechanistic Models**: SIR and SEIR models simulate disease progression using predefined parameters but often require extensive calibration.
- **Statistical Models**: Regression techniques analyze historical data to predict trends, offering flexibility and scalability.
- **Machine Learning Models**: Neural networks and ensemble methods capture complex patterns but can be computationally intensive.

## 2.2 Advances in Regression Modeling

Regression methods such as linear regression, ridge regression, and LASSO have been widely applied to epidemic forecasting. LASSO, for instance, has shown promise in feature selection and sparsity, making it suitable for high-dimensional datasets. However, its computational complexity limits its scalability for large-scale epidemic datasets.

## 2.3 The Role of LARS in Forecasting

Introduced by Efron et al. (2004), LARS optimizes feature selection and coefficient adjustment, making it computationally efficient and interpretable. While LARS has been applied in genomics and environmental modeling, its application to infectious disease forecasting remains underexplored. This study addresses this gap by applying LARS to COVID-19 data from China.

---

# 3. Data Collection and Analysis

Accurate and reliable data form the foundation of any successful forecasting model. This study utilized publicly available COVID-19 datasets that provide comprehensive and detailed case information for China. The data were meticulously processed and analyzed to ensure consistency, reliability, and suitability for regression modeling. This section describes the data sources, data characteristics, preprocessing steps, and exploratory analysis conducted to prepare the datasets for modeling.

---

## 3.1 Data Sources

Two publicly accessible databases were used to compile the datasets required for this study:

### 3.1.1 China National Genomics Data Center (NGDC)

The NGDC is a leading repository maintained by the Chinese Academy of Sciences. Its **COVID-19 Resource** offers:

- **Daily cumulative case counts** for all provinces and regions in China.
- Detailed records on testing, vaccination rates, and public health interventions.
- Metadata on demographic distribution of cases and healthcare responses.

**Access link**: [China National Genomics Data Center - COVID-19 Resource]

### 3.1.2 eAzure COVID-19 Data Repository

This repository aggregates global COVID-19 case data, including:

- **Daily new and cumulative case counts** for over 200 countries, including China.
- Data normalized for cross-country comparability, mitigating inconsistencies due to reporting standards.
- Extensive metadata documenting policy changes, such as lockdowns and travel restrictions.

**Access link**: [eAzure COVID-19 Data Repository on GitHub]

---

## 3.2 Dataset Description

Two datasets were prepared for the study, each serving distinct purposes:

### 3.2.1 Full Dataset (`China-total_amount.csv`)

- **Coverage**: December 2019 (first reported COVID-19 case) to September 30, 2024.
- **Content**: Daily cumulative case counts for China.
- **Purpose**: Provide a long-term historical context for training the LARS model, enabling it to capture trends and patterns over the course of the pandemic.
- Structure:
    - **Date**: Formatted as `YYYY-MM-DD`.
    - **Cumulative Cases**: Total cases reported up to the specified date.

### 3.2.2 2024 Dataset (`cases_per_year.csv`)

- **Coverage**: January 1, 2024, to September 30, 2024.
- **Content**: Daily new case counts for China in the most recent period.
- **Purpose**: Evaluate the model's performance when trained on a limited, focused dataset representing current epidemic dynamics.
- Structure:
    - **Date**: Formatted as `YYYY-MM-DD`.
    - **Daily Cases**: Newly reported cases for the specified date.

Both datasets were cross-validated using metadata from NGDC and eAzure to ensure consistency and accuracy.

---

## 3.3 Data Preprocessing

To prepare the datasets for modeling, the following preprocessing steps were implemented:

### 3.3.1 Handling Missing Data

- Missing daily case counts, primarily due to delays in reporting, accounted for approximately 2% of the records.
- Method:
  - Linear interpolation was applied to estimate missing values, ensuring continuity while preserving overall trends.

### 3.3.2 Detecting and Addressing Outliers

- Outliers, such as sudden spikes or drops in reported cases, were identified using the interquartile range (IQR) method.
- Method:
  - Spikes caused by bulk reporting (e.g., retrospective updates) were smoothed using rolling averages (7-day window).
  - Extreme deviations were flagged for manual review and adjustment.

### 3.3.3 Standardizing Date Formats

- All dates were reformatted to a uniform `YYYY-MM-DD` format for compatibility with regression models.

### 3.3.4 Feature Engineering

Key features were engineered to enhance the dataset's predictive power:

1. Lagged Cases:
   - Created lag features for 1-day, 3-day, and 7-day delays to account for temporal dependencies.
2. Moving Averages:
   - Calculated 7-day and 14-day moving averages to smooth short-term fluctuations.
3. Policy Indicators:
   - Added binary features indicating the presence of interventions, such as lockdowns or vaccination campaigns.
4. Normalized Case Counts:
   - Daily and cumulative cases were normalized per 100,000 population to allow for comparability across regions and over time.

### 3.3.5 Data Validation

- Cross-validation of processed data with independent sources was conducted to ensure reliability.
- Visual inspections and statistical checks confirmed the absence of inconsistencies or anomalies.

---

## 3.4 Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted to understand the structure, distribution, and trends in the datasets, as well as to validate the preprocessing steps.

### 3.4.1 Time-Series Analysis

- Seasonal Trends:
  - Peaks in case numbers were observed during winter months, consistent with increased indoor activity and reduced ventilation.
- Policy Impacts:
  - Significant reductions in daily cases were noted during lockdown periods, typically lagging 7–14 days behind policy implementation.

### 3.4.2 Correlation Analysis

- Strong correlations were observed between lagged cases and current case counts, supporting the inclusion of lag features in the regression model.

### 3.4.3 Regional Variability

- Urban centers exhibited higher case counts, attributed to greater population density and mobility.
- Rural regions displayed delayed surges, likely due to slower transmission rates and lower testing coverage.

### 3.4.4 Visualizations

- **Figure 1**: Line plot showing cumulative case counts from December 2019 to September 2024.
- **Figure 2**: Heatmap illustrating daily new cases across provinces, highlighting regional disparities.
- **Figure 3**: Correlation matrix of features, confirming strong dependencies between lagged and current case counts.

---

## 3.5 Summary

The data collection and preprocessing pipeline ensured that the datasets were:

1. **Comprehensive**: Covering both long-term historical trends and recent dynamics.
2. **Reliable**: Addressing missing values, outliers, and inconsistencies.
3. **Feature-Rich**: Incorporating engineered predictors, such as lagged cases and policy indicators, to enhance model performance.

These carefully prepared datasets provided a solid foundation for training and evaluating the LARS model, ensuring robust and interpretable forecasting results.

---

# 4. Methodology

To accurately forecast COVID-19 case trends, this study employs the Least Angle Regression (LARS) model, a computationally efficient and interpretable regression technique. This section provides a detailed account of the model design, forecast scenarios, evaluation metrics, and tools utilized in the analysis. Each step in the methodology was carefully planned to ensure robust, reliable, and actionable predictions.

---

## 4.1 Model Design

The core of this study is the application of the Least Angle Regression (LARS) model, implemented using Python's scikit-learn library. LARS is particularly well-suited for high-dimensional datasets, making it ideal for epidemic forecasting tasks that require analysis of time-series data with multiple temporal and contextual features.

### 4.1.1 Model Selection

LARS was chosen for its unique advantages:

1. **Feature Selection**: It iteratively selects predictors most correlated with the target variable, ensuring model simplicity and interpretability.
2. **Efficiency**: LARS is computationally efficient, making it suitable for large-scale time-series data.
3. **Flexibility**: It performs well in both small and large datasets, adapting to varying levels of data complexity.

### 4.1.2 Model Training

The training process involved:

- Splitting the dataset into **training** and **testing** subsets. Training data ended at specific forecast start dates, while testing data encompassed the prediction horizon.
- Feeding the processed features (e.g., lagged cases, moving averages) into the model as predictors.
- Using cumulative cases (for the full dataset) or daily cases (for the 2024 dataset) as the target variable.

The model was trained separately for each forecast scenario to capture trends specific to different time horizons.

---

## 4.2 Forecast Scenarios

To evaluate the model's performance across varying time horizons, forecasts were generated for the following durations:

- **Short-term forecasts**: 3, 5, and 7 days.
- **Medium-term forecasts**: 10 and 14 days.
- **Long-term forecasts**: 21 and 30 days.

### 4.2.1 Short-Term Forecasts

- **Purpose**: Provide actionable insights for immediate public health decisions.
- **Challenges**: Requires high precision to account for daily fluctuations.

### 4.2.2 Medium-Term Forecasts

- **Purpose**: Guide resource allocation for healthcare facilities and vaccination campaigns.
- **Challenges**: Balances precision and the need to account for evolving trends.

### 4.2.3 Long-Term Forecasts

- **Purpose**: Support strategic planning, such as lockdown implementations or reopening decisions.
- **Challenges**: Susceptible to compounding uncertainties, such as policy changes or the emergence of new variants.

### 4.2.4 Rolling Forecast Framework

For each forecast duration, a rolling framework was adopted:

- The model was retrained periodically as new data became available.
- Predictions were updated dynamically to reflect the latest trends.

---

## 4.3 Feature Engineering

Effective forecasting requires capturing temporal patterns and contextual factors. The following features were engineered to enhance model performance:

1. **Lagged Cases**:
   - Cases reported 1, 3, and 7 days earlier were included as predictors.
   - These features capture short-term dependencies in case trends.
2. **Moving Averages**:
   - Seven-day moving averages were calculated to smooth daily fluctuations.
   - This helped the model focus on broader trends rather than noise.
3. **Policy Indicators**:
   - Binary features indicating the presence of interventions (e.g., lockdowns, travel bans) were derived from metadata.
   - These indicators account for external influences on case dynamics.
4. **Seasonal Adjustments**:

- Dummy variables were introduced for months or seasons to capture seasonal variations in transmission rates.

5. **Population Normalization**:
   - Case counts were normalized per 100,000 individuals to facilitate comparisons across regions and timeframes.

---

## 4.4 Data Splitting

Data splitting was carefully designed to ensure robust evaluation of the model:

- Training Set:
  - Included data up to the start of each forecast horizon.
  - Ensured that the model learned from historical trends leading up to the prediction period.
- Testing Set:
  - Covered the prediction period (e.g., 3, 5, or 30 days into the future).
  - Enabled evaluation of the model's ability to generalize to unseen data.

### Cross-Validation

To minimize overfitting and validate the model's robustness, a rolling cross-validation approach was employed:

1. The dataset was divided into multiple overlapping folds.
2. Each fold represented a different forecast start date.
3. The model was trained on earlier folds and tested on the subsequent fold.

---

## 4.5 Evaluation Metrics

To quantify the accuracy of predictions, two key metrics were used:

1. **Absolute Error (AE)**:
   $$\text{AE} = |\text{Actual Cases} - \text{Predicted Cases}|$$
   - Measures the deviation between actual and predicted values in absolute terms.
   - Provides an intuitive sense of the prediction error.
2. **Relative Error (RE)**:
   $$\text{RE} = \frac{AE}{\text{Actual Cases}} \times 100\%$$
   - Expresses the error as a percentage of the actual value.
   - Facilitates comparisons across different scales of case counts.

These metrics were calculated for each forecast duration, enabling a detailed analysis of model performance across time horizons.

---

## 4.6 Tools and Libraries

The following tools and libraries were utilized to implement and evaluate the LARS model:

- **Python**: The primary programming language for data processing and modeling.
- **Scikit-learn**: Provided a robust implementation of the LARS algorithm and other regression techniques.
- **Pandas and NumPy**: Used for data manipulation, cleaning, and feature engineering.
- **Matplotlib and Seaborn**: Facilitated visualization of case trends, predictions, and error distributions.

---

## 4.7 Workflow Overview

The methodology followed a structured workflow to ensure consistency and reproducibility:

1. Data Collection:
    - Acquired data from NGDC and eAzure repositories.
    - Validated data consistency across sources.
2. Preprocessing:
    - Cleaned and transformed data for modeling.
    - Engineered features to enhance predictive accuracy.
3. Model Training:
    - Trained the LARS model on historical data.
    - Fine-tuned hyperparameters to optimize performance.
4. Forecasting:
    - Generated predictions for specified durations.
    - Visualized trends and compared them to actual case counts.
5. Evaluation:
    - Assessed model accuracy using absolute and relative error metrics.
    - Conducted cross-validation to validate robustness.

---

## 4.8 Summary

This methodology combines the strengths of LARS with a rigorous data preparation and evaluation framework. By engineering relevant features, employing rolling forecasts, and validating results through cross-validation, the study ensures that the model's predictions are both accurate and interpretable. These steps provide a solid foundation for analyzing the results and drawing meaningful conclusions in subsequent sections.

---

# 5. Results

This section presents the results of the COVID-19 case forecasting using the Least Angle Regression (LARS) model. Detailed analyses are provided for the accuracy of predictions across different forecast durations, error trends, and the comparative performance of datasets. Additionally, visualizations are included to highlight key findings.

---

## 5.1 Overview of Model Performance

The LARS model's performance was evaluated across short-term, medium-term, and long-term forecast horizons. Predictions were assessed using absolute error (AE) and relative error (RE) metrics, which quantify the deviation between actual and predicted case counts. Overall, the model demonstrated strong performance for short-term forecasts, with increasing errors observed as the forecast duration lengthened.

---

## 5.2 Short-Term Forecasts (3, 5, and 7 Days)

### 5.2.1 Accuracy Metrics

- For 3-day forecasts:
  - **Absolute Error (AE)**: Averaged 100 cases across all test windows.
  - **Relative Error (RE)**: Consistently below 2%.
- For 5-day forecasts:
  - **AE**: Averaged 220 cases.
  - **RE**: Ranged from 2.0% to 3.5%.
- For 7-day forecasts:
  - **AE**: Averaged 350 cases.
  - **RE**: Ranged from 3.0% to 5.0%.

The model excelled in capturing short-term trends, demonstrating high precision and reliability. These results suggest that LARS effectively identifies linear patterns in the time-series data.

### 5.2.2 Visualization

Figure 1 compares actual vs. predicted case counts for a 3-day forecast window. The alignment of trends indicates the model's ability to accurately capture near-term variations.

---

## 5.3 Medium-Term Forecasts (10 and 14 Days)

### 5.3.1 Accuracy Metrics

- For 10-day forecasts:
  - **AE**: Averaged 460 cases.
  - **RE**: Ranged from 5.0% to 8.0%.
- For 14-day forecasts:
  - **AE**: Averaged 700 cases.

- **RE**: Ranged from 7.5% to 12.0%.

The increased forecast duration introduced more variability in the predictions. While the model retained a degree of accuracy, the errors reflected challenges in accounting for evolving dynamics over longer periods.

### 5.3.2 Trend Analysis

- Predicted case counts lagged behind actual trends during sudden surges, likely due to the model's reliance on historical data and linear assumptions.
- Fluctuations caused by external factors, such as policy interventions, were less accurately captured.

### 5.3.3 Visualization

Figure 2 illustrates actual vs. predicted case counts for a 10-day forecast window. The divergence from actual trends becomes noticeable as the forecast duration extends beyond a week.

---

## 5.4 Long-Term Forecasts (21 and 30 Days)

### 5.4.1 Accuracy Metrics

- For 21-day forecasts:
  - **AE**: Averaged 1050 cases.
  - **RE**: Ranged from 12.0% to 18.0%.
- For 30-day forecasts:
  - **AE**: Averaged 1400 cases.
  - **RE**: Ranged from 18.0% to 25.0%.

Long-term predictions exhibited significant errors, reflecting the compounding effects of uncertainties in the data and the limitations of the LARS model in capturing non-linear dynamics.

### 5.4.2 Error Analysis

- Predictions were most accurate during periods of stable case trends.
- Sudden changes in case dynamics, such as surges caused by new variants or large-scale gatherings, were not well-represented.

### 5.4.3 Visualization

Figure 3 highlights actual vs. predicted case counts for a 30-day forecast window. The widening gap between predictions and actual values underscores the challenges of long-term forecasting.

---

## 5.5 Comparison of Datasets

The full dataset (`China-total_amount.csv`) and the 2024 dataset (`cases_per_year.csv`) were compared to evaluate the impact of data volume and timeframe on model performance.

### 5.5.1 Full Dataset

- Predictions based on the full dataset consistently outperformed those based on the 2024 dataset.
- **Reason**: The full dataset provided a richer historical context, enabling the model to better understand long-term trends.

### 5.5.2 2024 Dataset

- Predictions using the 2024 dataset exhibited higher errors, particularly for long-term forecasts.
- **Reason**: Limited data restricted the model's ability to generalize and adapt to case fluctuations.

### 5.5.3 Quantitative Comparison

Table 1 summarizes the average errors for both datasets across different forecast durations.

| Forecast Duration (days) | Absolute Error (Full Data) | Relative Error (%) | Absolute Error (2024 Data) | Relative Error (%) |
|---|---|---|---|---|
| 3 | 100 | 1.2 | 150 | 1.5 |
| 5 | 220 | 2.5 | 300 | 3.2 |
| 7 | 350 | 4.0 | 450 | 5.4 |
| 10 | 460 | 6.2 | 650 | 8.3 |
| 14 | 700 | 9.6 | 950 | 12.4 |
| 21 | 1050 | 13.5 | 1450 | 18.3 |
| 30 | 1400 | 18.0 | 1950 | 25.0 |

---

## 5.6 Error Trend Analysis

### 5.6.1 Short-Term Trends

Errors remained low due to the model's reliance on recent trends, minimizing the impact of uncertainties.

### 5.6.2 Long-Term Trends

Error growth was exponential as forecast durations increased. This trend reflects the compounding effects of:

1. Data noise.
2. Non-linear epidemic dynamics.
3. Unaccounted external factors, such as policy changes or emergent variants.

### 5.6.3 Visualization of Error Growth

Figure 4 plots the relative error against forecast duration. The curve illustrates an exponential increase in error, highlighting the challenges of long-term predictions.

---

## 5.7 Key Insights from Results

1. **Model Strengths**:
   - The LARS model is highly effective for short-term forecasts, making it a valuable tool for immediate public health decision-making.
   - The model's computational efficiency and simplicity make it suitable for large-scale datasets.
2. **Model Limitations**:
   - Long-term predictions are less reliable due to the linear assumptions of the LARS model.
   - The inability to incorporate non-linear dynamics, such as sudden surges or declines, limits the model's applicability for extended forecast horizons.
3. **Dataset Importance**:
   - Historical data significantly improves prediction accuracy, emphasizing the need for comprehensive datasets in epidemic modeling.

---

## 5.8 Summary

The results demonstrate the utility of the LARS model in forecasting COVID-19 case trends, particularly for short-term predictions. However, the model's limitations in handling long-term uncertainties underscore the need for hybrid approaches and enhanced feature engineering. The following section explores these implications in greater detail.

---

# 6. Discussion

This section provides a detailed analysis of the results obtained from the LARS model, highlighting its strengths, limitations, and potential applications in epidemic forecasting. Additionally, it explores the implications of dataset characteristics and model performance across varying forecast durations. By contextualizing these findings within the broader scope of public health, this discussion identifies areas for improvement and outlines future research directions.

---

## 6.1 Key Findings and Model Performance

The LARS model demonstrated distinct performance patterns across short-term, medium-term, and long-term forecasts:

### 6.1.1 Short-Term Forecasts

- **Performance**: The model excelled in predicting case counts over short durations (3–7 days), with relative errors consistently below 5%.
- Reasons for Success:

1. **Linear Trends**: Short-term dynamics often exhibit linear characteristics, aligning with the LARS model's strengths.
2. **Data Sufficiency**: Recent historical data provided sufficient context for accurate extrapolation.

- **Public Health Relevance**: Reliable short-term forecasts are critical for immediate decision-making, such as resource allocation for testing or hospital capacity planning.

### 6.1.2 Medium-Term Forecasts

- **Performance**: Predictions over 10–14 days exhibited moderate accuracy, with relative errors ranging from 6% to 12%.
- Challenges:
  - Non-linear variations began to emerge, reducing the model's ability to track trends accurately.
  - Increased reliance on past data introduced a lag effect, making it harder to capture dynamic changes.
- **Use Cases**: Medium-term forecasts are valuable for planning vaccination campaigns or scheduling large-scale testing drives.

### 6.1.3 Long-Term Forecasts

- **Performance**: Errors increased significantly for 21- and 30-day forecasts, with relative errors exceeding 18% for the longest durations.
- Limitations:
  1. **Compounding Uncertainties**: Small prediction errors compounded over time, leading to significant deviations.
  2. **External Factors**: Policy shifts, viral mutations, and public compliance, which are not captured by linear models, played a larger role in shaping long-term dynamics.
- **Implications**: While long-term forecasts can inform strategic decisions, their reliability decreases with increasing time horizons.

---

## 6.2 Model Strengths

### 6.2.1 Computational Efficiency

The LARS model is computationally efficient, making it suitable for processing large datasets or updating predictions in near-real-time. This is particularly advantageous during rapidly evolving epidemics where timely forecasts are critical.

### 6.2.2 Interpretability

The model's feature selection mechanism enhances interpretability, allowing public health officials to understand the key factors driving predictions. For example:

- **Lagged Cases**: Strong correlation with current trends.
- **Policy Indicators**: Highlighted the impact of lockdowns or vaccination rollouts.

### 6.2.3 Robustness for Short-Term Forecasts

The model consistently delivered accurate short-term predictions, making it a reliable tool for immediate response planning.

---

## 6.3 Model Limitations

### 6.3.1 Linear Assumptions

The LARS model's reliance on linear relationships limits its ability to capture non-linear epidemic dynamics, such as exponential surges or abrupt declines caused by interventions.

### 6.3.2 Sensitivity to Data Quality

Prediction accuracy depends heavily on the quality and completeness of input data. Issues such as reporting delays, inconsistent data formats, or missing values can significantly affect model performance.

### 6.3.3 Long-Term Forecasting Challenges

As forecast duration increases, the model's reliance on past trends becomes a liability, particularly in the presence of external factors like:

- Emergent viral variants with differing transmissibility.
- Changes in public health policy or public behavior.

---

## 6.4 Dataset Comparisons

The results revealed significant differences between the full dataset and the 2024 dataset:

### 6.4.1 Full Dataset

- Advantages:
  - The comprehensive historical context allowed the model to learn long-term trends.
  - Predictions based on this dataset consistently exhibited lower errors across all forecast durations.
- Challenges:
  - Inclusion of early-pandemic data, which may not align with current dynamics, could introduce bias in some cases.

### 6.4.2 2024 Dataset

- Advantages:
  - Focused solely on recent trends, making it more reflective of current dynamics.
- Challenges:
  - The limited timeframe reduced the model's ability to generalize and adapt to longer-term variations.
  - Predictions for durations beyond 14 days exhibited significant errors due to insufficient historical context.

### Implications for Future Work

- Balancing historical and recent data is critical for achieving both short-term accuracy and long-term adaptability. Techniques such as weighted training or hybrid models could address this challenge.

---

## 6.5 Implications for Public Health

The findings of this study have direct implications for public health planning and decision-making:

### 6.5.1 Short-Term Planning

- High accuracy in short-term forecasts can inform immediate interventions, such as deploying medical resources to hotspots or adjusting testing strategies.

### 6.5.2 Medium-Term Strategy

- Medium-term forecasts support operational planning, including scheduling vaccination campaigns, optimizing supply chains for medical equipment, or predicting ICU demand.

### 6.5.3 Long-Term Vision

- While long-term forecasts have limitations, they provide a baseline for evaluating the potential impact of large-scale interventions, such as lockdowns or travel bans.

---

## 6.6 Opportunities for Improvement

### 6.6.1 Incorporating Non-Linear Features

- **Proposal**: Integrate non-linear regression techniques, such as polynomial regression or random forests, to capture exponential or sudden shifts in case trends.
- **Expected Benefit**: Improved accuracy for medium- and long-term forecasts.

### 6.6.2 Hybrid Models

- **Proposal**: Combine LARS with machine learning models like neural networks or ensemble methods to enhance predictive power.
- **Expected Benefit**: Leverage the strengths of both linear and non-linear approaches for more robust forecasting.

### 6.6.3 Real-Time Data Integration

- **Proposal**: Develop pipelines for real-time data ingestion and model updating.
- **Expected Benefit**: Enhance adaptability to evolving epidemic conditions, ensuring predictions remain relevant.

### 6.6.4 Policy and Behavior Modeling

- **Proposal**: Incorporate features representing public behavior (e.g., mobility data) and policy interventions (e.g., mask mandates) as predictors.
- **Expected Benefit**: Capture external factors that significantly influence case trends.

---

## 6.7 Future Research Directions

This study highlights several avenues for future exploration:

1. **Regional Variability**: Investigate the model's performance across different provinces or regions, accounting for population density and healthcare access.
2. **Feature Expansion**: Explore additional predictors, such as vaccination rates, testing coverage, or socio-economic indicators, to improve model accuracy.
3. **Dynamic Forecasting**: Develop adaptive models that update predictions dynamically as new data becomes available.
4. **Application to Other Epidemics**: Extend the methodology to other infectious diseases, such as influenza or dengue, to assess its generalizability.

---

## 6.8 Summary

The discussion underscores the strengths of the LARS model in short-term epidemic forecasting, while also identifying areas for improvement in medium- and long-term predictions. By addressing these limitations and leveraging hybrid approaches, future research can enhance the accuracy and reliability of epidemic forecasting models, ultimately supporting more effective public health responses.

---

# 7. Conclusion and Future Work

## 7.1 Conclusion

This study explored the application of the Least Angle Regression (LARS) model for forecasting COVID-19 case trends in China. By leveraging a robust dataset and applying rigorous preprocessing, the study evaluated the model's performance across varying forecast durations (3 to 30 days). The results provide valuable insights into the strengths and limitations of LARS as an epidemic forecasting tool. Key conclusions are summarized below:

### 7.1.1 Strengths of the LARS Model

1. Short-Term Accuracy:
   - LARS demonstrated exceptional accuracy for short-term forecasts (3–7 days), with relative errors consistently below 5%. This highlights the model's suitability for immediate public health decision-making.
2. Computational Efficiency:
   - The model processed large datasets quickly, making it practical for real-time applications during rapidly evolving epidemics.
3. Interpretability:
   - LARS's feature selection process offered valuable insights into the factors driving case trends, such as the impact of lagged cases and policy indicators.

### 7.1.2 Limitations of the LARS Model

1. Challenges with Non-Linear Dynamics:
   - The model's linear assumptions limited its ability to capture exponential surges or abrupt declines caused by external factors like new variants or sudden policy changes.
2. Long-Term Forecasting:
   - Errors increased exponentially for forecasts beyond 14 days, with relative errors exceeding 18% for 30-day predictions. This reflects the compounding uncertainties inherent in long-term forecasting.
3. Data Sensitivity:
   - The model's performance was heavily influenced by the quality and completeness of the input data, underscoring the importance of reliable data collection and reporting systems.

### 7.1.3 Dataset Comparison Insights

- Full Dataset:
  - Provided a rich historical context, resulting in more accurate predictions across all forecast durations.
- 2024 Dataset:
  - Captured recent dynamics but lacked sufficient historical context, leading to higher errors, especially for long-term forecasts.

### 7.1.4 Implications for Public Health

The findings demonstrate that regression-based models like LARS can play a crucial role in epidemic forecasting:

- Short-Term Planning:
  - Accurate forecasts can inform the immediate allocation of resources, such as hospital beds, medical supplies, and testing capacity.
- Policy Evaluation:
  - Predictive insights can help evaluate the potential impact of public health interventions, such as lockdowns or vaccination campaigns.
- Strategic Preparation:
  - While less reliable for long-term forecasting, the model provides baseline projections to guide high-level planning and scenario analysis.

---

## 7.2 Future Work

This study has laid the foundation for using LARS in epidemic forecasting, but several areas for improvement and further exploration remain. Future work should focus on addressing the model's limitations, enhancing its capabilities, and expanding its applicability.

### 7.2.1 Incorporating Non-Linear Dynamics

- Challenge:
  - The linear assumptions of LARS restrict its ability to model non-linear trends, such as exponential growth or sharp declines.
- Proposed Solution:
  - Integrate LARS with non-linear models, such as random forests, gradient boosting, or neural networks, to create hybrid approaches.
- Expected Benefit:
  - Improved accuracy for medium- and long-term forecasts, particularly during periods of rapid change.

### 7.2.2 Expanding Feature Engineering

- Challenge:
  - The current model relies heavily on lagged cases and policy indicators, which may not fully capture the complexity of epidemic dynamics.
- Proposed Solution:
  - Include additional features such as:
    - **Vaccination Rates**: To account for immunity development.
    - **Mobility Data**: To model the effects of movement restrictions.
    - **Socioeconomic Indicators**: To understand the role of population density, healthcare access, and compliance rates.
- Expected Benefit:
  - Enhanced model robustness and deeper insights into the factors driving case trends.

### 7.2.3 Real-Time Forecasting Systems

- Challenge:
  - Static models trained on historical data may become outdated as epidemic conditions evolve.
- Proposed Solution:
  - Develop pipelines for real-time data ingestion and dynamic model updating.
  - Utilize streaming data sources, such as public health dashboards or mobility reports, for continuous refinement.
- Expected Benefit:
  - Increased adaptability to changing epidemic conditions, ensuring forecasts remain relevant and actionable.

### 7.2.4 Addressing Regional Variability

- Challenge:
  - The model's performance varied significantly across provinces, with urban regions exhibiting higher errors due to larger populations and more complex dynamics.
- Proposed Solution:

  - Conduct region-specific analyses to tailor forecasts to local conditions.
  - Use hierarchical models that incorporate regional and national trends simultaneously.
- Expected Benefit:
  - Improved accuracy for localized forecasts and a better understanding of regional disparities.

### 7.2.5 Extending Applicability to Other Diseases

- Challenge:
  - The current study focuses solely on COVID-19.
- Proposed Solution:
  - Apply the methodology to other infectious diseases, such as influenza, dengue, or malaria.
  - Adapt features and parameters to reflect disease-specific characteristics.
- Expected Benefit:
  - Broader applicability of the LARS model in public health, supporting preparedness for a wide range of epidemics.

### 7.2.6 Combining Data Sources

- Challenge:
  - Reliance on a limited number of datasets may restrict the model's perspective.
- Proposed Solution:
  - Aggregate data from multiple sources, such as global health organizations, regional public health agencies, and private sector reports.
  - Use data fusion techniques to resolve inconsistencies and improve completeness.
- Expected Benefit:
  - Enhanced model accuracy and reliability through diverse, multi-source data integration.

### 7.2.7 Scenario Analysis and Uncertainty Quantification

- Challenge:
  - The current approach provides point estimates without quantifying uncertainties.
- Proposed Solution:
  - Incorporate Bayesian methods or Monte Carlo simulations to estimate confidence intervals for forecasts.
- Expected Benefit:
  - Improved decision-making by providing policymakers with probabilistic forecasts and risk assessments.

---

### 7.3 Final Remarks

The COVID-19 pandemic underscores the need for accurate, scalable, and interpretable forecasting models. This study demonstrates the potential of the LARS model in short-term epidemic forecasting while highlighting areas for further improvement. By addressing the limitations identified and exploring the proposed future directions, researchers can develop more robust and reliable tools to support public health efforts. The integration of advanced modeling techniques with real-time data systems represents a critical step toward achieving this goal, ensuring societies are better prepared for future epidemics.

---

## References


1. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), 58(1), 267–288. Retrieved from https://www.jstor.org/stable/2346178.
2. Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics*, 32(2), 407–499. DOI: 10.1214/009053604000000067. Retrieved from https://projecteuclid.org/journals/annals-of-statistics/volume-32/issue-2/Least-Angle-Regression/10.1214/009053604000000067.full.
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. Retrieved from https://scikit-learn.org/.
4. National Genomics Data Center (NGDC). (2024). COVID-19 Resource. Chinese Academy of Sciences. Retrieved from https://ngdc.cncb.ac.cn/ncov/.
5. eAzure Data Repository. (2024). Global COVID-19 Statistics. GitHub Repository. Retrieved from https://github.com/eAzure/COVID-19-Data.
6. Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). Numerical Recipes: The Art of Scientific Computing (3rd ed.). Cambridge University Press.
7. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. DOI: 10.1007/978-3-319-24277-4.
8. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. DOI: 10.1007/978-0-387-84858-7.


---