ROBUST REGRESSION AS A BENCHMARK FOR REGRESSION BASED FEDERATED LEARNING

Karunarathna J.H.S.P

219441K

Degree of Master of Science

Department of Electronic and Telecommunication
Faculty of Engineering

University of Moratuwa
Sri Lanka

April 2025

ROBUST REGRESSION AS A BENCHMARK FOR REGRESSION BASED FEDERATED LEARNING

Karunarathna J.H.S.P

219441K

Dissertation submitted in partial fulfillment of the requirements for the degree
Degree of Master of Science

Department of Electronic and Telecommunication
Faculty of Engineering

University of Moratuwa
Sri Lanka

April 2025

DECLARATION

I declare that this is my own work and this Dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 2025.07.02

The supervisor should certify the Dissertation with the following declaration.

The above candidate has carried out research for the Degree of Master of Science Dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr.Upeka Premaratne

Signature of the Supervisor: Date: 2025.07.23

ACKNOWLEDGEMENT

This MSc was carried out from 2021 to 2024 in the Department of Electronic and telecommunication Engineering University of Moratuwa Sri Lanka. I express my heartfelt gratitude to everyone who supported and guided me throughout this MSc journey. In particular, I extend my deepest gratitude to my supervisor, Dr. Upeka Premaratne and course coordinator, Dr. Kasun Hemachandra for their unwavering support, insightful advice, and invaluable guidance. Their expertise, dedication and encouragement have been instrumental in the successful completion of this research.

Further, I am grateful to the University of Moratuwa and the Department of Electronic and Telecommunication Engineering for providing facilities, resources and a supportive academic environment to carry out this research. Also, I would like to thank all the owners of the materials that have been referred to in this research work. Finally, I want to express my gratitude to my wife, family, friends, and colleagues for the support provided to complete this research.

ABSTRACT

This research evaluates the effectiveness of robust regression techniques, specifically Theil-Sen and Repeated Median Regression (RMR) as benchmarks for regression-based Federated Learning (FL) across a range of simulated sensor configurations. FL is a privacy-preserving paradigm that enables collaborative model training across distributed nodes without centralizing the raw data. Standard regression techniques such as linear regression, which is not robust against outliers and heterogeneity are commonly employed in real world FL environments. In contrast, RMR demonstrates the highest resilience and lowest error rates, with Theil-Sen also showing strong performance as a more computationally efficient alternative. RMR achieves the overall highest performance gains, with the best efficiency improvement with respect to the linear federated regression. The findings of the study support the use of robust regression as a dependable alternative in distributed, real-world sensor networks.

Keywords: Federated Learning, Repeated Median Regression, Linear Regression, Theil-Sen Regression

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Description
FL	Federated Learning
IID	Independent and Identically Distributed
MAE	Mean Absolute Error
MSE	Mean Square Error
OLS	Ordinary Least Squares
RFE	Relative Federated Error
RMR	Repeated Median Regression

CHAPTER 1

INTRODUCTION

FL is a key machine learning technique which can use as a solution to critical concerns in the latest context, such as data privacy, distributed computation, and model personalization. FL enables decentralized training by aggregating model updates from local devices without transferring raw data, while conventional machine-learning frameworks rely on centralized data storage[1]. This centralized approach is important in sensitive domains that contain confidential data, like finance, healthcare, business data, and mobile applications, where safeguarding data privacy and security is of utmost importance [2]. Regression analysis is a statistical method used in statistical modelling to identify the relationships between the variables. This method plays a crucial role in FL because it supports decentralized predictive modelling, handling heterogeneous data, privacy-preserving forecasting and proving robustness against outliers. However, it needs advancements in robust regression methodologies and the development of techniques to minimize statistical biases in distributed settings.

Conventional regression models frequently assume that the data for all training examples is uniformly distributed. But in a federated environment, data is usually dispersed among several devices, each with its unique distribution determined by variables including device capabilities, user demographics, and geographic location[3]. Because of the statistical bias that emerges, traditional regression procedures are less useful in federated contexts. Robust regression techniques are showing promise as solutions to these issues, guaranteeing more accurate and dependable regression models in FL[4]. The goal of this thesis is to assess robust regression as a standard for regression-based FL, with an emphasis on how well it handles situations, including heterogeneous and non- Independent and Identically Distributed (IID) data.

1.1 Research Problem

FL has shown promise in a number of fields in the present context, but regression-based FL has some serious drawbacks in the handling of noisy and skewed data distributions. Since data collected on various client devices and device types frequently exhibits non-uniformity and follows diverse statistical distributions due to the data heterogeneity. The assumption that data set which consider is independently and identically distributed is commonly made by conventional regression models, although it is not true in FL situations. Due to this reason, normal regression procedures in these situations frequently result in less generalization and less than ideal model performance.

Outliers in practical scenarios are a key concern which needs to be addressed when working with the FL implementations. Due to theis noise, variations in device quality,

user behaviour, or inadequate data collection, data gathered from a variety of client devices may contain notable variations in real-world FL situations. Further, conventional regression models are extremely susceptible to outliers due to those mentioned reasons, which frequently leads to skewed predictions and lower model accuracy. Improving the dependability of FL-based regression models requires the development of strong regression strategies that can reduce and manage the impact of such outliers.

In FL, it entails repetitive interactions for data transfer between a central server and numerous dispersed clients. Due to this high number of transactions, communication efficiency needs to be optimized to reduce the over resource utilization. Normally, FL systems are vulnerable to latency and bandwidth issues as they require frequent model updates. Robust regression techniques can be effectively used to sustain model stability and convergence while lowering communication overhead. However, in FL, achieving a balance between resilience and communication efficiency is critical for practical deployment in bandwidth-constrained contexts.

Privacy restriction in FL is a key consideration, and regression techniques need to be created to function well inside these privacy-preserving frameworks [5]. Mainly FL was specifically developed to safeguard user data privacy by storing raw data on local devices. Due to that reason, it can sometimes be difficult to maintain the accuracy of robust regression models while adhering to privacy laws in each scenario. This mainly occurs when modifications to the model incorporate sensitive data or unintentionally reveal patterns of private information throughout the FL process. During the FL process, all the users may provide the equal amount or quality of data. This imbalance of data also can lead to biases in the global model, making it unfair or skewed toward certain user groups.

In the client data unbalanced or non-representative scenarios, standard regression models may unintentionally introduce or magnify such biases, resulting in unfair and distorted forecasts. To solve this issue and guarantee equitable performance across various client demographics, fairness considerations must be incorporated into reliable regression techniques. The objective of this thesis is to thoroughly examine, evaluate, and contrast several robust regressions approaches in the framework of FL. Through a thorough assessment of these resilient regression techniques, this study aims to set a standard for their performance in regression-based FL applications, ultimately advancing the creation of FL models that are more trustworthy, equitable, and privacy-preserving.

1.2 Scope and Objectives of the Research

Regression-based FL is the subject of this study, with a focus on robust regression methods as a standard. This research objective is to explore how robust regression methods can strengthen federated learning models when dealing with varied and noisy

data. It focuses on generating synthetic datasets that reflect real-world differences across sensors by comparing techniques such as linear regression, Theil-Sen, and RMR. The goal is to understand and evaluate how these approaches perform at scale, measuring accuracy, efficiency, and their ability to handle outliers. This study aims to offer practical guidance and approach for building more reliable and fair regression models in distributed environments. The study highlights the difficulties presented by decentralized and varied data sources by methodically examining the effects of data heterogeneity and non-IID data distributions on regression performance in federated systems. Regression-based FL is the subject of this study, with a focus on robust regression methods as a standard. The study highlights the difficulties presented by decentralized and varied data sources by methodically examining the effects of data heterogeneity and non-IID data distributions on regression performance in federated systems. Furthermore, because device quality, user behavior, and data gathering methods vary widely in decentralized contexts, it assesses how resilient various regression models are to outliers and noisy data. The evaluation of computational and communication efficiency is another essential component of this study because reliable regression techniques in FL need to strike a compromise between accuracy and real-world factors like latency and bandwidth consumption.

It should be noted that this study primarily examines how robust regression can improve predictive accuracy and fairness in decentralized machine learning frameworks, thereby focusing exclusively on regression-based FL models and purposefully excluding classification-based FL approaches.

1.3 Contribution

This thesis addresses important issues and advances the status of research in the fields of robust regression and FL, making several important contributions and improvements. One of the main contributions is the methodical benchmarking of reliable regression algorithms in federated contexts. This study creates a framework for performance comparison by thoroughly analyzing several robust regression techniques and determining their advantages and disadvantages when used with decentralized data. In this work, non-IID data distributions that are frequently found in practical applications are modelled to explore the impact of data heterogeneity. This research highlights the importance and significance of model selection when working with heterogeneous datasets by offering insightful information. Further, it discusses how various robust regression models react to data variability through in-depth analysis.

Due to the variations in device quality, user behaviour, and data collection techniques, real-world data in federated contexts is frequently susceptible to outliers. This research work addresses the significant issue of outlier resilience in FL by carefully examining the impacts of outliers on FL-based regression models and determining the

best robust regression strategies for reducing these effects. To address the scalability and efficiency concerns in FL, this study evaluates the impact of robust regression techniques on FL's performance. It provides helpful suggestions for maximising effectiveness without sacrificing robustness. Furthermore, this work covers the gap between the traditional regression models and FL environments by tackling these important concerns highlighted previously. The results of this research and approaches presented in this thesis help to improve the robustness, scalability, and adaptability of FL-based regression algorithms to practical applications, which in turn improves the dependability and equity of decentralised machine learning systems.

1.4 Organization of the thesis

This thesis consists of five main chapters and is organized as follows. Chapter 1 is "Introduction" and discussed background of the topic,the research problem, scope and the objectives of the research work and the contribution.

In Chapter 2, "Literature Review" the literature on FL, regression models, and resilient regression procedures is thoroughly examined. This chapter examines earlier research on regression-based FL and talks about different strategies meant to solve issues with model robustness and data heterogeneity. This chapter lays the groundwork for the research challenge this thesis addresses by critically assessing the corpus of existing knowledge.

In Chapter 3, "Methodology," the experimental design and theoretical underpinnings of the robust regression approaches used in the study are described. The theoretical formulations, dataset selection criteria, and FL model implementation are presented, together with the evaluation measures that are used to gauge the models' effectiveness. This chapter also outlines the simulation environment designed to test the efficacy of several robust regression techniques in a FL setting.

"Experimental Results", the fourth chapter, provides the empirical results of applying robust regression techniques to the FL framework. A thorough examination of performance measures, method comparisons, and in-depth explanations of the outcomes are all included in this chapter. It demonstrates how well each robust regression technique handles non-IID data, assesses communication efficacy, and guarantees robustness against outliers. The conclusions derived from these findings are examined closely in order to comprehend their practical significance .

In Chapter 5, "Conclusion and Future Work," the main conclusions drawn from the study are outlined. In addition to outlining possible directions for further research, this chapter considers the contributions made to the field of regression-based FL. It recommends investigating cutting-edge privacy-preserving strategies to improve regression accuracy and model fairness in decentralized contexts, as well as expanding the strong regression benchmarking framework to support deep learning models within FL sce-

narios. By structuring the thesis in this manner, the research progresses from foundational concepts and literature to methodological implementation, empirical analysis, and future perspectives, providing a holistic understanding of robust regression as a benchmark for regression-based FL.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction to Federated Learning and Robust Regression

Mobile phones, tablets laptops and many of the recent devices have become essential part of daily life of the most of the people [6, 7]. The online connectivity increases day by day and the amount of data generated in those devices also increases in a higher phase. From staying connected with loved ones to managing work, fitness,communties and even entertainment, these digitally empowered devices handle it all. As highlighted, these devices collect a massive amount of personal data including photos, messages, health stats, locations, user browsing behaviors and many personal and highly confidential information. Most apps collect and share user data to improve functionality and enhance the overall user experience [8, 9]. While this can lead to better services, it also poses a significant risk especially when sensitive or confidential information is involved. For instance, imagine a user working on a highly confidential project at their workplace, storing critical documents alongside everyday files on their device. If an app is granted permission to share data for optimization purposes, there's a risk that confidential files could be included in the shared data. This creates a serious vulnerability if the shared data is compromised, sensitive information could fall into the wrong hands, posing risk of potential data breaches, financial losses, or reputational damage. This highlights the urgent need for privacy preserving technologies like FL. FL allows apps to improve their performance without directly accessing or transmitting raw user data.

FL is a decentralized machine learning technique that facilitates collaborative training of a model across multiple clients while ensuring the data remain localized. This innovative approach ensures that training data remains decentralized, thereby maintaining the privacy and confidentiality of data on each individual device [10]. As shown in the 2.1 training data is distributed across the number of end devices, which act as individual workers. In the process, a central aggregator collects local model updates from these devices, using their locally stored training data to refine and update a global model [11] iteratively. FL focuses on protecting privacy by allowing models to train directly on user devices, eliminating the need to share raw data with a central server. In Some FL scenarios, only essential updates are shared to improve the model while strengthening privacy. However, the level of privacy maintained depends largely on how these updates are designed and handled during the process [12].

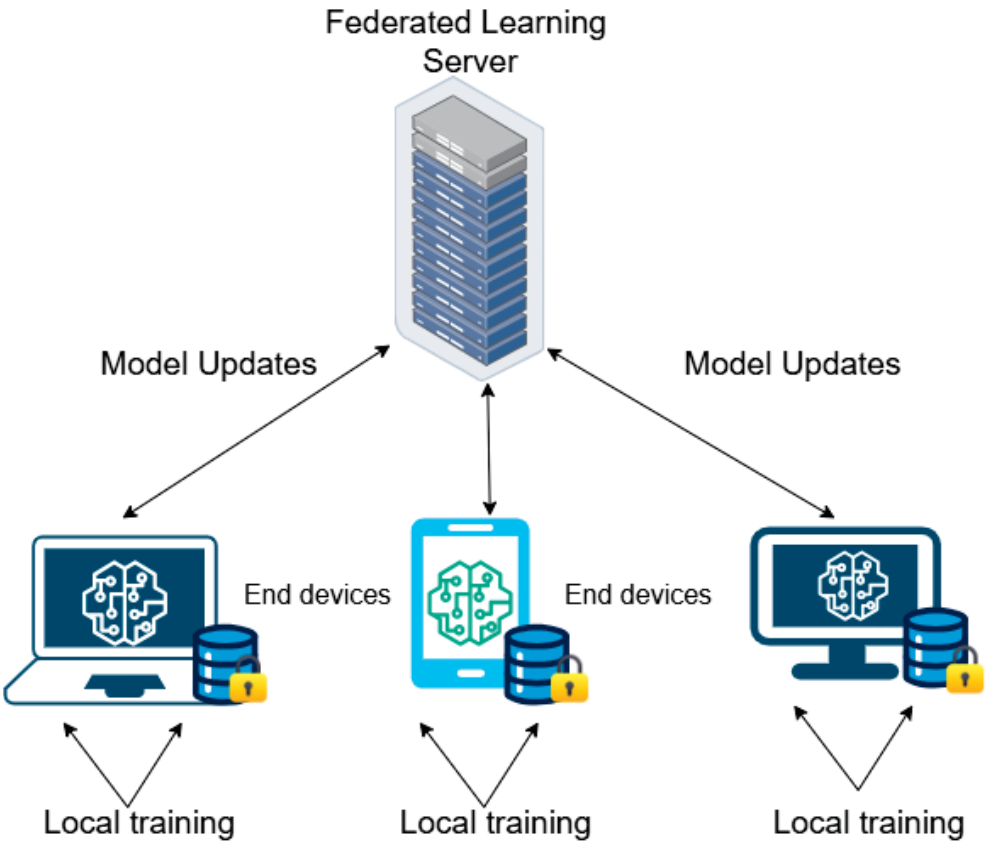

Fig. 2.1: Federated Model Architecture

2.2 Elimination of Statical bias

Statical bias is a critical challenge in FL, and one of the reasons for that is the non-independent and identically distributed nature of data across different clients. This imbalance available in data distribution can lead to biased model updates, reducing fairness and generalization across diverse populations [13]. Addressing the statistical bias in FL is extremely important to obtaining fair and robust machine learning models [14]. This chapter of the thesis reviews existing research regarding the statistical bias in FL and a few methods discussed in this work to mitigate its impact.

2.2.1 Sources of Statistical Bias in Federated Learning

One of the main reasons for bias in FL occurs due to data heterogeneity, where individual clients possess datasets with varying distributions [15],[16]. In normal centralized learning, data is aggregated into a common repository. FL operates with local data, and that may differ in size, quality, and feature representation across clients. This reason can lead clients to drift and models favoring dominant data distributions under representing minority groups and needing to be properly addressed throughout a FL-based

implementation[17].

In FL, imbalanced participation or uneven contribution also result the over representation of their data in the final model. Recent research demonstrates that clients with limited computational capacity or connectivity may engage less often, which can cause bias in federated models [12]. furthermore, clients with different data collection practices may introduce additional bias, as observed in real-world applications such as healthcare and finance [18, 19].

2.2.2 Methods for Eliminating Statistical Bias in FL

Statical bias elimination methods are critical to obtaining accurate predictions and results. Below are some of the methods that can be used for bias elimination.

2.2.2.1 Data Reweighting and Resampling

Data reweighting and resampling is a widely adopted method to address the bias in FL. That process involves assigning different weights to client updates based on their data distribution. For example, [17] presented FedProx, a method using a proximal term to vary each client's contribution, hence reducing the effect of biased updates.

2.2.3 Personalized Federated Learning

Personalized FL methodologies aim to establish models customized for unique client distributions while preserving a global model. One such approach is meta-learning, where models are trained to adapt to client-specific distributions [20, 21]. Another method is clustered FL, where clients with similar data distributions are grouped, allowing models to learn from more homogeneous subsets [22, 23].

2.2.4 Fair Aggregation Algorithms

Bias can also be mitigated at the aggregation level using fair aggregation techniques. For instance, FedAvg [12] was an early approach that averaged model updates, but it often amplified bias due to data imbalance. Later methods, such as FedOpt [24] introduced adaptive learning rates for different clients to equalize their influence. Another technique, q-Fair FL [25] prioritizes the worst-performing clients during aggregation to ensure fairer model updates.

2.2.5 Differential Privacy and Bias Correction

Although differential privacy is mostly applied in security, it can also help reduce bias by stopping overfitting dominant client distributions. Researchers have combined differencial privacy with bias correction methods as an improvement. That approach

works as fair representation learning to ensure models are less sensitive to disparities in client data [26].

2.2.6 Conclusion- Elimination of Bias Federated Learning

Eliminating statistical bias in FL is essential for building fair and reliable artificial intelligence systems. Researchers have explored various approaches, such as data reweighting, personalised learning, fair aggregation, and privacy-preserving techniques, to address this challenge. Although these strategies have significantly enhanced model fairness, practical applications still encounter challenges because of different client participation rates and inconsistent data distributions. As FL evolves, future research must focus on developing adaptive bias mitigation strategies that can dynamically respond to changing data patterns, ensuring that models remain equitable.

2.3 Regression

2.3.1 Introduction

Regression models are frequently used in various industries and domains such as healthcare, finance, telecommunications, manufacturing, and many other industries for predictive modeling and analysis [27]. However, data imbalance, omitted variable bias, and measurement errors are some issues associated with regression [28]. Bias in regression can lead to some issues such as incorrect inferences, poor generalization, and unfair, incorrect decision-making, particularly in critical applications such as medical diagnostics [29]. Statical bias in regression models and methods to mitigate those are discussed in this chapter.

2.3.2 Sources of Statistical Bias in Regression Models

One of the main causes of statistical bias in regression analysis is omitted variable bias and it results from the exclusion of a significant explanatory variable from the statistical model. The omission that appears can lead to biased and inconsistent parameter estimates [30]. As an example, in a salary prediction model the education level is a key factor and if fail to include it in the model may result in an overestimation of the effect of work experience on salary. Furthermore, that predicted model is also a biased result due to the omission. For this kind of salary estimation, both education and work experience contribute, and omitting one differs the results and skews the precision of the model predictions that leading to incorrect and misleading conclusions [31].

Mainly data imbalance and skewed distribution occur when the specific data dominate in the data set. In some practical situations, data can be unbalanced such that one group has way more examples than another. This is a problem because the model

learns mostly from the bigger group and struggles to make accurate predictions for the smaller group. As a result, the model becomes biased and less reliable [32]. Similarly, measurement errors in input variables can introduce bias, as inaccuracies in data collection propagate through the model and affecting regression coefficients and predictions [33].

Endogeneity is another issue that contributes to bias in regression models[34]. It arises when an independent variable is correlated with the error term, leading to biased and inconsistent estimate results. This can be mitigated and addressed by using the instrumental Variable methods, which are used to estimate the causal effect of a variable when there is a potential for endogeneity, meaning the variable of interest is correlated with the error term in a regression model [35].

2.3.3 Methods for Eliminating Statistical Bias in Regression

Statical bias elimination methods are critical to obtaining accurate predictions and results. Below are some of the methods that can be used for bias elimination.

2.3.4 Regularization Techniques

There are few methods used to eliminate statical bias and regularization technique is one of the widely used approach[36]. This technique involves adding penalty terms to the regression coefficients to reduce the over-lifting and improve the generalization of the model. Ridge regression [37] and Lasso regression [38] apply L2 and L1 penalties, respectively, to shrink coefficient values. This helps reduce variance and address bias caused by multicollinearity. Elastic Net, a hybrid of Ridge and Lasso, further enhances stability in cases with highly correlated predictors [39].

2.3.5 Robust Regression Methods

One of the main issues in traditional least squares regression is that it is susceptible to outliers and data anomalies, which introduce bias to models. Robust regression techniques such as Huber regression and quantile regression [40] can be used for more improved results, which help mitigate the influence of outliers by using alternative loss functions that are less sensitive to extreme values. These robust regression methods improve the reliability of regression estimates, especially in datasets with heavy-tailed distributions [41].

2.3.6 Instrumental Variable Approach

The instrumental variable statical bias elimination method is used to address the bias that arises due to endogeneity, which involves identifying external instruments that are related to endogenous regressors but have no association with the error term [42].

After identification, the rest of the filtered variation is used to estimate the causal effect of the independent variable on the dependent variable. The Two-Stage Least Squares method is commonly used to obtain unbiased estimates. In such cases, the first stage predicts the endogenous variable utilizing the instrument, and the second stage uses this prediction to estimate the causal effect [43].

2.3.7 Bias Correction Techniques

A number of bias correction techniques have been developed to improve the accuracy of regression estimates and results by addressing potential estimation errors. The post-stratification weighting introduced by Little [44] is one of the widely used methods that are highly useful in the correction of bias in survey-based regression models. In this technique, the weights of the survey responses adjust based on the known population characteristics, supporting the correction of sample bias and ensuring that the results more accurately reflect the broader population.

Bayesian hierarchical modelling is another practical bias elimination approach, which incorporates prior distributions to regularise parameter estimates [45]. This method is mainly beneficial when working with small sample sizes or sparse data and helps stabilise estimates and reduce the impact of extreme values. Significantly, Bayesian models can mitigate and improve the robustness of the statical inferences by combining the prior knowledge and hierarchical structures.

These statical bias elimination methods play an important role in enhancing the reliability of regression analysis In cases where typical techniques may not yield trustworthy or accurate findings.

2.3.8 Fairness-Aware Regression

Recent research has increasingly focused on developing fairness-aware regression models to mitigate bias, especially in social and economic applications [46]. Fair regression is one of the widely used approaches under fairness-aware regression, which introduces fairness constraints to ensure that model predictions remain equitable across different demographic groups [47]. By including fairness constraints directly into the regression framework, these fairness-aware regression models help reduce disparities in outcomes. This approach is beneficial in domains such as hiring, lending, and healthcare decision-making [48, 49]. Adversarial debiasing is another fairness-aware regression that utilizes supplementary models to detect and mitigate biased representations within regression predictions. This method was initially proposed by Zemel [50], and this method utilizes adversarial networks to enforce fairness by minimizing the extent to which protected attributes, such as race or gender, influence model predictions. By integrating an adversarial element that identifies and rectifies bias during the training

of the primary model, adversarial debiasing improves the fairness and generalizability of regression models, making them more effective in real-world scenarios. These fairness-aware regression techniques are essential for addressing unbiased issues in predictive modelling. Research related to this area is continuously evolving, and further advancements in algorithmic fairness and debiasing strategies will be critical for contributing to transparency and accountability.

Addressing statical bias in regression modelling is critical for ensuring accurate results and reliable, fair predictions. There are a lot of challenges that remain in real-world applications with large data sets and, especially with dynamic evolving data and hidden biases included. Future research in this area should focus on adaptive bias mitigation strategies that integrate fairness and interpretability in regression modelling.

2.4 Robust regression

Although FL offers significant potential, one of the biggest challenges in the FL is addressing the non-independent and identically distributed nature of client data sets. The existence of outliers and noise in data can introduce biases and instability in models if not properly managed. To address this, robust regression techniques play a crucial role in maintaining the reliability of the learning process, even when data quality varies significantly among clients. This chapter explores the advancement of robust regression, its incorporation into FL, and the practical impact of these methodologies in real-world applications.

2.4.1 Fundamentals of Robust Regression

Mainly robust regression evolved from the classical statical methodologies to address the limitations of techniques like Ordinary Least Squares (OLS) which more vulnerable towards the outliers. Researches have demonstrated that even a few irregular data points can heavily skew OLS estimates, leading to unreliable and biased outcomes [51]. Huber introduced m-estimators an approach that modifies the loss function to lessen the impact of large deviations to mitigate this bias issue [52] . Building upon these foundational insights, Rousseeuw and Leroy developed high breakdown point estimators, such as Least Trimmed Squares by further strengthening the field [53]. These estimators are especially useful when data quality is compromised. They can manage large volumes of corrupted data while preserving the reliability of regression models. Over time, robust regression techniques have been widely adopted in various fields, including econometrics and biomedical research. Their broad applicability has proven effective in addressing real-world data challenges. Robust regression enhances the accuracy and stability of estimating underlying trends in data by reducing the excessive influence of outliers. This characteristic makes robust regression an essential tool in

both theoretical and applied statistical studies. It helps preserve model integrity even in complex, data-driven environments where anomalies and noise are prevalent.

2.4.2 Integrating Robust Regression into Federated Learning

Robust regression algorithms are highly beneficial in the context of FL. Heterogeneity is inherent in FL systems since different clients may collect data under different circumstances [54]. Because of this variability, some clients' datasets frequently have noise or outliers that, if carelessly aggregated, could have a detrimental effect on the global model. Before the model changes are aggregated at the central server, the impact of these anomalies can be mitigated by using strong regression procedures at the client level. The use of robust loss functions, like the Huber loss, which has been modified for FL scenarios, is another intriguing discovery by Huber and Leroy et al [52]. The resulting updates are more dependable when each client is given the opportunity to train their local model using a robust loss function. By making sure that no single client's data significantly skews the aggregated model, this approach offers two benefits: it protects the global model from outliers and promotes equity.

2.4.3 Methodological Advances and Practical Considerations

A number of methodological advancements have been presented in recent years, including robust regression approaches into the FL framework, and a few of them are discussed below.

1. Robust Loss Adaptation: Robust loss functions are now directly incorporated into the training process of many FL systems These functions help to mitigate the adverse effects of outliers by down-weighting their influence during the learning process. Each client can dynamically adapt to the quality of their own data thanks to the adaptive nature of these loss functions as per the studies of [52], Huber *et al.*

2. Adaptive Aggregation Schemes: Traditional aggregation techniques, like FedAvg, consider all client updates equally. However, in a robust setting, adaptive aggregation schemes weigh updates based on their estimated reliability. In [55], Yin *et al.* developed such an approach, which actively diminishes the influence of erratic updates while promoting those that are consistent with the broader trend. This results in a more stable and fairer global model.

3. Personalization and Clustered Learning: Robust regression in FL is also being paired with personalization strategies. By clustering clients with similar data distributions, models can be trained to be finely tuned to specific groups while still benefiting from global insights. This is particularly valuable in settings like healthcare, telecommunication or finance, where data heterogeneity is the norm [22].

CHAPTER 3

METHODOLOGY

3.1 Introduction

The growing use of FL in the real world has increased adoption and brought attention to the necessity of strong regression techniques, especially when managing adversarial impacts, data heterogeneity, and statistical bias. Conventional regression techniques in FL frequently provide less-than-ideal models that lack generalization and fairness because of noisy settings, outliers, and non-independent and identically distributed data. To address these challenges, this study explores robust regression as a benchmark for regression-based FL, evaluating its effectiveness in mitigating statistical bias and improving model stability.

This methodological chapter covers the procedures for integrating, validating, and contrasting robust regression approaches inside the FL framework. This methodology procedure ensures an evaluation of robust regression in FL, which includes data collection, processing, model selection, federated training, evaluation, and comparative analysis. The process is presented step by step in the following subsections.

3.2 Data Generation and Noise Model

In the context of sensor data analysis, generating synthetic data is a critical step for testing and evaluating the performance of various regression methods. The synthetic data simulates real-world scenarios, where sensor measurements are subject to noise and inconsistencies. In this study, synthetic data is generated using a combination of linear, Gaussian, and exponential models within the Scilab platform. Each model is designed to exhibit distinct characteristics that simulate various real-world phenomena commonly observed in sensor data collection.

The linear model is the simplest and most fundamental form of data representation. It assumes a linear relationship between the input variable x and the output variable y. The linear equation used for data generation is given by:

$$y = mx + c + \epsilon \tag{3.1}$$

Here, m represents the slope, c is the intercept, and ϵ denotes the noise component. Linear data models are particularly useful when the sensor output is expected to vary proportionally with the input, such as in temperature sensors under controlled conditions.

Noise is inherent in any real-world data acquisition system and is crucial to simulating synthetic data generation. In this study, three types of noise models are incor-

porated, namely Gaussian, Exponential, and Non-Linear Quadratic noise, as discussed below.

3.2.1 Gaussian Noise

Gaussian noise, also known as normal noise, follows a normal distribution with a mean of zero and a specified standard deviation ϵ. It is mathematically represented as:

$$\epsilon \sim N(0, \sigma^2) \tag{3.2}$$

Gaussian noise is prevalent in many real-world applications due to the Central Limit Theorem, where the aggregated effect of numerous small disturbances leads to a normal distribution. This type of noise models scenarios where random fluctuations affect the sensor readings.

3.2.2 Exponential Noise

Exponential noise is characterized by a distribution where the probability decreases exponentially as the deviation from the mean increases. It is modeled as:

$$\epsilon \sim \text{Exp}(\lambda) \tag{3.3}$$

This noise model is suitable when sensor errors or measurement anomalies exhibit an exponentially decaying distribution, often seen in signal attenuation and decay processes.

3.2.3 Non-Linear Quadratic Noise

To introduce non-linearity, a quadratic term is added to the synthetic data generation model. The corresponding equation is given below.

$$y = mx + c + \alpha x^2 + \epsilon \tag{3.4}$$

Here, α represents the non-linearity coefficient. This model is used when the relationship between the sensor input and output is not purely linear, such as in systems affected by physical deformation or varying environmental factors.

3.3 Data Generation

Synthetic data is generated using a combination of linear equations integrated with the previously described noise models. Each dataset consists of 100 samples per sensor, with the slope and intercept randomly selected from a predefined range to reflect variations among different sensors. Noise is then incorporated into each data point to mimic

real-world uncertainties and signal disturbances. Furthermore, outliers are intentionally introduced by modifying certain data points with an increased noise factor. This ensures that the dataset can effectively test the resilience of regression methods when dealing with noisy conditions.

3.3.1 Experimental Design

This study simulates a FL environment in which multiple sensors generate data based on slightly different linear models. The goal is to analyze how different regression methods perform under various levels of sensor heterogeneity and noise, across a range of sensor counts from as few as 10 to as many as 10,000. Each sensor collects data independently and trains a local model, which is then aggregated globally in a federated manner.

Each sensor follows a general model given by:

$$y = a \cdot x + b + \epsilon + \eta x^2 \tag{3.5}$$

where a and b are the unique slope and intercept for each sensor, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the Gaussian noise with $\sigma = 1$, and $\eta = 0.05$ introduces a minor non-linearity into the otherwise linear model. In addition, an outlier noise factor of 0.2 is incorporated to simulate unexpected disturbances, thereby increasing the realism of the dataset.

To evaluate how the system performs as the sensor network expands, the total number of deployed sensors, denoted by N, was systematically adjusted across a series of predetermined values. The range started with a minimal configuration of 10 sensors and was incrementally increased through representative steps such as 20, 50, 100, 200, 300 up to a maximum of 10,000 sensors. This gradual scaling approach was intentionally chosen to thoroughly examine the model's performance, scalability, and adaptability as the network size grew from small-scale to large-scale environments.

3.3.2 Sensor Parameter Scenarios

3.3.2.1 Scenario 1: Low Heterogeneity (Narrow Uniform Range)

In this baseline scenario, sensor parameters (slopes and intercepts) are drawn from a narrow uniform distribution within the interval $[0.9, 1.1]$:

$$\theta_i \sim \mathcal{U}(0.9, 1.1) \tag{3.6}$$

This setting represents a controlled environment where all sensors are well-calibrated and operate under similar conditions. It simulates deployments such as lab-based experiments, newly manufactured devices, or systems with high synchronization and frequent maintenance. Low heterogeneity is essential for benchmarking algorithm per-

formance under ideal conditions, helping isolate the impact of increasing diversity in later scenarios [56].

3.3.2.2 Scenario 2: Moderate Heterogeneity (Wider Uniform Range)

In this scenario, greater variation among sensor parameters is introduced. Slopes and intercepts are uniformly sampled from the range $[0.5, 2.0]$:

$$\theta_i \sim \mathcal{U}(0.5, 2.0) \tag{3.7}$$

This broader uniform distribution models moderately inconsistent sensors, which is common in distributed or aging networks. It reflects real-world cases where sensor wear, minor miscalibration, or environmental differences introduce variation without skew or bias. This is typical in large-scale, lightly maintained networks.

3.3.2.3 Scenario 3: Normal Distribution (Centered Variation)

Here, sensor parameters are sampled from a normal distribution centered at 1.0 with a standard deviation of 0.1:

$$\theta_i \sim \mathcal{N}(1.0, 0.1^2) \tag{3.8}$$

This configuration mimics sensor deployments where most devices perform consistently, but small deviations exist due to manufacturing tolerances, calibration shifts, or environmental noise. Such distributions are widely used in simulations of realistic sensor behavior due to their symmetry and well-understood statistical properties [57].

3.3.2.4 Scenario 4: Skewed Parameters (Exponential Distribution)

This final scenario introduces asymmetry by sampling sensor parameters from an exponential distribution with rate $\lambda = 2$:

$$\theta_i \sim \text{Exponential}(\lambda = 2) \tag{3.9}$$

The exponential distribution captures long-tailed, skewed behavior, typical of systems under stress or degradation. It is especially suitable for simulating faults, harsh operating environments, or sporadic but severe deviations in sensor output. These conditions challenge regression robustness and help evaluate model performance under extreme heterogeneity [58].

3.4 Implementation of Robust Regression Algorithms

This study discusses and analyses linear regression, Theil-Sen regression, and repeated median regression under different conditions. Linear regression is a traditional regression method that minimizes the sum of squared errors to determine the best-fit line. Theil-Sen regression is a robust regression method that calculates the median slope from all possible data point pairs [59]. Normally Theil-Sen regression is less sensitive to the outliers. RMR increases robustness by using an iterative process based on medians. This approach offers stronger resistance to extreme values and noisy data. It is useful when datasets contain irregular or outlier points. By comparing RMR with other regression methods, this study evaluates how well each technique performs under such challenging conditions.

3.4.1 Least Squares Linear Regression

Linear regression is a simple and widely used method for modeling the relationship between an independent and a dependent variable. It represents this relationship as below:

$$y = mx + c + \epsilon \tag{3.10}$$

In this expression, m denotes the slope of the regression line, c is the intercept, and ϵ captures the random error present in the observations. The parameters m and c are typically estimated using the least squares approach, which seeks to minimize the total squared differences between observed and predicted values. This is mathematically formulated as:

$$\min_{m,\,c} \sum_{i=1}^{n} \left(y_i - (mx_i + c)\right)^2 \tag{3.11}$$

This minimization ensures that the line best fits the given data by reducing the overall prediction error.

This optimization ensures the best-fitting line that minimizes the error between actual and predicted values. However, standard linear regression is highly sensitive to outliers, making it less suitable in federated settings where sensor noise and adversarial conditions are prevalent.

3.4.2 Theil-Sen Robust Regression

Theil-Sen regression is a non-parametric method that estimates the slope as the median of all pairwise slopes:

$$m = \text{median}\left(\frac{y_j - y_i}{x_j - x_i}\right),\ \forall\ i < j \tag{3.12}$$

The intercept is then estimated as:

$$c = \text{median}(y - mx) \tag{3.13}$$

The Theil-Sen robust regression method is highly robust to noise and outliers. To incorporate various sensor models in a FL framework, the regression parameters are aggregated across all sensors. The federated regression model is calculated using a weighted average of the individual sensor models. For both linear and robust regression, the federated slope and intercept are computed as follows.

$$m_{\text{fed}} = \frac{\sum w_i m_i}{\sum w_i}, \quad c_{\text{fed}} = \frac{\sum w_i c_i}{\sum w_i} \tag{3.14}$$

Here regression parameters are aggregated across all sensors to include various sensor models in a FL framework. In this approach, the federated regression model is calculated using a weighted average of the individual sensor models.

3.4.3 Repeated Median Regression

The Repeated Median method extends Theil-Sen by computing local slopes for each point which formulate as:

$$m_{RMR} = \text{median}\left(\text{median}_{j \neq i}\left(\frac{y_j - y_i}{x_j - x_i}\right)\right) \tag{3.15}$$

and the intercept as:

$$c_{RMR} = \text{median}(y_i - m_{RMR} x_i) \tag{3.16}$$

This regression method provides increased robustness by eliminating the influence of extreme outliers. By computing the median of medians, RMR enhances stability, making it more resilient to both noise and adversarial data variations. The key advantage of RMR lies in its ability to mitigate the effects of local fluctuations while maintaining reliable parameter estimation.

The FL model aggregates the regression parameters from multiple sensors to generate a global model. The key aspects of this implementation are as follows

1. Model Parameter Aggregation : In this approach, main regression parameters including slope and intercept are aggregated using a weighted averaging approach. This approach improves the accuracy of the global parameter estimates by inte-

grating insights from multiple sensors. At the same time, it reduces the impact of outliers, resulting in a more stable and reliable model.

2. Equal Contribution of Sensors : All sensors are assigned equal weight in this aggregation approach to ensure uniform contribution to the resulting federated model. This method helps mitigate the risk of skewed parameter estimation that could arise from disparities in individual sensor data by promoting fairness and consistency in the model.

3. Robustness to Local Fluctuations : The federated regression model improves stability against fluctuations by aggregating outputs from multiple sensors. This approach minimizes the effects of local data variations and ensures the consistency of the global model and it makes more resilient to sensor-specific noise and fluctuations.

This regression method reduces the effect of local data fluctuations which is commonly appear in real scenarios from individual sensors and guarantees robustness in the federated regression model.

3.5 Federated Regression

Federated regression is a specialized regression method within FL designed to conduct regression analysis. Data privacy can be mainained and data is decentralized. Federated regression differs from traditional machine learning by keeping raw data on local devices rather than transferring it to a central server. In federated regression only model updates are shared by preserving data privacy while enabling collaborative training. This approach is important when working with sensitive or widely distributed datasets. Data collected from personal devices or remote sensing equipment are examples for this kind of privacy related data sets. Each device can perform computations locally without transferring raw data, and only the resulting model parameters are shared with a central aggregator. This ensures the learning process stays decentralized and secure. It is particularly useful in areas such as healthcare, sensor networks, and telecommunications, where protecting data and reducing exposure are key priorities.

3.5.1 Federated Regression Formulation

In the context of federated learning for regression, a distributed setting is considered where multiple clients collaboratively contribute to learning a global regression model. In this federated regression, let the local dataset for each client i be defined as:

$$D_i = \{(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \ldots, (x_{in_i}, y_{in_i})\} \tag{3.17}$$

as per the above formulation x_{ij} represents the feature vector, and y_{ij} is the corresponding target value. The objective is to establish a global regression model, given by:

$$y = f(x) + \epsilon \tag{3.18}$$

where $f(x)$ is the regression function (e.g. linear or polynomial), and ϵ is a noise term. Here,each client independently trains a local regression model by minimizing its local loss function expressed as:

$$\theta_i = \arg\min_{\theta} \frac{1}{n_i} \sum_{j=1}^{n_i} L(y_{ij}, f(x_{ij}; \theta)) \tag{3.19}$$

Once local training is completed, each client transmits its computed model parameters θ_i to a central server. The server then aggregates these parameters to construct a global model which formulate as:

$$\theta_{\text{global}} = \frac{1}{N} \sum_{i=1}^{N} \theta_i \tag{3.20}$$

If client datasets are imbalanced, a weighted averaging strategy can be employed during aggregation to ensure fair model representation. This FL strategy allows decentralized data to contribute to the global model while preserving data privacy and minimizing communication overhead.

3.5.2 Federated Regression Techniques

Federated regression can be applied using various regression techniques, each tailored to specific data characteristics and robustness requirements. This study explores three key techniques such that linear regression, Theil-Sen regression, and RMR, each offering distinct advantages in handling heterogeneous data.

3.5.2.1 Linear Regression

Linear regression is among the simplest techniques for modeling the relationship between input and output variables, assuming a direct linear dependency. In a federated setting, each client independently learns a local model using its private data, which takes the form:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{3.21}$$

In this equation:

β_0 is the intercept term, β_1 denotes the slope or weight assigned to the input variable x, ϵ is the error term capturing randomness or noise in the observations.

After local training, each participant (client) transmits the estimated coefficients $\beta_0^{(i)}$ and $\beta_1^{(i)}$ to the coordinating server. The global model parameters are then computed using a sample-size-weighted average as follows:

$$\beta_0^{\text{global}} = \sum_{i=1}^{N} \alpha_i \cdot \beta_0^{(i)}, \quad \beta_1^{\text{global}} = \sum_{i=1}^{N} \alpha_i \cdot \beta_1^{(i)} \tag{3.22}$$

Here, $\alpha_i = \frac{n_i}{\sum_{j=1}^{N} n_j}$ represents the weight of each client based on its local dataset size n_i relative to the total number of samples across all N clients.

3.5.2.2 Theil-Sen Regression

The Theil-Sen estimator offers a robust alternative to OLS by calculating the slope as the median of all pairwise slopes formed from the dataset. This non-parametric approach is known for its resistance to the influence of outliers and skewed data distributions.

In a FL framework, each client node evaluates its local slope by computing the set of slopes between every distinct pair of data points:

$$m_{jk}^{(i)} = \frac{y_k^{(i)} - y_j^{(i)}}{x_k^{(i)} - x_j^{(i)}}, \quad \text{for all } j < k \tag{3.23}$$

Here, $m_{jk}^{(i)}$ represents the slope between the j-th and k-th points in client i's dataset. After calculating all such pairwise slopes, the client determines its local Theil-Sen slope estimate $\tilde{\beta}_1^{(i)}$ as:

$$\tilde{\beta}_1^{(i)} = \text{median}\left(\left\{m_{jk}^{(i)}\right\}\right) \tag{3.24}$$

The client also computes the corresponding intercept $\tilde{\beta}_0^{(i)}$ as the median of the adjusted values:

$$\tilde{\beta}_0^{(i)} = \text{median}\left(y_j^{(i)} - \tilde{\beta}_1^{(i)} x_j^{(i)}\right) \tag{3.25}$$

These robust local estimates $(\tilde{\beta}_0^{(i)}, \tilde{\beta}_1^{(i)})$ are then transmitted to the central server, which aggregates them by taking the component-wise median across all clients:

$$\tilde{\beta}_1^{\text{global}} = \text{median}\left(\left\{\tilde{\beta}_1^{(i)}\right\}_{i=1}^{N}\right), \quad \tilde{\beta}_0^{\text{global}} = \text{median}\left(\left\{\tilde{\beta}_0^{(i)}\right\}_{i=1}^{N}\right) \tag{3.26}$$

This median-based fusion helps mitigate the effect of anomalous or skewed local data distributions, making Theil-Sen Regression particularly effective in heterogeneous and noisy data environments.

4.3.3 Repeated Median Regression

RMR is a robust extension of the Theil-Sen method, specifically tailored to withstand both vertical outliers and leverage points. This approach enhances stability in the presence of irregular or noisy data by applying a two-level median filtering process.

For each observation (x_j, y_j) in the dataset, the method computes the slope to every other point (x_k, y_k), where $j \neq k$, as follows:

$$m_{jk} = \frac{y_k - y_j}{x_k - x_j} \tag{3.27}$$

The median of these values for a fixed point j gives the repeated median estimate for that observation:

$$\tilde{m}_j = \text{median}\left(\{m_{jk} \mid k \neq j\}\right) \tag{3.28}$$

The overall slope estimate $\tilde{\beta}_1$ is then calculated by taking the median across all repeated median values:

$$\tilde{\beta}_1 = \text{median}\left(\{\tilde{m}_j\}_{j=1}^{n}\right) \tag{3.29}$$

Once the slope $\tilde{\beta}_1$ is established, the intercept $\tilde{\beta}_0$ is obtained using the median of the adjusted residuals:

$$\tilde{\beta}_0 = \text{median}\left(y_j - \tilde{\beta}_1 x_j\right) \tag{3.30}$$

In a FL context, each client i computes its local parameter estimates $(\tilde{\beta}_0^{(i)}, \tilde{\beta}_1^{(i)})$ using the repeated median method on its private dataset. These estimates are sent to the central server, where a second round of aggregation takes place:

$$\tilde{\beta}_1^{\text{global}} = \text{median}\left(\left\{\tilde{\beta}_1^{(i)}\right\}_{i=1}^{N}\right), \quad \tilde{\beta}_0^{\text{global}} = \text{median}\left(\left\{\tilde{\beta}_0^{(i)}\right\}_{i=1}^{N}\right) \tag{3.31}$$

This dual-median process enables RMR to offer high resilience in federated environments where client data may contain significant noise, any kind of variability or

outliers.

Federated regression provides multiple advantages, specially in contexts where data sensitivity is a concern. Data transmission transmits full datasets, and clients share only model parameters or updates such that substantially lowers communication requirements and conserves bandwidth. Another significant benefit is the ability of federated regression to accommodate data diversity. In here, each participating node trains its model based on its unique local data and the system is inherently capable of handling non-uniform distributions across clients. Scalability is also a natural advantage of this approach. With more devices contributing, the aggregated global model becomes increasingly robust and representative of a broader data spectrum. Furthermore, the architecture is resilient to partial client participation even if some nodes are temporarily offline or fail to contribute. Also the overall model quality remains largely unaffected due to the distributed and redundant nature of the training process.

3.6 Evaluation Metrics

Evaluating the effectiveness of regression models, particularly in decentralized settings like FL, requires reliable performance indicators. In this study, mainly three evaluation metrics used as Mean Square Error (MSE) , Mean Absolute Error (MAE) , and Relative Federated Error (RFE) . These metrics help assess how well the models predict target values under different learning configurations.

3.6.1 Mean Absolute Error

The MSE quantifies the average of the squares of the differences between predicted values and actual observations. It is defined mathematically as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.32}$$

This metric penalizes larger errors more heavily due to the squaring operation, making it suitable for applications where significant deviations are particularly undesirable.

3.6.2 Relative Federated Error

RFE quantifies the relative error of the federated model compared to individual sensor models. It is calculated as:

$$RFE_{MSE} = \frac{MSE_{fed}}{MSE_{avg}} \times 100 \tag{3.33}$$

$$RFE_{MAE} = \frac{MAE_{fed}}{MAE_{avg}} \times 100 \tag{3.34}$$

Where:

1. MSE_{fed} = MSE of the federated model
2. MSE_{avg} = Average MSE of individual sensor models
3. MAE_{fed} = MAE of the federated model
4. MAE_{avg} = Average MAE of individual sensor models

This metric indicates the relative difference in error between the federated model and the average of standalone models. A lower RFE value suggests that the federated model performs comparably or better, affirming the success of the aggregation process. This methodology ensures a rigorous evaluation of regression models in both individual and federated settings. By incorporating robust regression techniques and federated learning principles, the approach enhances model resilience against noise, outliers, and sensor heterogeneity, making it highly effective for real world applications.

3.6.3 Efficiency Metric for Federated Regression Model Evaluation

To comprehensively assess the performance of regression models in a federated learning setting, an efficiency metric is employed that integrates both Mean Squared Error (MSE) and Mean Absolute Error (MAE). This provides a unified, scale-invariant measure of model robustness.

Importantly, the reported MSE and MAE values are not from individual sensor models but are aggregated from the federated models after the learning process. This ensures that the efficiency reflects the global performance across the distributed sensor network.

$$\text{Efficiency} = \frac{\text{MSE}}{(\text{MAE})^2} \tag{3.35}$$

Where MSE measures the average squared difference between predicted and actual values, thereby penalizing larger errors more severely. On the other hand, MAE quantifies the average absolute difference between predicted and actual values, offering a more intuitive sense of model accuracy.

The squared MAE in the denominator brings both metrics to a comparable scale. A lower value of Efficiency implies a better-performing model, as it maintains low squared error relative to its average absolute error.

This calculation was consistently applied for each value of N and for all three federated regression models: linear regression , Theil–Sen, and RMR.

Percentage Improvement

In this study, where efficiency is derived from error-based metrics such as MAE and MSE, lower values indicate better model performance as discussed. Therefore, the percentage improvement is computed using the following formula:

$$\text{Percentage Improvement} = \left(\frac{\text{Method Value} - \text{Linear Value}}{\text{Linear Value}} \right) \times 100 \tag{3.36}$$

A negative percentage value signifies that the method under evaluation achieves a lower error than the baseline linear regression, indicating a better performance. Conversely, a positive percentage implies higher error and hence poorer performance compared to the baseline. In this context, more negative percentage improvements reflect greater robustness and accuracy relative to the baseline.

CHAPTER 4

EXPERIMENTAL RESULT

4.1 Results and Discussion

This section presents a comprehensive analysis of the results obtained from simulating different sensor configurations under three primary regression techniques: linear regression, Theil-Sen regression, and RMR. The behavior of the system was investigated under varying numbers of sensors and across three statistical distributions: uniform, normal (Gaussian), and exponential. The performance of each regression method was evaluated based on the accuracy of trend estimation and robustness to outliers and noise.

4.2 Initial Evaluation Under Uniform Sensor Parameters

To evaluate the performance of federated regression methods under a controlled and consistent setup, simulations are first conducted using a uniform distribution for sensor model parameters. Specifically, slope and intercept values are generated from a uniform range between 0.9 and 1.1. These parameters govern the behavior of each synthetic sensor's linear trend, while controlled Gaussian noise, synthetic outliers, and a small non-linearity component are incorporated to better reflect real-world sensor variability. Each sensor generated 100 data points, which were then used to fit three types of regression models independently:

1. **Linear Regression**: Standard least-squares fitting.
2. **Theil-Sen Regression**: A median-based robust estimator.
3. **Repeated Median Regression**: A highly robust estimator less sensitive to outliers.

For each regression method, local model parameters (slope and intercept) were calculated individually for each sensor. The federated model was obtained by averaging the local parameters across all sensors. To assess the accuracy of the federated models, Mean Squared Error (MSE) and Mean Absolute Error (MAE) are computed based on the true underlying model, which is defined by the average of the base parameters. Furthermore, RFE is computed as the ratio (expressed as a percentage) of the federated model error to the average error of individual sensor models. The results of these experiments are presented across a range of sensor counts, increasing from N = 10 to N = 10,000 in logarithmic increments. Each figure illustrates the generated sensor data and

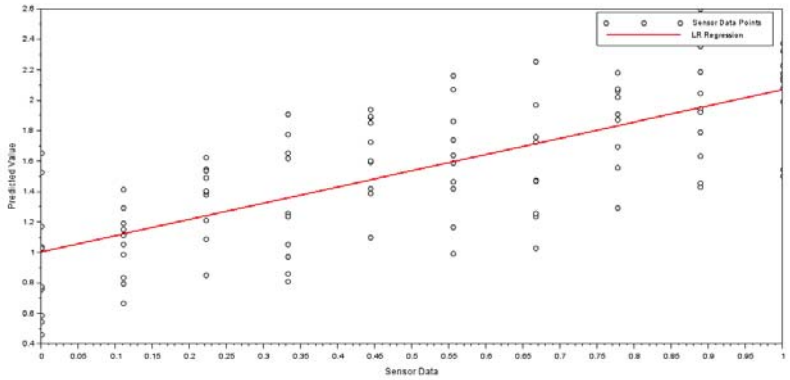

Fig. 4.1: Linear regression federated model for N=10

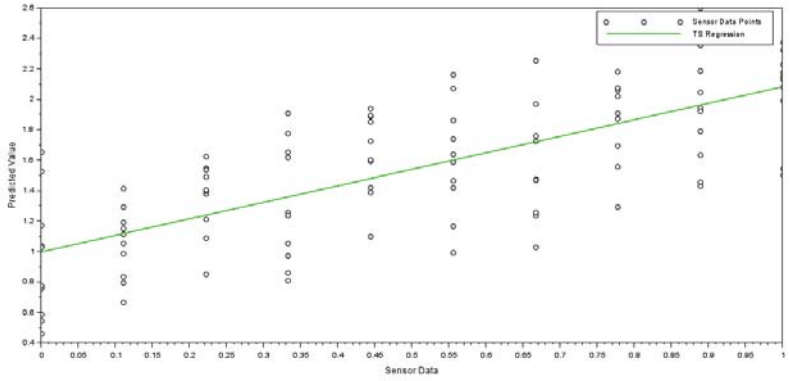

Fig. 4.2: Theil-sen regression federated model for N=10

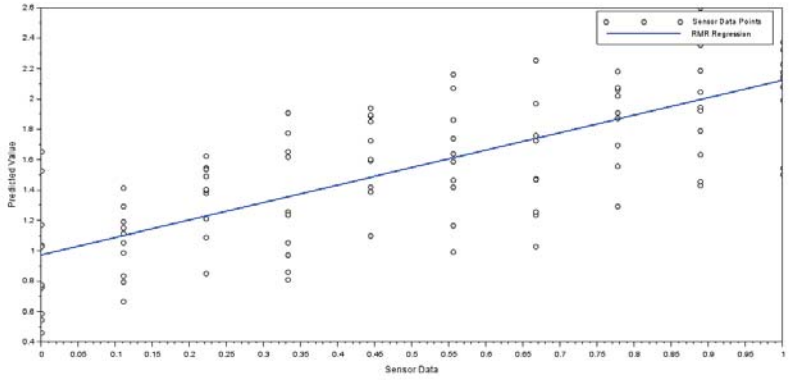

Fig. 4.3: Repeated mean regression federated model for N=10

the regression line produced by each federated method. Figure 4.1 Figure 4.2 and Figure 4.3 respectively show the regression line produced by each federated method and Figure 4.4 shows the three federated regression lines in a single figure. Separations

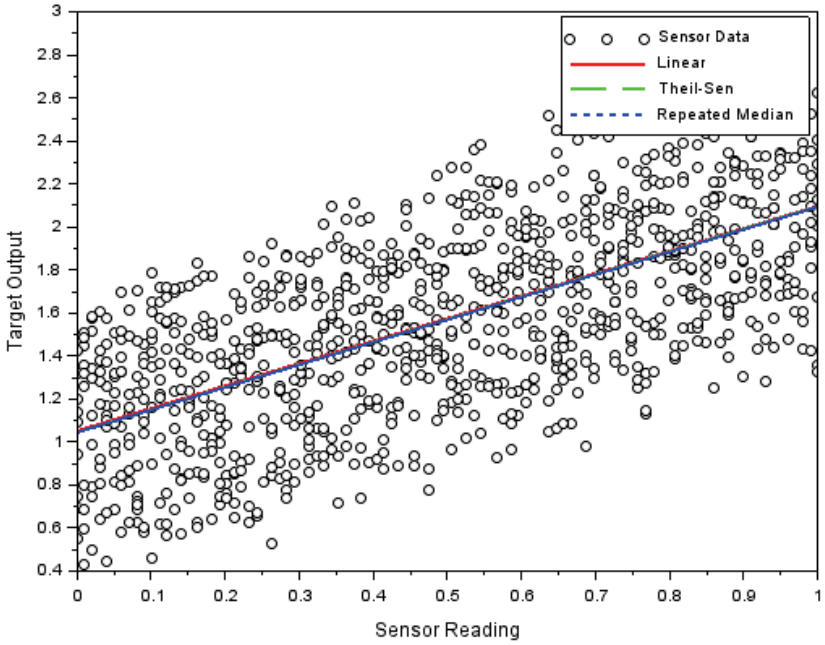

Fig. 4.4: Regression models federated model for N=10

of the lines are difficult to observe due to the federated model of each lying nearby. This initial uniform configuration forms the baseline for comparison with further evaluations under normal and exponential parameter distributions, discussed in the next section.

4.2.1 Performance of Regression Techniques

The performance comparison among the three regression methods reveals distinct advantages and limitations associated with each approach. Linear Regression, being the most basic of the three, performed adequately in well-distributed datasets, particularly under the normal distribution.

However, due to its sensitivity to outliers, linear regression showed noticeable deviation when sensor values exhibited large variance as shown in the Figure 4.5 and Figure 4.6. This behavior was specially prominent in scenarios where sensor readings included extreme values, which skewed the regression line, leading to inaccurate trend estimation. Theil-Sen Regression offered a significant improvement over linear regression, particularly in its ability to handle noisy and skewed datasets. Its median-based approach ensured better resistance to outliers and allowed more consistent trend estimation even when the sensor distribution included high variance or irregular sampling. This robustness was most evident under exponential distribution scenarios, where Theil-Sen produced smoother and more stable fits compared to linear

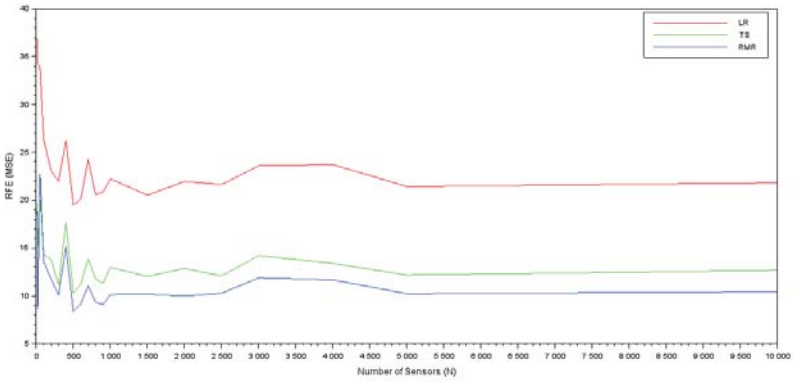

Fig. 4.5: Relative federated error (MSE) of uniform distribution method 1

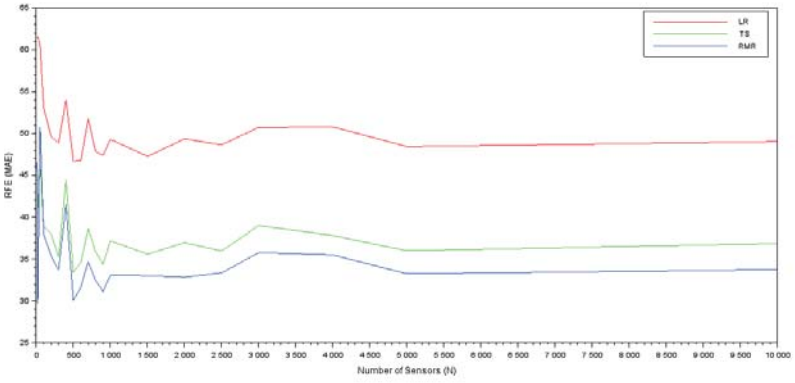

Fig. 4.6: Relative federated error (MAE) of uniform distribution method 1

regression.

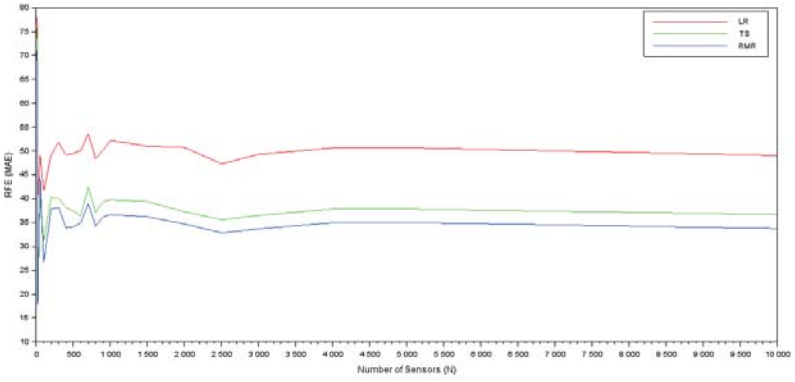

Fig. 4.7: Relative federated error (MAE) of exponential distribution

RMR emerged as the most robust technique in the study. By computing the me-

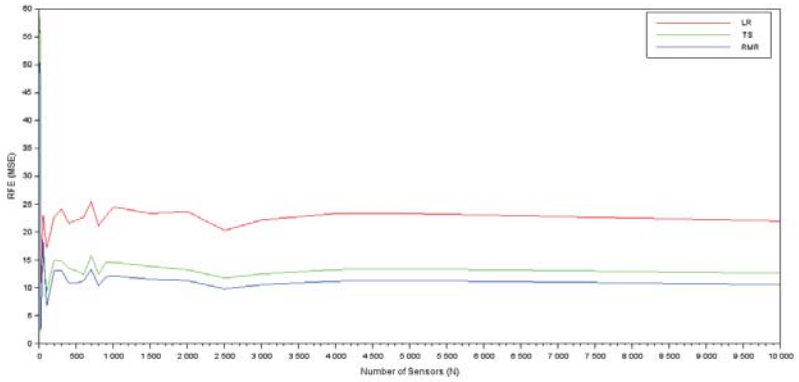

Fig. 4.8: Relative federated error (MSE) of exponential distribution

dian of pairwise slopes, RMR effectively nullified the influence of extreme data points. This characteristic made it exceptionally reliable under all three sensor configurations. As per the results plotted in the in the Figure 4.5 and Figure 4.6 exponential distribution setting where most regression models struggled with stability RMR consistently produced the most accurate and stable trend lines. Its performance advantage was especially clear in scenarios involving smaller sample sizes or non-uniform data spacing, highlighting its suitability for real-world systems where noise and data imbalance are common. From a computational perspective, RMR does have slightly higher overhead compared to linear regression and Theil-Sen, but the trade-off is justified by the accuracy gains and the increased resilience to corrupted data. In resource-constrained environments, Theil-Sen provides a middle ground, offering better robustness than linear regression with lower computational cost than RMR.

4.2.2 Effect of Sensor Configuration Distributions

The behavior of the regression models was further analyzed under three sensor deployment configurations: uniformly spaced sensors, normally distributed sensors (concentrated around a mean), and exponentially distributed sensors (densely packed in low-value regions with a long tail).

In the uniform distribution, which is shown by Figure 4.9 and Figure 4.10, all three regression methods showed comparable performance with minimal deviations. The evenly spaced sensors provided a balanced dataset that aligned well with the assumptions of standard regression techniques. Here, linear regression produced relatively reliable estimates as the absence of outliers minimized its weaknesses. Theil-Sen and RMR maintained their robustness, although their advantages were not as apparent due to the uniformity of the data.

The normal distribution introduced mild clustering around the center as shown by

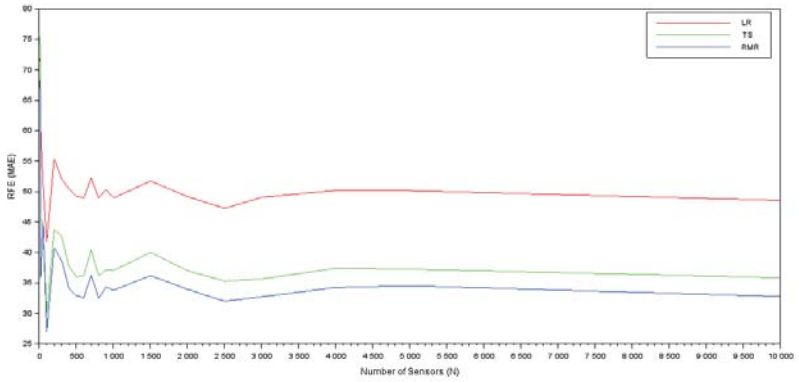

Fig. 4.9: Relative federated error (MAE) of uniform distribution method 2

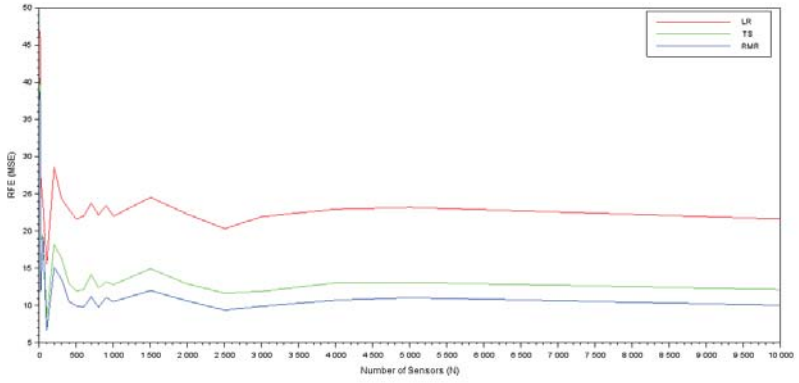

Fig. 4.10: Relative federated error (MSE) of uniform distribution method 2

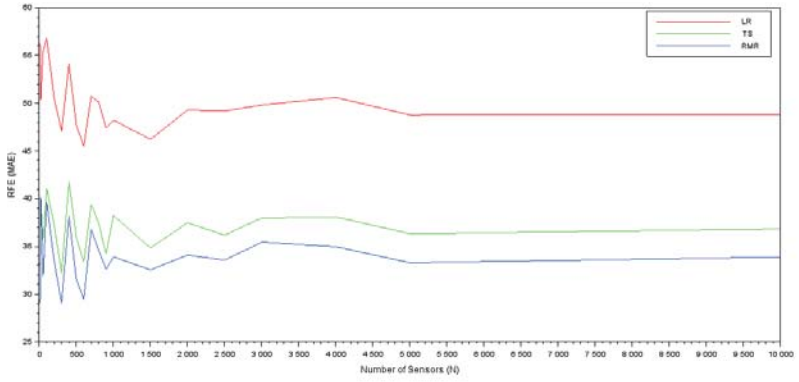

Fig. 4.11: Relative federated error (MAE) of Gausian distribution

Figure 4.9 and Figure 4.10, simulating real-world scenarios where sensors are denser in high-interest zones. Linear regression started showing slight deviations due to occasional high-value outliers. Theil-Sen demonstrated improved trend consistency and

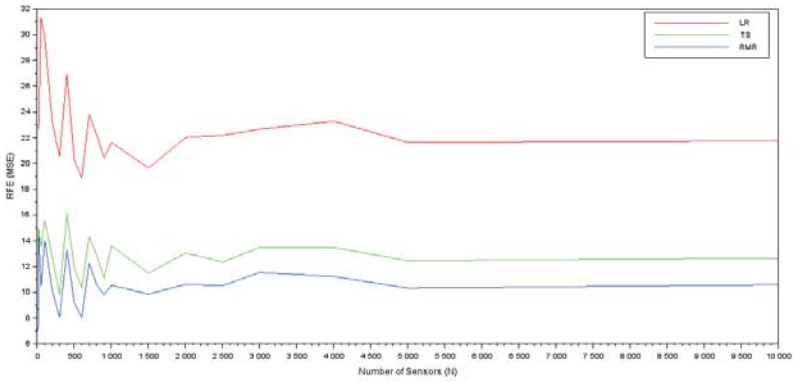

Fig. 4.12: Relative federated error (MSE) of Gausian distribution

was less influenced by these outliers. RMR maintained the most stable results, demonstrating a reliable pattern that converged well with increased sensor numbers.

The exponential distribution posed the greatest challenge due to the long-tailed nature of the data. A high density of sensors near the origin and sparse data in higher values caused regression lines to shift and stretch non-linearly. Linear regression, in this setting, struggled the most, producing unstable and erratic lines that failed to capture the central trend accurately. Theil-Sen fared better, although some instability remained at higher ranges. RMR, once again, outperformed the other methods by consistently modeling the trend line with the least sensitivity to extreme values.

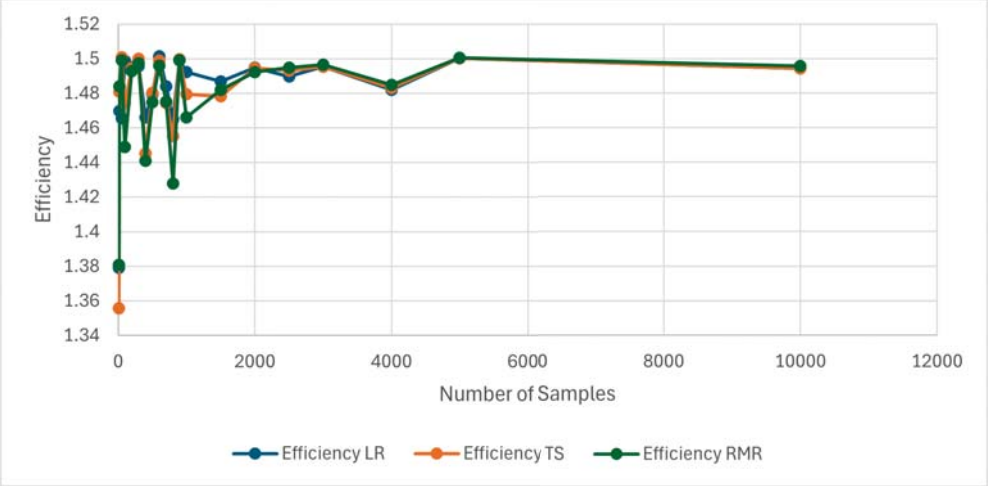

Fig. 4.13: Efficiency comparison for exponential method across sample sizes

This study further investigates how the number of participating sensors influences the trade-off between efficiency and predictive accuracy. As shown in Figure 4.13, the findings of exponential model indicate that model performance generally improves as the sensor count increases. Understanding this trend provides practical guidance

for identifying an approximate minimum sensor count required to achieve a desired accuracy level, thereby informing how to balance resource allocation with predictive performance in real-world federated learning deployments. Each method shows an initial improvement in accuracy and efficiency as the number of sensors increases. This consistent pattern across models further supports the insight that careful consideration of sample efficiency is essential when designing scalable federated learning systems.

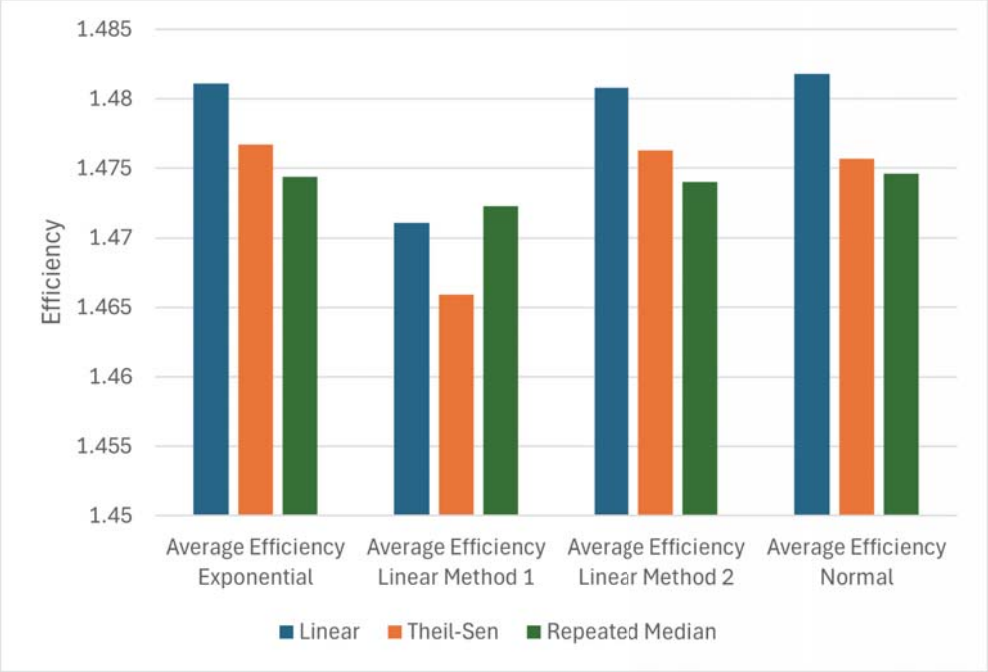

Fig. 4.14: Efficiency of federated regression models

As described in the methodology, efficiency is calculated across all sensor distributions, as shown in Figure 4.14. Here, lower values indicate better performance, as they reflect lower average errors and better model robustness. RFE mainly focuses on relative variability, and efficiency metrics focus on the outliers and the error distribution. Across all distribution types, Theil-Sen and RMR methods consistently yield lower efficiency values than standard linear regression, indicating superior robustness and predictive accuracy in federated model settings.

With the efficiency calculation data , a further percentage increment percentage is calculated, and the result is shown in figure 4.15. Percentage improvement is computed as the relative difference between a method and the Linear baseline, scaled by 100. The efficiency improvement analysis compares Theil-Sen and RMR methods against the baseline linear regression across four metrics. As per the results, RMR is the most robust method in noisy, federated settings.

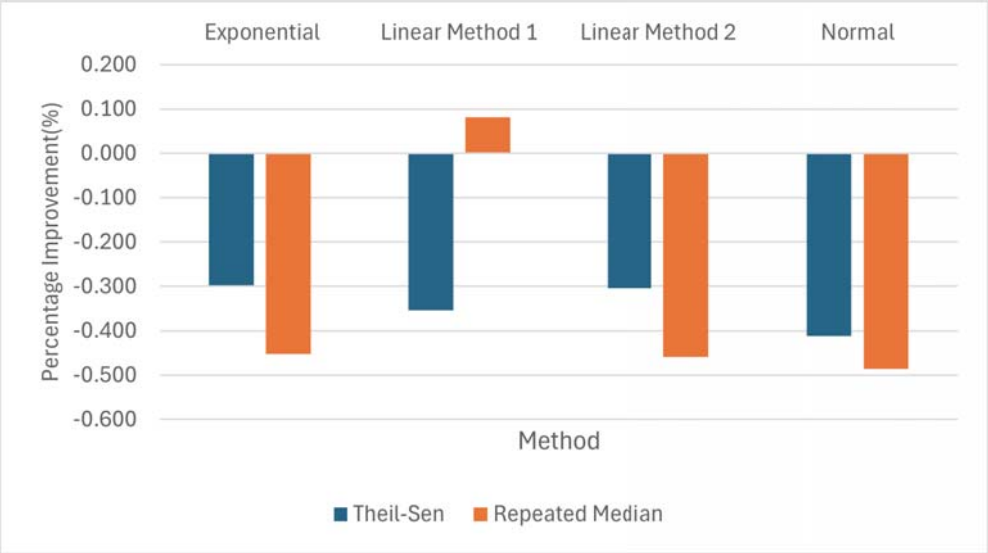

Fig. 4.15: Efficiency improvement with respect to Theil-Sen method

4.2.3 Impact of Sensor Quantity and Convergence Trends

Another important aspect explored in this study was the influence of the number of sensors on regression accuracy. As expected, all methods exhibited improved accuracy and convergence with increasing sensor count. However, the rate of convergence and the stability of the estimation varied significantly depending on both the regression technique and sensor distribution. Linear regression showed high sensitivity to small sample sizes, especially under exponential and skewed distributions, where even a single outlier could drastically alter the trend line. In contrast, both Theil-Sen and RMR demonstrated better convergence characteristics. RMR, in particular, converged faster and exhibited less fluctuation, even with a relatively low number of sensors. Interestingly, a saturation point was observed where adding more sensors beyond a certain threshold resulted in diminishing improvements. This asymptotic behavior suggests that after a certain number of sensors, the additional information becomes redundant. For uniformly distributed sensors, this threshold was reached earlier, while for exponential distributions, more sensors were needed to compensate for the sparse tail regions.

4.2.4 Selecting the Best Method

Based on empirical results, RMR stands out as the most effective approach across all distributions and sample sizes. Its high robustness, especially under noisy and skewed data, makes it ideal for real-world applications involving sensor networks. Theil-Sen offers a solid compromise between robustness and computational efficiency and is suitable where processing resources are limited. Linear Regression, while simplest and fastest, should be reserved for controlled environments where data is well-distributed.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This research aimed to evaluate and compare the performance of various regression methods. For the research scope selected linear regression Theil-Sen regression, and RMR methods with a FL context over generated sensor data. Through simulation experiments conducted using Scilab, a synthetic environment was created to mimic real-world sensor behavior. The sensor data was generated with intentional incorporation of noise, variance in slope and intercept, and non-linear distortion. Furthermore, allows for a thorough assessment of each regression technique's performance under different conditions.

A key takeaway of this study is the importance of robust regression techniques in federated settings. Linear regression has remained appealing due to its simplicity and low computational requirement. However, its performance degrades significantly with increasing data heterogeneity and sensor count continuously as per the results of this study. This was clearly reflected in the rising MSE and MAE values, particularly once the network scaled to hundreds or thousands of sensors. Theil-Sen and specially RMR maintained stable, low error rates even under highly noisy and non-linear conditions, proving their robustness.

TABLE 5.1: Comparison of Regression Methods

Method	**Strengths**	**Weaknesses**	**Best Use Cases**
Linear	Fast, simple, uses low computational resources	Sensitive to noise and outliers; weak under skewed data	Controlled environments with clean, well-behaved data
Theil-Sen	Robust to outliers, relatively efficient	May miss nonlinear patterns; slower than linear regression	Real-world applications where moderate robustness is needed
Repeated Median	Highly robust to noise; stable under data heterogeneity	Heavier computation; less ideal for linear-only patterns	Noisy, distributed sensor networks in federated learning

Efficiency analysis added another layer of insight beyond the error metric analysis. Efficiency, defined as the ratio of error metrics MSE and MAE of the method under test to those of an ideal baseline. Importantly, lower efficiency values indicate better performance and the higher efficiency values demonstrate a less performance. The results showed that RMR consistently achieved the lowest average efficiency scores,

particularly under exponential and normal distributions. Even the efficiency increment percentage metric revealed that robust regression methods yield measurable improvements over standard linear regression. In large federated systems, their benefits are notable even if their percentage increases in efficiency seem to be small. Minor improvements in dependability can result in significant advantages in overall model performance when performed across multiple distributed nodes.

The findings of the study also highlighted how the federated averaging process contributes positively to model generalization. Despite the local inconsistencies in model parameters due to distributed, noisy data, the aggregated global models closely approximated central trends properly. This supports the idea that FL can build accurate global models without centralized data sharing, and it is essential for privacy-preserving applications. However there are some tradeoffs in the well performing models as well. while robust methods like RMR provide superior accuracy and resilience, they come with greater computational demands. This is mainly relevant in real-world applications where edge devices operate under tight resource constraints. Theil-Sen regression, offering a good balance between computational efficiency and robustness and possible to use as a strong alternative where RMR is too costly to deploy.

In conclusion, the study demonstrates that RMR is the most reliable and efficient model for federated sensor networks, mainly under conditions such as high noise, outliers, and data skew. Theil-Sen regression is a viable compromise but offers improved performance over linear regression with moderate computational needs. Linear regression is fast but not performed accuralty in a environment where outliers and variations exists. These insights provide a strong empirical foundation for choosing regression methods in future deployments of large-scale federated sensor systems, where accuracy, robustness, and efficiency must all be carefully balanced.

Overall, this research work contributes not only to the methodological understanding of regression modeling in federated settings, but also paves the way for more scalable, resilient, and privacy-preserving data analysis frameworks for future distributed systems. Overall comaprison summary of regression methods are summarized by the table 5.1.

5.2 Future Work

Current research has successfully demonstrated and discussed the value of robust regression techniques in federated sensor environments. However, several directions remain open for future exploration, which are connected to different areas. A promising direction for future research lies in exploring heterogeneous data environments. In this heterogeneous data environments sensor nodes vary not only in noise characteristics but also in aspects such as feature distributions, sampling frequencies, and sensing technologies. Recognizing these variations can result in more resilient and applicable

federated regression models that can adjust to various real-world applications. Due to the heterogeneous behaviors of data, the collective approach needs to be implemented to find a correlation between each feature. Such kind of diversity reflects real-world deployments more accurately and presents unique challenges in model aggregation. Another important area to study is using adaptive federated aggregation methods. In this method of approach, each sensor's model is given a weight based on its trust level, reliability, or how effective it is on local data before aggregating it with others. This could lead to more personalized and context-aware global models. Furthermore, verifying these regression methods on actual data sets from fields like wearable health sensors or environmental monitoring would provide useful information and assist in the identification of limits not evident in synthetic simulations.

Another important scope that can further improve the current modeling framework is including time-based patterns in the data. This could be achieved using time-series models like federated auto regressive integrated moving average or recurrent neural networks. Improving efficiency in communication is another key area that can focused on due to the importance of energy reserving. Developing lighter, more efficient regression models would make the system better suited for low-power environments. Light and efficient regression models would help systems to perform robustly and save energy in end devices. Finally, integrating privacy-enhancing technologies such as differential privacy or secure multiparty computation would strengthen the confidentiality of sensor data in federated systems. This is a very important aspect for applications that handle sensitive personal or operational data.

REFERENCES

[1] E. T. M. Beltrán, M. Q. Pérez, P. M. S. Sánchez, S. L. Bernal, G. Bovet, M. G. Pérez, G. M. Pérez, and A. H. Celdrán, "Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2983–3013, 2023.

[2] A. Rauniyar, D. H. Hagos, D. Jha, J. E. Håkegård, U. Bagci, D. B. Rawat, and V. Vlassov, "Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions," *IEEE Internet of Things Journal*, vol. 11, no. 5, pp. 7374–7398, 2023.

[3] H. Cho, A. Mathur, and F. Kawsar, "Flame: Federated learning across multi-device environments," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–29, 2022.

[4] A. Ghosh, J. Hong, D. Yin, and K. Ramchandran, "Robust federated learning in a heterogeneous environment," *arXiv preprint arXiv:1906.06629*, 2019.

[5] Z. Liu, J. Guo, W. Yang, J. Fan, K.-Y. Lam, and J. Zhao, "Privacy-preserving aggregation in federated learning: A survey," *IEEE Transactions on Big Data*, 2022.

[6] J. Poushter *et al.*, "Smartphone ownership and internet usage continues to climb in emerging economies," *Pew research center*, vol. 22, no. 1, pp. 1–44, 2016.

[7] S. F. A. Zaidi, M. A. Shah, M. Kamran, Q. Javaid, and S. Zhang, "A survey on security for smartphone device," *International journal of advanced computer science and applications*, vol. 7, no. 4, 2016.

[8] M. Van Kleek, I. Liccardi, R. Binns, J. Zhao, D. J. Weitzner, and N. Shadbolt, "Better the devil you know: Exposing the data sharing practices of smartphone apps," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 5208–5220.

[9] B. Struminskaya, V. Toepoel, P. Lugtig, M. Haan, A. Luiten, and B. Schouten, "Understanding willingness to share smartphone-sensor data," *Public Opinion Quarterly*, vol. 84, no. 3, pp. 725–759, 2020.

[10] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106775, 2021.

[11] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani, "Reliable federated learning for mobile networks," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 72–80, 2020.

[12] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[13] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[14] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *International conference on machine learning*. PMLR, 2021, pp. 6357–6368.

[15] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[16] S. Vahidian, M. Morafah, C. Chen, M. Shah, and B. Lin, "Rethinking data heterogeneity in federated learning: Introducing a new notion and standard benchmarks," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 3, pp. 1386–1397, 2023.

[17] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[18] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018.

[19] M. Duan, D. Liu, X. Chen, R. Liu, Y. Tan, and L. Liang, "Self-balancing federated learning with global imbalanced data in mobile systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 1, pp. 59–71, 2020.

[20] I. Jeon, M. Hong, J. Yun, and G. Kim, "Federated learning via meta-variational dropout," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[21] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," *arXiv preprint arXiv:2002.07948*, 2020.

[22] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 8, pp. 3710–3722, 2020.

[23] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 586–19 597, 2020.

[24] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," *arXiv preprint arXiv:2003.00295*, 2020.

[25] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International conference on machine learning*. PMLR, 2019, pp. 4615–4625.

[26] M. Padala and S. Gujar, "Fnnc: Achieving fairness through neural networks," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,{IJCAI-20}, International Joint Conferences on Artificial Intelligence Organization*, 2020.

[27] E. Nunez, E. W. Steyerberg, and J. Nunez, "Regression modeling strategies," *Revista Española de Cardiología (English Edition)*, vol. 64, no. 6, pp. 501–507, 2011.

[28] J. Friedman, "The elements of statistical learning: Data mining, inference, and prediction," *(No Title)*, 2009.

[29] G. James, D. Witten, T. Hastie, R. Tibshirani *et al.*, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[30] W. Greene, "H.(2012): Econometric analysis," *Journal of Boston: Pearson Education*, pp. 803–806, 2012.

[31] J. H. Stock and M. W. Watson, "Introduction to econometrics (3rd updated edition)," *Age (X3)*, vol. 3, no. 0.22, 2015.

[32] G. King and L. Zeng, "Logistic regression in rare events data," *Political analysis*, vol. 9, no. 2, pp. 137–163, 2001.

[33] W. A. Fuller, *Measurement error models*. John Wiley & Sons, 2009.

[34] R. Blundell and J. L. Powell, "Endogeneity in nonparametric and semiparametric regression models," *Econometric society monographs*, vol. 36, pp. 312–357, 2003.

[35] J. D. Angrist and J.-S. Pischke, *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

[36] E.-M. Brinkmann, M. Burger, J. Rasch, and C. Sutour, "Bias reduction in variational regularization," *Journal of Mathematical Imaging and Vision*, vol. 59, no. 3, pp. 534–566, 2017.

[37] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[38] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[39] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[40] R. Koenker and G. Bassett Jr, "Regression quantiles," *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

[41] R. Wilcox, *Modern statistics for the social and behavioral sciences: A practical introduction*. Chapman and Hall/CRC, 2017.

[42] J. D. Angrist and A. B. Krueger, "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics*, vol. 106, no. 4, pp. 979–1014, 1991.

[43] J. M. Wooldridge, *Econometric analysis of cross section and panel data*. MIT press, 2010.

[44] R. J. Little, "Post-stratification: a modeler's perspective," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 1001–1012, 1993.

[45] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[46] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompatsiaris, "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 853–862.

[47] A. Agarwal, M. Dudík, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *International Conference on Machine Learning*. PMLR, 2019, pp. 120–129.

[48] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *2011 IEEE 11th international conference on data mining workshops*. IEEE, 2011, pp. 643–650.

[49] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 3, p. e1452, 2022.

[50] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[51] R. D. Cook, "Detection of influential observation in linear regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, 1977.

[52] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 492–518.

[53] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. John wiley & sons, 2003.

[54] B. Luo, W. Xiao, S. Wang, J. Huang, and L. Tassiulas, "Tackling system and statistical heterogeneity for federated learning with adaptive client sampling," in *IEEE INFOCOM 2022-IEEE conference on computer communications*. IEEE, 2022, pp. 1739–1748.

[55] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International conference on machine learning*. Pmlr, 2018, pp. 5650–5659.

[56] J.-J. Lee, B. Krishnamachari, and C.-C. Kuo, "Impact of heterogeneous deployment on lifetime sensing coverage in sensor networks," in *2004 First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON 2004.* IEEE, 2004, pp. 367–376.

[57] G. Zhou, T. He, S. Krishnamurthy, and J. A. Stankovic, "Models and solutions for radio irregularity in wireless sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 2, no. 2, pp. 221–262, 2006.

[58] W. Sun, F. Brännström, and E. G. Ström, "On clock offset and skew estimation with exponentially distributed delays," in *2013 IEEE International Conference on Communications (ICC)*. IEEE, 2013, pp. 1872–1877.

[59] I. Lavagnini, D. Badocco, P. Pastore, and F. Magno, "Theil–sen nonparametric regression technique on univariate calibration, inverse regression and detection limits," *Talanta*, vol. 87, pp. 180–188, 2011.