Comprehensive Study on Estimating S-Wave Intensity in New Zealand Earthquakes: A Mixed Effect Modelling Approach

By

E. V. T. Eranthi

The dissertation is submitted as a part of the M.Sc. in Applied Statistics

Department of Statistics,
University of Colombo, Sri Lanka

August 2024

Declaration

This thesis is my original work and has not been submitted previously for a degree at this or any other university/institute. To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.

Author's name: E. V. T. Eranthi Date…………………………

Signature ………………………..

This is to certify that this dissertation is based on the work carried out by Mrs. E. V. T. Eranthi under my supervision. The dissertation has been prepared according to the format stipulated and is of an acceptable standard.

Supervisor: Dr. (Mr.) G. P. Lakraj Date…………………

Senior Lecturer

Department Of Statistics Signature ………………

University of Colombo

Coordinator: Dr. (Mrs.) K. A. D. Deshani Date…………………

Senior Lecturer

Department Of Statistics Signature ………………

University of Colombo

Acknowledgements

This research project has been the most challenging yet the most interesting project of my academic career. It would not have been possible without the encouragement and kind assistance from many helping hands.

First and foremost, I extend my deepest gratitude to my supervisor, Dr. (Mr.) G. P. Lakraj. His patient guidance, clear explanations, and encouragement to critically analyse have been invaluable. His support not only helped me navigate this project but also broadened my interest in the field of statistics and encouraged me to pursue publications. I truly respect and am indeed grateful for his wonderful supervision.

I am profoundly grateful to Dr. (Mrs.) K.A.D. Deshani, Coordinator of the MSc in Applied Statistics 2020 program, for her unwavering help and guidance over the last few years. She has always been there for us, regardless of the challenges, even beyond regular working hours. Her dedication is deeply appreciated.

I am grateful to our department head, Dr. Rushan A. B. Abeygunawardana, and extend my sincere appreciation to all the lecturers in the MSc in Applied Statistics 2020 program. The foundation of my achievements is built upon their instruction and support. I offer special thanks to the lecturers who have guided and assisted me throughout these past few years.

I am also thankful to Mr. Chanthujan Chandrakumar, Doctoral Researcher at the Joint Centre for Disaster Research, Massey University, New Zealand. His assistance in helping me understand seismology and earthquake science related to New Zealand earthquakes was invaluable.

Last but not least, I convey my greatest gratitude and deepest appreciation to my husband and my mother for their unwavering support, encouragement, and motivation throughout all these years. I am forever grateful for their countless commitments and endeavors.

Thank you to everyone who contributed to the success of this research project.

Abstract

Earthquake Early Warning Systems (EEWS) are essential in seismically active regions like New Zealand, where timely alerts can mitigate potential damage and save lives. The objective of this study is to identify the most appropriate predictors for estimating S-wave intensity and to develop a robust nationwide model that accounts for data complexities. Estimating S-wave intensity is critical for predicting ground shaking and potential damage. This research specifically explores the role of P-wave parameters, station characteristics, earthquake ID, and year index as random effects in enhancing the model's accuracy.

This study initially analyzed 54,728 earthquake waveforms from 12,502 events in New Zealand (2013–2022). After filtering for waveforms within 100 km of the epicenter and removing outliers, the dataset was refined to 29,057 waveforms from 9,206 earthquakes recorded by 293 stations. The dataset includes records of earthquakes with magnitudes greater than 3.0 on the Richter scale. To handle the non-independence of these data due to repeated measurements from the same station and event, a Linear Mixed-Effects Model (LMM) was adopted, incorporating random effects for stations, earthquake events, and years. The choice of these random effects is justified by the observed variability in ground motion across different stations and times.

Log-transformed values were used for both the P-wave peak ground velocity (Pv) and S-wave peak ground velocity (PGV) to address skewed data distributions and improve model stability by standardizing variable ranges, making relationships more linear and reducing outlier impact. This transformation enhanced the model's ability to capture trends and patterns, improving its reliability and interpretability. Correlation analysis confirmed that log(Pv) was the most reliable predictor for log(PGV) compared to other P-wave parameters. The distribution of random effects revealed significant variance between stations (Intra-class correlation coefficient, ICC = 0.459), earthquake events (ICC = 0.610), and years (ICC = 0.162), validating the inclusion of these factors as random effects.

The LMM demonstrated superior model performance (conditional $R^2 = 0.826$, marginal $R^2 = 0.718$) compared to traditional regression models, reducing Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The results demonstrate that the LMM provides a better fit and applying this model at the station level can enhance the accuracy of ground-shaking intensity predictions, potentially improving the reliability of warning systems and public safety.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIC	- Akaike information criterion
BIC	- Bayesian information criterion
CAV	- Cumulative absolute velocity
CNN	- Convolutional neural network
EEWS	- Earthquake early warning systems
ES	- Earthquake source
FDSN	- The international federation of digital seismograph networks
GNS	- Institute of geological and nuclear sciences limited
ICC	- Intra-class correlation
IQR	- Interquartile range
LMM	- Linear mixed-effect modelling
LRT	- Likelihood ratio test
MAPE	- Mean absolute percentage error
MLP	- Multilayer perceptron classifier
MMI	- Modified mercalli intensity
MSE	- Mean square error
NEMA	- National emergency management agency
OLS	- Ordinary least squares
Pa	- Peak ground acceleration of P-wave
Pd	- Peak ground displacement of P-wave
PGA	- Peak ground acceleration of S-wave
PGD	- Peak ground displacement of S-wave
PGV	- Peak ground velocity of S-wave
Pv	- Peak ground velocity of P-wave
P-waves	- Primary or Pressure waves
RMSE	- Root mean square error
STEAD	- Stanford earthquake data set
SVM	- Support vector machine
S-waves	- Secondary or Shear waves

1. Introduction

Earthquakes present a formidable natural hazard, especially for regions situated along active fault lines (Stoppa & Berti, 2013). The sudden onset of seismic events, often with little to no prior warning, poses significant challenges for both the public and authorities responsible for disaster response and mitigation(Mousavi et al., 2020). As technology advances, the development of robust Earthquake Early Warning Systems (EEWS) has become a critical focus for enhancing preparedness and reducing the impact of earthquakes. These systems, designed to provide alerts seconds before the arrival of destructive seismic waves, can significantly improve response times and potentially save lives by enabling timely protective actions(Allen & Melgar, 2019).

New Zealand experiences frequent seismic activities, making the implementation of an effective EEWS particularly crucial. Despite the country's advanced seismic monitoring capabilities, including the GeoNet network, a nationwide EEWS has not yet been established(Becker et al., 2020). The current warning systems are primarily focused on post-event response rather than providing immediate alerts. Given New Zealand's unique geological setting and the diverse range of seismic events, conducting research on reliable EEW methodologies is imperative(Vinnell et al., 2023).

A key component of an effective EEWS is the accurate estimation of earthquake intensity of the location of monitoring station(Sarkar et al., 2022a). Accurately estimating earthquake intensity, which measures the severity of ground shaking, is essential for assessing potential damage and issuing timely warnings. Traditional methods often rely on empirical models or ground motion prediction equations, which can be limited by data availability and quality(Hsu & Huang, 2021a). Recent advancements in Machine Learning offer promising new approaches for intensity estimation, leveraging large datasets and sophisticated algorithms to enhance predictive accuracy(Sarkar et al., 2022a). However, these models frequently encounter challenges related to station variability, which can impact their generalizability and reliability(Kuehn & Scherbaum, 2015).

The research presented in this thesis aims to address these challenges by developing a generalized earthquake intensity estimation model for New Zealand. By utilizing a comprehensive dataset from GeoNet(GeoNet, n.d.), which includes over 54,000 earthquake waveforms from more than 12,000 events, this study seeks to overcome the limitations of single-station and region-specific models. The proposed approach employs Linear Mixed-Effects Modelling (LMM) to account for data non-independence and station-specific variations, enabling the creation of a robust, generalized model applicable across multiple stations nationwide.

1.1 Research gap

1. Single-Station focus: Numerous studies have concentrated on single-station data, emphasizing the significance of station-specific characteristics in estimating earthquake intensity(Hsu & Pratomo, 2022). However, building separate models for each station is impractical due to the limited availability of station-specific data in New Zealand. This lack of extensive data for individual stations hampers the development of accurate and reliable intensity estimation models tailored to each station(Abdalzaher et al., 2024).
2. Regional limitations: Existing research on earthquake intensity estimation in New Zealand is sparse and primarily region-specific. These studies often face limitations such as data non-independence and the failure to account for the nested structure of seismic data. The regional focus of these studies restricts their applicability to a broader, nationwide context(Chandrakumar et al., 2024).
3. Generalized model deficiency: There is a notable absence of a comprehensive, generalized intensity estimation model that can be applied across different stations throughout New Zealand. Current methodologies do not adequately address the need for a model that considers station variability and can utilize historical data from multiple stations(Fayaz & Galasso, 2023).
4. Technological gaps/ Methodological gap: The lack of advanced, robust models capable of integrating diverse seismic data from various stations underscores a significant technological gap. Existing approaches often rely on complex neural networks tailored to specific datasets, which are not designed for generalized application across New Zealand(Sarkar et al., 2022a),(Abdalzaher et al., 2023a).

1.2 Research objectives

1. Determine the best factors for accurately estimating S-wave intensity.
2. Explore the impact of various parameters (P-wave parameters, station characteristics, earthquake ID, and year index) on S-wave intensity and develop a generalized earthquake intensity estimation model for predicting S-wave intensity at any station in New Zealand.

1.3 Research significance

The proposed research addresses existing gaps by leveraging a comprehensive dataset from GeoNet, encompassing over 54,000 earthquake waveforms from more than 12,000 events recorded across multiple stations in New Zealand. By applying Linear Mixed-Effects Modelling, this study effectively handles data non-independence and station-specific variations, facilitating the development of a robust, generalized intensity estimation model.

This model has the potential to significantly enhance the effectiveness of EEWS in New Zealand by providing accurate intensity estimation. Beyond filling a critical gap in New Zealand's earthquake intensity estimation, this research also advances the broader field of earthquake warning. By introducing a generalized model approach that can be applied at any earthquake station, this research opens new avenues for future studies aimed at enhancing earthquake warning systems.

Integrating advanced statistical techniques with comprehensive seismic data, this study aims to significantly contribute to earthquake impact reduction efforts. It can improve disaster preparedness and mitigation, reducing the impact of earthquakes on communities and infrastructure through quicker and more reliable warnings.

2. Literature Review

Throughout human history, disasters have consistently caused havoc on both people and critical infrastructure(Mousavi et al., 2020). Among these hazards, earthquakes stand out as the most harmful, especially in regions near active fault lines on land or offshore subduction zones(Stoppa & Berti, 2013). Unlike other natural hazards like cyclones or tsunamis, earthquakes remain unpredictable hours in advance, with detection only possible during the actual seismic event, as earthquake alerts are generated within seconds. Therefore, having a robust earthquake early warning system (EEWS) is of paramount importance given the nature of this event. However, this inherent unpredictability of the event and the limited time available between earthquake occurrence and the subsequent destructive impact poses a substantial challenge in the development of EEWS (Xia et al., 2021).

Despite the considerable challenges involved, EEWS have emerged as valuable tools for alerting both the public and authorities to take appropriate safety measures during an earthquake(Allen & Melgar, 2019). Offering critical advance warnings to regions susceptible to significant ground shaking, EEWS provides a brief yet crucial window for individuals to undertake simple yet potentially life-saving actions, such as the "drop-cover-hold" protocol, and mentally prepare for an impending earthquake(Nakayachi et al., 2019). Additionally, these precious seconds allow automated systems to initiate emergency measures(Strauss & Allen, 2016). In the early stages of EEWS development, the complexity associated with earthquake-related data processing posed substantial challenges, making the generation of reliable alerts a formidable task (Böse et al., 2008). However, in the contemporary context, substantial technological progress in seismic instrumentation, data processing, digital communication and algorithmic capabilities has opened the way for the implementation of a robust and reliable EEWS(Kanamori et al., 1997).

The operation of EEWSs relies on two fundamental principles. First, the transmission of information across communication networks is significantly faster than the propagation of seismic waves, including P-waves (primary or pressure waves) and S-waves (secondary or shear waves)(Festa et al., 2022). Second, it is important to note that the most significant damage during an earthquake occurs with S-waves, which arrive later than the initial P-waves(Hsu & Nieh, 2020a). The duration of the early warning window for earthquakes can vary from just a few seconds to several tens of seconds, based on factors such as the geometry of the specific earthquake and the design of the sensor stations integrated into the EEWS(Wu et al., 2007). In practical terms, EEWS utilise a single sensor, or a network of sensors distributed across a defined geographical area to swiftly detect earthquake activity and relay real-time alerts(Peng et al., 2019).

Depending on the geography of the spread of an EEWS, they can be categorised as On-site and Regional EEWs(Caruso et al., 2017). The regional warning system relies on abundant station information to provide a more accurate estimate of the earthquake source (ES) parameters. However, this system typically requires a longer processing time and cannot provide timely warnings in areas close to the epicentre. Conversely, the on-site warning system uses the initial part of the P wave waveforms observed by a single or few adjacent stations to predict subsequent intensity measurement values at the same site. Therefore, the on-site warning system is commonly deployed for critical targets, with surrounding stations used to predict the intensity of the impending seismic waves. Due to its rapid warning capability, it is more suitable to provide early warnings to the area around the epicentre(Zollo et al., 2010).

Earthquake intensity measures the strength of ground shaking during an earthquake, indicating the seismic event's "power level" and its impact on ground movement(Wu et al., 2003). This intensity is closely linked to ground motion factors such as peak ground acceleration (PGA), peak ground velocity (PGV) and Peak Ground Displacement (PGD), rather than earthquake source parameters like earthquake magnitude in different scales(Wu et al., 2004). Earthquake intensity is often measured by the Modified Mercalli Intensity (MMI) Scale(Wald et al., 1999). At the lower levels of the scale, the intensity is generally assessed in terms of how the shaking is felt by people. Higher levels of the scale are based on observed structural damage surveyed by professionals(Wu et al., 2003). In this research, we used Peak Ground Velocity (PGV) as the most suitable parameter for measuring S-wave intensity, which is discussed in detail in section 3.3. PGV provides critical information about the maximum velocity of ground motion during an earthquake, making it particularly relevant for assessing the intensity of S-wave effects.

For onsite EEWS, accurately estimating intensity is crucial for quick assessments of shaking levels, enabling timely alerts to protect people and critical structures. Studies highlight the significance of determining earthquake intensity from initial P-wave signals to predict damaging S-wave amplitudes before their arrival, providing valuable insights for EEW systems and seismic hazard assessment (Ahmadzadeh et al., 2020) (Chandrakumar et al., 2024). Additionally, accurate intensity estimation enhances the reliability of warnings by reducing unnecessary panic caused by false alarms.

Various methods are employed for earthquake intensity estimation, including ground motion prediction equations, empirical relationships, and machine learning techniques. Ground motion prediction equations, such as those based on spectral acceleration equations, provide a framework for estimating intensity by analyzing ground motion characteristics like spectral accelerations and spectrum intensities(Bradley et al., 2009) (Jayaram & Baker, 2009). Empirical relationships utilize past earthquake data to establish correlations between ground motion parameters and intensity measures, offering insights into the expected shaking levels. Machine learning approaches, such as support vector machines, neural networks, and random forests, leverage computational algorithms to estimate earthquake intensity based on input P-wave data, enhancing the effectiveness of onsite EEWS.

To provide a comprehensive overview of onsite earthquake intensity estimation studies conducted worldwide, the following table 2.1 summarizes various research efforts that utilize diverse models, technique, country region, method and comparison with this research. These studies explore different parameters and approaches to predict seismic intensity, contributing valuable insights into enhancing EEWS.

Table 2.1 : Overview of earthquake intensity estimation research world-wide: models, approaches, regions, and techniques

Research Title	**Technique**	**Region**	**Method**	**Comparison with this Research**
Seismic Intensity Estimation for Earthquake Early Warning Using Optimized Machine Learning Model (Abdalzaher et al., 2023)	Extreme Gradient Boosting	Italy	The paper emphasizes rapid intensity determination methodology using deep neural networks (within 2 seconds) for an EEW system using Italian data.	While this research focus on immediate, localized intensity prediction within 2 seconds using Italian data, it does not address the generalization across various stations.

Development of On-Site Earthquake Early Warning System for Taiwan(J. et al., 2012)	Artificial neural network (ANN)	Taiwan	This research focuses on a smartphone-based crowdsourced EEW system with PGA predictions.	While this research aims for rapid, localized earthquake early warning with immediate alerts based on Taiwanese data, it does not address the development of a generalized model for multiple stations nationwide.
Multilayer Perceptron Based Early On-Site Estimation of PGA During an Earthquake(Sarkar, Kumar, et al., 2023)	Multilayer perceptron (MLP) neural network	Japan	Use the initial three-component p-wave features. These features are extracted from a one-second-long accelerogram sampled from 0 to 7 seconds after the p-wave arrival. The features are used to train a multilayer perceptron (MLP) neural network for estimating on-site PGA.	Sarkar et al. use a multilayer perceptron to estimate PGA from p-wave data at individual sites for immediate local predictions, but do not address generalization across broader regional applicability
On-Site Earthquake Early Warning Using Smartphones(Hsu & Nieh, 2020)	Artificial neural network	Taiwan	Use the interquartile range (IQR) between the 25th and 75th percentile of the acceleration vector sum of the three-component acceleration, the zero-crossing rate from the component with the highest value (ZC), and the cumulative absolute velocity (CAV) of the acceleration vector sum of the 3-component acceleration.	This approach uses smartphones and neural networks for local earthquake detection and PGA prediction but does not account for station-specific factors

			These parameters are used to train an ANN classifier to distinguish between earthquakes and human activities.	
Neural Network-Based Strong Motion Prediction for On-Site Earthquake Early Warning(Chiang et al., 2022)	Convolutional neural network (CNN)	Taiwan	Use peak ground displacement (Pd) from the initial P-wave, the period parameter of the initial portion τc of the P-wave, and the PGA of the seismic waves. This model is employed to extract relevant features from the initial P-waves and predict whether the peak ground acceleration of subsequent waves surpasses a pre-selected threshold.	The model focuses on predicting strong ground motion at individual sites using neural networks and initial P-wave signals for immediate, localized early warning, without addressing generalization across multiple stations.
Neuroevolution-Based Earthquake Intensity Classification for Onsite Earthquake Early Warning (Sarkar, Roy, et al., 2023)	Multilayer Perceptron neural network	Japan	The parameters used to estimate earthquake intensity include the number of input features, the number of hidden layers, the number of perceptron in hidden layers, the number of output neurons, and the output classes of warning.	This method enhances onsite early warning using neuroevolutionary techniques for intensity classification based on PGA, lacking broader applicability across multiple stations. Also consumes earthquake magnitude factor in the model.

Earthquake Early Warning Starting From 3 s of Records on a Single Station with Machine Learning (Lara et al., 2023)	Extreme Gradient Boosting (XGB)	Global dataset (STEAD, Clile, Japan)	Use magnitude, distance, depth, and back azimuth. These are estimated using time windows that contain 7 seconds of noise and 3 seconds of P-wave signal extracted from the earthquake database. The P-wave signal is preprocessed and filtered using a fourth-order Butterworth band-pass filter from 1 to 45 Hz. The estimation models are trained independently, and the source characterization algorithm uses these parameters to forecast ground shaking intensity measures such as PGA.	This model is optimized for real-time early warning at single stations using P-wave data for quick magnitude and source estimation but does not address the generalization across multiple stations.
Support Vector Machine-Based On-Site Prediction for China Seismic Instrumental Intensity from P-Wave Features(Hou et al.)	Support vector machine (SVM)	China	The parameters used to estimate earthquake intensity include PGA, PGV, and the evolution of PGA. These parameters are used to calculate the seismic instrumental intensity, which is a representative parameter for the application of EEWS systems.	This model focuses on predicting PGV regional applications but does not consider station specific factors or data across multiple stations nationwide. Also consumes earthquake magnitude factor in the model.
Seismic Intensity Estimation Using Multilayer Perceptron for Onsite Earthquake	Multilayer Perceptron classifier (MLP)	India	The parameters used to estimate earthquake intensity include PGA, PGV, and PGD. These parameters are directly related to ground	This model predicts PGA based on initial seismic data but does not address broader regional applicability or integration

Early Warning (Sarkar et al., 2022)			motion and are used to assess the severity of shaking at a location and the probability of damage.	of historical data across multiple stations.
Onsite Early Prediction of PGA Using CNN With Multi-Scale and Multi-Domain P-Waves as Input(Hsu & Huang, 2021)	Convolutional neural network (CNN)	Taiwan	Use magnitude, predominant frequency of P wave, fitting parameter of the waveform envelope, P-wave amplitude, peak displacement amplitude of the P wave, PGV, predominant period of seismic waves, and the inner product of acceleration and velocity. These parameters are used to establish empirical functions between the extracted P-wave parameters and the source parameters or seismic intensity.	This model predicts PGA from initial P-wave data at a single site but does not address broader regional applicability or station-specific data limitations

Here, almost all on-site intensity estimation methods in the world rely on complex neural network techniques to estimate earthquake intensity, but these models are not designed for site-specific applications.

2.1 Importance of earthquake intensity estimation model for New Zealand

New Zealand, situated within the Pacific Ring of Fire, experiences frequent seismic activity due to its position on the boundary of the Pacific and Australian tectonic plates (*GeoNet Recent Quakes*, n.d.). Approximately 14-15,000 occur in and around the country each year. Most earthquakes are too small to be noticed, but between 150 and 200 are large enough to be felt (Learn NZ, n.d.). The country is susceptible to various types of earthquakes, ranging from shallow crustal events to deeper subduction zone quakes. Notably, the Alpine Fault on the South Island and the Wellington Fault near the capital are continuously monitored due to their potential for significant seismic events (GNS Science, n.d.). The Canterbury region, particularly Christchurch, suffered profound devastation from the destructive earthquake of 2011, which claimed 185 lives and caused widespread damage to infrastructure (National Emerency Management Agency, n.d.). Despite advances in seismic monitoring and stringent building regulations, New Zealand remains at risk of earthquakes, necessitating ongoing research, preparedness efforts, and community resilience initiatives nationwide(www.building.govt.nz, n.d.).

However, New Zealand currently lacks a nationwide EEWS provided by an authorized warning agency (Becker et al., 2020b). While GeoNet monitors earthquakes in real-time, it does not offer EEW. The National Emergency Management Agency (NEMA) handles national-level warnings, excluding earthquakes, with local civil defence groups managing community alerts. Other agencies like GNS Science and MetService issue alerts for geological and meteorological hazards. Despite the absence of a nationwide EEWS, there have been a few studies in New Zealand focused on earthquake intensity estimation. One study, "A Novel Method for Predicting Local Site Amplification Factors Using 1-D Convolutional Neural Networks," focused on the Lower Hutt area (Yang et al., 2021). However, this study uses earthquake source parameters like magnitude as inputs, making it unsuitable for EEW scenarios. Another study estimated S-wave amplitude using a simple linear regression model based on data from the Canterbury region(Chandrakumar et al., 2024). This method is not suitable for non-independent data and does not account for station variability. Both studies are region-specific and do not cover the entire New Zealand area.

3. About Data and Data preprocessing

3.1 Data collection

The data was collected from the publicly available GeoNet FDSN web service by GNS Science (GeoNet, n.d.). Several steps were required to obtain the final data set. Python scripts, supported by the ObsPy library, facilitated the data download process. All codes are available in Appendix 2.

1. Initially, a list of events occurring between 2013 and 2022 was obtained using the event service. The data was then downloaded in station view as earthquake catalogs. Due to the large size of the requests, the data had to be downloaded in chunks. Each catalog is defined by an earthquake magnitude range, iterating from 3 to 8 in increments of 0.1, resulting in 50 sub-catalogs.

2. Once this was completed, all the data was compiled into a CSV file. This file included information such as the earthquake ID, event start time, station name, P-wave pick time, and S-wave pick time. Data entries lacking either P or S-wave pick times were excluded, as these wave picks were made by GeoNet itself. Initially, the file contained 65,124 rows, but after removing these missing values, the number was reduced to 61,245.

3. The next step was to download the complete waveform data. The previously created CSV file from the second step served as the base event dataset. Using the FDSN station view, raw acceleration data from sensor stations was downloaded. A 90-second waveform window was used for the data, capturing 30 seconds before and 60 seconds after the wave start time. The downloaded data was saved in the MiniSEED (mseed) format, a standard for storing acceleration data. The waveform data itself contains other metadata such as epicentral distance and magnitude.

4. The next goal was to identify the peak ground acceleration, peak ground velocity, and peak ground displacement values in the S and P waveform data. The calculation methods of these parameters are explained in the section "3.3 Description of the Parameters Used."

Data from one station (TUVZ) had to be removed due to corruption, resulting in a final dataset of 54,728 rows. All code used for data retrieval and preprocessing is included in Appendix 2.

3.2 About data

In summary, recordings spanning nine years, from 2013 to 2022, were extracted to construct the final dataset. This dataset comprises 54,728 earthquake waveforms recorded by various stations, documenting 12,502 earthquakes. Only waveform data associated with earthquake magnitudes exceeding 3M (on the Richter scale) are included. For this study, 13 variables aligned with the research objectives were selected. Erroneous records lacking identified P and S-wave picks were excluded from the dataset.

Table 3.1: Considered variables from the GeoNet Database

Numerical	Categorical	DateTime
Pa	EarthquakeKey	P_Wave Pick Time
Pv } P-wave parameters	EarthquakeID	S_Wave Pick Time
Pd	Station Name	
PGA		
PGV } S- wave parameters		
PGD		
Epicentral Distance		
Magnitude		

The following section provides an in-depth explanation of the data retrieval and parameter calculation processes.

3.3 Description of the parameters used

In this research, relationship between P-wave parameters and S-wave intensity were established using computed seismic parameters. Calculations for estimating the intensity characteristics of the P-wave window that has been selected involve determining the Peak Ground Acceleration

(Pa), Peak Ground Velocity (Pv), and Peak Ground Displacement (Pd). The calculations were performed using peak data on vertical acceleration, velocity, and displacement, as P-waves primarily generate motion in the vertical direction(Wu et al., 2019).

Next, the peak ground acceleration (PGA), peak ground velocity (PGV), and peak ground displacement (PGD) of the S-waves were computed for the S-wave window to evaluate their intensity. Since S-waves predominantly exhibit motion in the horizontal direction, we used data from the HNE (east-west direction) and HNN (north-south direction) channels to capture these peak values(Shearer, 2009). Some researchers are utilizing the maximum value of the vector norm derived from the directional channels of the S-wave yields optimal estimates for PGA, PGV and PGD for S-waves.

The most suitable parameter for measuring S-wave intensity among PGA, PGV and PGD is Peak Ground Velocity (PGV). PGV provides information on the maximum velocity of ground motion during an earthquake, which can be particularly relevant for assessing the intensity of S-wave effects. Research by has highlighted the spatial variability of PGV, indicating its sensitivity to ground motion variations(Johnson et al., 2020). This variability can offer insights into the dynamic behavior of S-waves and their impact on structures. Furthermore, PGV has been shown to correlate better with MMI levels compared to PGA, especially for high-intensity events(Grandin et al., 2007). This correlation underscores the effectiveness of PGV in capturing the intensity of ground shaking experienced during earthquakes, particularly in relation to human perception and structural response. Additionally, PGV can be a stable candidate for ground motion intensity measures in seismic assessment methods, as suggested by Akkar & Özen (Akkar & Özen, 2005).

Moreover, PGV is a direct and effective parameter for determining earthquake intensity(Idha et al., 2023). Its ability to provide real-time information on ground motion characteristics, along with its inclusion in earthquake early warning systems, showcases the practical utility of PGV in seismic monitoring and hazard mitigation efforts(Wu & Mittal, 2021).

In this study, "Station Name" refers to the unique identifier assigned to each seismic station within the monitoring network. These stations are equipped with instruments that record seismic activity, capturing ground motions during earthquakes. It is important to note that the geotechnical properties of the site can influence the amplification or attenuation of seismic waves. Consequently, the location of each station can significantly impact the intensity and characteristics of the seismic shaking experienced during an earthquake.

Earthquake ID" represents a specific earthquake event. In this dataset, multiple records can share the same Earthquake ID because the same event is captured by different stations. "Earthquake Key," on the other hand, represents the unique identifier for each individual record in the dataset.

3.4 Data preprocessing

To accurately predict S-wave PGV values from P-wave parameters like Pa, Pv, and Pd, it is crucial to consider the findings from seismic studies. Research has shown that the P-peak initial amplitude can reliably indicate the late maximum amplitude of seismic records within an epicentral distance (the horizontal distance between a specific observation point and the hypocentre of the earthquake, where the seismic waves originate.) range of less than 100 km(Zollo et al., 2023). Therefore, for precise predictions of S-wave intensity values based on P-wave parameters, data within this epicentral distance limit should be utilized. This filtering process led to the final selection of 9,206 earthquake events, producing 29,058 earthquake ground motion waveforms recorded by 293 stations, suitable for comprehensive analysis. Identified potential outliers are removed after careful consideration, as discussed in the preliminary analysis section.

Duplicate earthquake records were carefully assessed and handled to ensure data accuracy. Same earthquake records from different stations were not treated as duplicates because each station records unique ground motion characteristics based on its location, local soil conditions, and other geotechnical factors. These variations are crucial for understanding how the same earthquake affects different areas. Therefore, rather than being redundant, these records provide valuable information about spatial differences in seismic intensity, contributing to a more comprehensive and accurate model of S-wave intensity estimation across different stations.

When predicting the S-wave intensity parameter PGV using P-wave parameters (Pa, Pv, Pd), the raw parameter values are often transformed into logarithmic values. The rationale and further details regarding this transformation are also covered in the preliminary analysis section.

All categorical variables in the dataset were transformed into factors to ensure more appropriate handling and proper encoding for future data modelling. Year and month indexes were created from the P-wave pick time variable for future analysis purposes.

4. Methodology

4.1 Context and research question

In this research, the historical earthquake dataset used for this analysis of NZ. Classical regression models are unsuitable for this dataset due to data non-independence and the influence of station parameters like geographical location. Single-station data models are impractical due to insufficient data points for some stations and the complexity of building and fine-tuning models for over 300 stations. Instead, a Linear Mixed-Effects Modelling (LMM) is appropriate. LMM uses all available data, accounting for correlations between records from different stations, and incorporates variables such as station, earthquake event, and time of occurrence as random effects. LMM effectively handles unbalanced data by incorporating random effects, ensuring better model fit, accurate estimations, accurate standard errors, and reliable statistical inferences.

4.2 Limitations of traditional linear regression models

Traditional linear regression models, while useful, have significant limitations in this context:

Ignoring data non-independency: Standard linear regression assumes that all observations are independent. This assumption is violated in our dataset, where observations within the same group (such as earthquake records from the same stations or repeated seismic records of the same earthquake event) are correlated. Ignoring this correlation can result in incorrect standard errors and significance tests(Baltagi & Liu, 2020).

Ignoring Within-Group Correlation: Traditional models do not account for the correlation of observations within the same group or cluster. This can result in inflated Type I error rates and unreliable confidence intervals because the model treats correlated data as if they were independent (Huang et al., 2012).

Overlooking Hierarchical Variability: Different seismic stations may exhibit distinct baseline intensities and relationships between P-wave parameters and S-wave PGV. Ignoring this hierarchical structure can lead to biased estimates and incorrect inferences (Kubokawa, 2010).

Homogeneity Assumption: Standard regression models assume that the effect of the predictors is the same across all groups. This assumption is unrealistic in our context where the influence of P-wave parameters on S-wave PGV may differ between stations (Yuniarti et al., 2022).

Inflexibility with Unbalanced Data: Traditional linear regression models assume that each group in the data (e.g., seismic stations) contributes a similar number of observations. When the data is unbalanced: meaning some stations have significantly more seismic records than others, these models can struggle to provide reliable estimates. In real-world scenarios, like in this dataset, some stations may have recorded many earthquake events, while others have fewer records. Linear regression models do not inherently account for this variability in data distribution, potentially leading to biased estimates and reduced model performance when applied to unbalanced data (RUSU & ROMAN, 2018).

To address these limitations, an LMM is applied. An LMM extends the standard linear regression model by including both fixed and random effects, making it well-suited for handling repeated measures and unbalanced data structures.

4.3 Concept of linear mixed effect model

This approach incorporates both fixed effects, which describe the average relationship between the predictors and the response variable across all groups, and random effects, which account for variations within groups.

Fixed Effects: These coefficients describe the average relationship between the predictors and the response variable across all groups, similar to standard linear regression coefficients.

Random Effects: These account for the variability within groups (such as seismic stations) by introducing random deviations from the fixed effects, allowing intercepts and slopes to vary across groups.

4.4 General formulation

The linear mixed-effects model is an extension of classical linear regression models, designed to account for hierarchical or nested structure of the data. In this model, the interest lies in understanding the relationship between an independent variable x and a dependent variable y, where observations (x1, y1), . . . (xn, yn) are collected for each subject. A standard linear regression model can be represented as:

$$y_j = \alpha + \beta x_j + \varepsilon_j \qquad \text{---------------------} \ (1)$$

When such regression data are available for multiple subjects, a model that accommodates different regression lines for each subject is formulated as:

$$y_{ij} = \alpha_i + \beta_i x_{ij} + \varepsilon_{ij} \qquad \text{------------------} \ (2)$$

or, in more general notation:

$$y_i = \alpha(\text{subject}_i) + \beta(\text{subject}_i)x_i + \varepsilon_i \quad \text{------------------} \ (3)$$

In this context, the regression relationships vary across subjects. However, if the primary interest lies in estimating the average relationship across all subjects.

To model this scenario more appropriately, the subject-specific effects (both intercepts and slopes) are treated as random variables:

$$y_i = a(\text{subject}_i) + b(\text{subject}_i)x_i + \varepsilon_i \quad \text{---------} \ (4)$$

where:

$$a(k) \sim N(\alpha, \sigma^2),\ \ b(k) \sim N(\beta, \sigma^2), \varepsilon_i \sim N(0, \sigma^2)$$

Here, $k = 1, \ldots, K$ with K representing the number of subjects, and α and β are the unknown population intercept and slope, respectively. The expected value and variance of the response y_i can then be written as:

$$\mathrm{E}(y_i) = \alpha + \beta x_i \qquad \text{------------------------} \ (5)$$

$$\mathrm{Var} y_i = \sigma_a^2 + \sigma_b^2 x_i^2 + \sigma^2 \qquad \text{------------------------} \ (6)$$

Thus, the model can be expressed as follows, explicitly separating the fixed and random components:

$$y_i = \alpha + \beta x_i + a(\mathrm{subject}_i) + b(\mathrm{subject}_i)x_i + \varepsilon_I \quad \text{-------} \ (7)$$

where:

$$a(k) \sim N(0, \sigma^2),\ \ b(k) \sim N(0, \sigma^2), \varepsilon_i \sim N(0, \sigma^2)$$

This formulation highlights the structure of the linear mixed model. It is assumed that the random effects *a(k)*, $b(k)$ and ε_i are mutually independent. However, in some cases, the intercept and slope values might be correlated. To account for this, a bi-variate normal distribution can be assumed for the random effects:

$$(a(k), b(k)) \sim N(0, \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}),\ \epsilon_i \sim N(0, \sigma^2)$$

4.5 Metrics for variability in grouped data

4.5.1 Intra-class correlation (ICC)

The Intra-Class Correlation (ICC) is a measure used to describe how strongly units in the same group resemble each other in mixed effect models. It quantifies the proportion of total variance that is attributable to the grouping structure of the data.

ICC is calculated as follows:

$$\text{ICC} = \frac{\sigma_{group}^2}{\sigma_{group}^2 + \sigma_{residual}^2} \qquad \text{-----------} \quad (8)$$

Where σ_{group}^2 is the variance between groups and $\sigma_{residual}^2$ is the variance within groups.

4.5.2 Design effect value

The Design Effect quantifies the impact of violations of independence on standard error estimates. It provides an estimate of the multiplier required to adjust standard errors to account for the negative bias introduced by nested data structures. The Design Effect is computed as follows:

$$\text{Design Effect} = 1 + (n_c - 1)\text{ICC} \qquad \text{-------------} \quad (9)$$

where n_c is the average cluster size and ICC is the Intra-Class Correlation.

4.6 Measures of performance

Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Likelihood Ratio Tests are employed as evaluation criteria for assessing the performance of the techniques discussed in advanced analysis.

4.6.1 Root mean square error (RMSE)

Root Mean Square Error measures the average squared difference between the actual value and its estimated value at the test points. It provides information about the magnitude of the average error. RMSE is calculated as follows, where the estimated value is denoted by $\widehat{T}(x_i)$, the actual value is denoted by $T(x_i)$, and the number of test points is denoted by n:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (T(x_i) - \widehat{T}(x_i))^2} \qquad \text{---------} \quad (10)$$

4.6.2 Mean absolute percentage error (MAPE)

Mean Absolute Percentage Error expresses the Mean Absolute Error (MAE) as a percentage. This measure is undefined if the actual value $T(x_i)$ is zero. For estimates that are too low, MAPE cannot exceed 100%, but for estimates that are too high, MAPE can exceed 100% since there is no upper limit to the percentage error. Despite this, MAPE is useful for comparing two methods as it calculates the error relatively. MAPE is calculated as follows:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\widehat{T}(x_i) - T(x_i)}{T(x_i)}\right| \times 100\% \qquad \text{----------} \ \textit{(11)}$$

where $\widehat{T}(x_i)$ and $T(x_i)$ are the estimated and actual values, respectively.

4.6.3 Akaike information criterion (AIC)

Akaike Information Criterion is a measure used for model comparison. It quantifies the trade-off between the goodness of fit of the model and its complexity. A lower AIC value indicates a better model, as it suggests a good fit with fewer parameters. The formula for AIC is as follows:

$$\text{AIC} = 2k - 2\ln(L) \qquad \text{----------------} \ \textit{(12)}$$

where k is the number of parameters in the model and L is the likelihood of the model given the data.

4.6.4 Bayesian information criterion (BIC)

Bayesian Information Criterion is another criterion for model comparison, similar to AIC but with a stronger penalty for models with more parameters. BIC aims to balance model fit and complexity, with lower BIC values indicating better models. The formula for BIC is:

$$\text{BIC} = k\ln(n) - 2\ln(L) \qquad \text{-----------------} \ \textit{(13)}$$

where k is the number of parameters, $n$is the number of observations, and L is the likelihood of the model.

4.6.5 Likelihood ratio test

Likelihood Ratio Test (LRT) is a statistical test used to compare the goodness of fit between two nested models. It evaluates whether adding more parameters to a model significantly improves the fit. The test statistic is:

$$\mathrm{LR} = -2(\ln(L_0) - \ln(L_1)) \quad \text{----------------} \quad (14)$$

where L_0 is the likelihood of the simpler model and L_1 is the likelihood of the more complex model. This test statistic follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models.

5. Preliminary Analysis

This chapter comprises exploratory data analyses conducted prior to applying the Linear Mixed-Effects Model. The aim is to understand the structure and characteristics of the data, identify patterns or anomalies, and ensure the data meets the assumptions necessary for further statistical modelling. This process includes examining the distributions of key variables, assessing relationships between predictors and the response variable, and identifying potential outliers. Descriptive statistics and visualizations provide insights into the dataset, guiding subsequent modelling decisions and improving the reliability of the final analysis.

5.1 Association between PGV and Pa

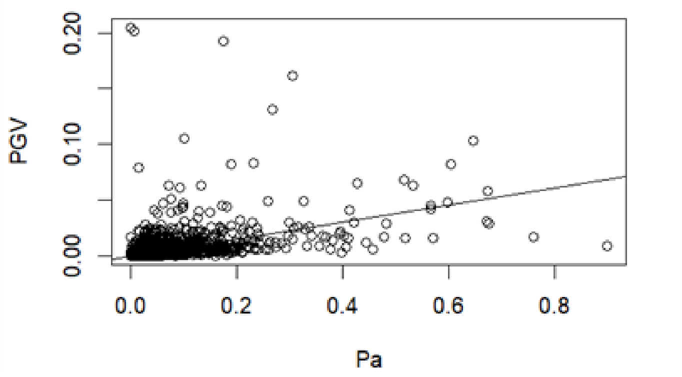


Figure 5.1: Scatter Plot: PGV vs Pa

Based on the scatter plot, there appears to be a positive linear relationship between PGV and Pa. The R-squared value of 0.2887 indicates that approximately 28.87% of the variation in PGV can be explained by Pa. However, the data points exhibit some scatter around the trend line, suggesting that other factors beyond Pa might influence PGV. Additionally, a few outliers can be observed, indicating potential unusual events or conditions that require further investigation.

5.2 Association between PGV and Pv

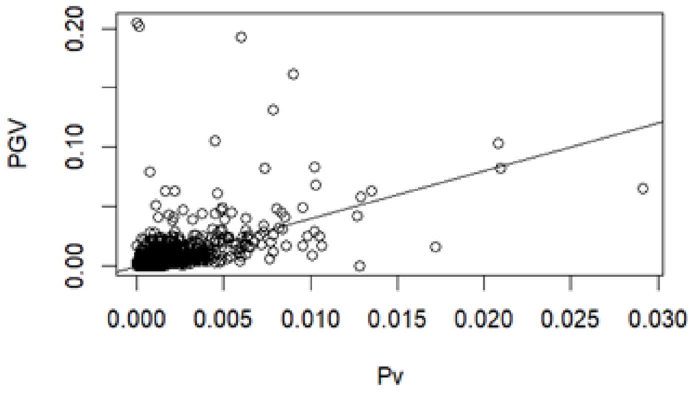


Figure 5.2: Scatter Plot: PGV vs Pv

The scatter plot illustrates a positive linear relationship between PGV and Pv. An R-squared value of 0.3879 indicates that approximately 38.79% of the variation in PGV can be explained by Pv. However, the data points exhibit some scatter around the trend line, suggesting that other factors beyond Pv might influence PGV. Additionally, a few outliers can be observed, indicating potential unusual events or conditions that require further investigation.

5.3 Association between PGV and Pd

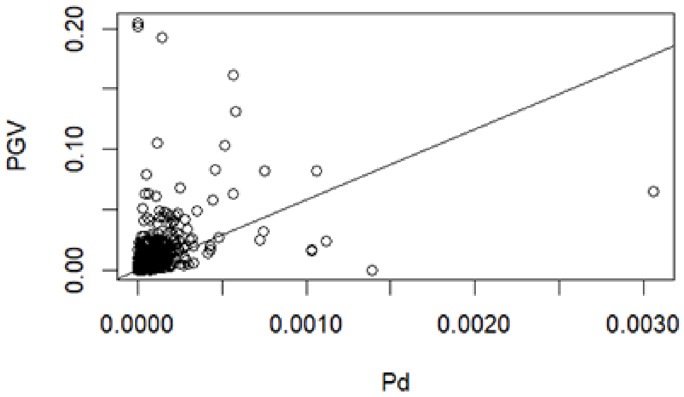


Figure 5.3: Scatter Plot: PGV vs Pd

The scatter plot illustrates a positive linear relationship between PGV and Pd. An R-squared value of 0.2448 indicates that approximately 24.48% of the variation in PGV can be explained by Pd. However, the data points exhibit some scatter around the trend line, suggesting that other factors beyond Pd might influence PGV. Additionally, a few outliers can be observed, indicating potential unusual events or conditions that require further investigation.

5.4 Investigating the Correlation Between Pa and Pv, Pa and Pd, and Pv and Pd

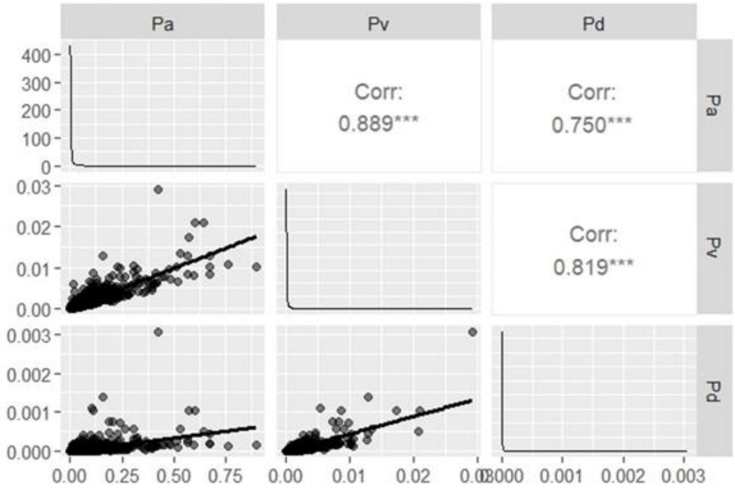


Figure 5.4: Correlation matix for Pa, Pv, Pd

The correlation matrix visually represents the pairwise relationships between the three variables Pa, Pv, and Pd. The diagonal elements display the correlation of each variable with itself (1.000), while the off-diagonal elements show the correlations between different variables. The numerical values in the upper right corner of each cell indicate the correlation coefficients, and the asterisks denote the statistical significance of these correlations. From the matrix, it can be observed that there are strong positive correlations between all three variables. The highest correlation is between Pa and Pv, with a coefficient of 0.889, indicating a strong positive relationship. The correlations between Pa and Pd, and between Pv and Pd, are also high, at 0.750 and 0.819 respectively. These strong correlations suggest that there are overlapping relationships between these variables, and it might be necessary to consider multicollinearity when analyzing their effects on other variables.

5.5 Identifying outliers

The boxplot presents the distribution of Peak Ground Velocity (Pv) values for P-waves in the dataset.

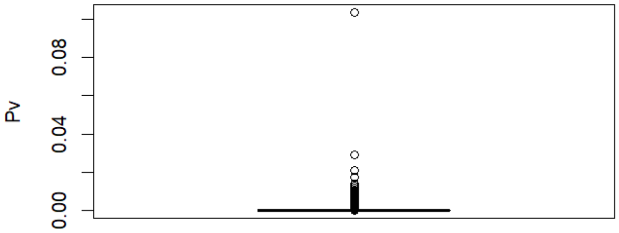


Figure 5.5: Box Plot: Distribution of Pv

The central thick black line within the box plot represents the median Pv value, which is 0.0000295. The distribution exhibits right skewness, suggesting that the majority of Pv values are relatively low, with a few exceptionally high values. Among the outliers, one notably highest Pv value was identified and removed after careful examination, as it was deemed unusual. The remaining outliers were verified as actual records and retained in the dataset.

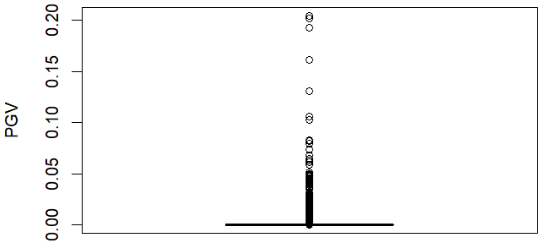


Figure 5.6: Box plot: Distribution of PGV

The boxplot illustrates the distribution of PGV values for S-waves in the earthquake dataset. The median PGV value, depicted by the central thick black line inside the box, represents the median value. PGV values span from 0 to just above 0.20. The distribution is skewed to the right, with most PGV values clustering towards the lower end of the range and a few notably higher values. Above the whiskers, several outliers are visible, indicating PGV values significantly higher than the majority of the data. These outliers were carefully verified as actual records and were consequently retained in the dataset rather than being removed.

5.6 Appling log transformation

In seismic analysis literature, logarithmic transformations are commonly applied to address such data imbalances. When predicting S-wave intensity parameter PGV using P-wave parameters (Pa, Pv, Pd), raw parameter values are often transformed into logarithmic values(Chandrakumar et al., 2024). This transformation is employed for several reasons based on existing research findings. Logarithmic transformations are utilized to standardize highly varying values in raw data, ensuring proper training of predictive models(Franco et al., 2022). In the context of seismic analysis, generalized logarithm transformations have been shown to produce values with approximately constant variance when the raw data variance has both additive and multiplicative components (Carle et al., 2013). Logarithmic transformations effectively compress data with a wide range of values, reducing the impact of extreme outliers and enhancing the robustness of the relationships(Liang et al., 2015). This transformation is crucial for improving the reliability and interpretability of seismic relationships. By utilizing logarithmic scales, data representation becomes more linear, simplifying the analysis process and aiding in establishing meaningful correlations. The figure 5.7 clearly illustrates the relationship between PGV and Pv following logarithmic transformation.

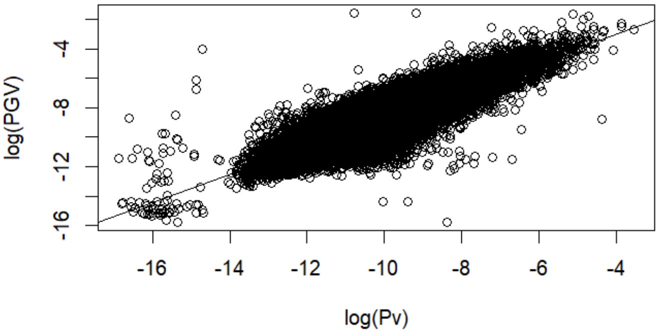


Figure 5.7: Scatter Plot: log(Pv) vs log(PGV)

After applying a logarithmic transformation, a discernible positive correlation between log(PGV) and log(Pv) is evident. This relationship suggests that as log(Pv) increases, log(PGV) also tends to increase. The transformed data points form a dense, stretched cluster with a general upward trend, indicating a strong linear relationship between these variables. The correlation coefficient of 0.87 further supports the strong positive relationship between them. Notably, some outliers are present, especially at the lower ranges of log(Pv) and log(PGV), deviating significantly from the main cluster. This scatter plot enhances the understanding of the relationship between the two variables compared to the previous non-transformed version.

Similarly, after this logarithmic transformation was applied to the relationships PGV~Pa and PGV~Pd, an improved understanding of the relationships was evident compared to the non-transformed versions. These relationships are shown in Figure 5.8, both of which display strong positive correlations. Some outliers were also present in these figures.

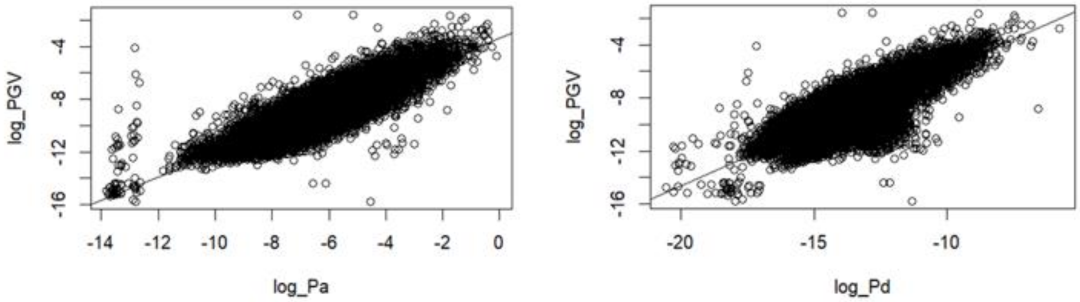


Figure 5.8: Scatter Plots - log(PGV) Vs log(Pa) and log(PGV) Vs log(Pd)

5.7 Checking linearity of data

This plot shows the residuals against the fitted values from a linear model where log(PGV) is regressed on log(Pv).

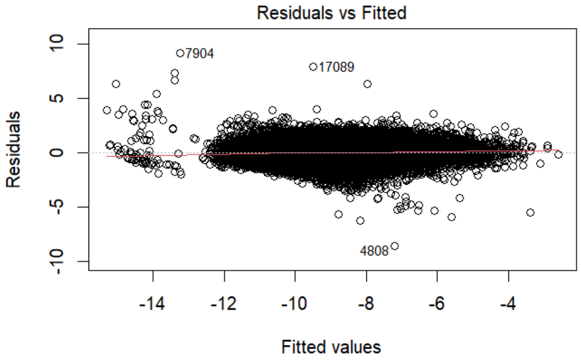


Figure 5.9: Residual Plot: lm(log(PGV) ~ log(Pv))

The x-axis represents the fitted values, and the y-axis represents the residuals. The residuals exhibit a random scatter around the horizontal line at zero, indicating that the linear model captures the general trend between log(PGV) and log(Pv). However, the spread of residuals is not uniform across the range of fitted values. The plot reveals some error variance heterogeneity, where the variance of residuals changes with fitted values. Specifically, there appears to be a funnel shape, with greater variance in residuals at the lower end of the fitted values. The presence of heteroscedasticity suggests that the linear model may not fully capture the relationship between log(PGV) and log(Pv) and that there may be other underlying factors or nonlinear relationships that are not accounted for. The outliers indicate that some observations do not fit the general pattern and may require further investigation.

5.8 Association between log(PGV) and Station

The below plot shows boxplot of ‘log(PGV)’ for 50 stations (randomly selected).

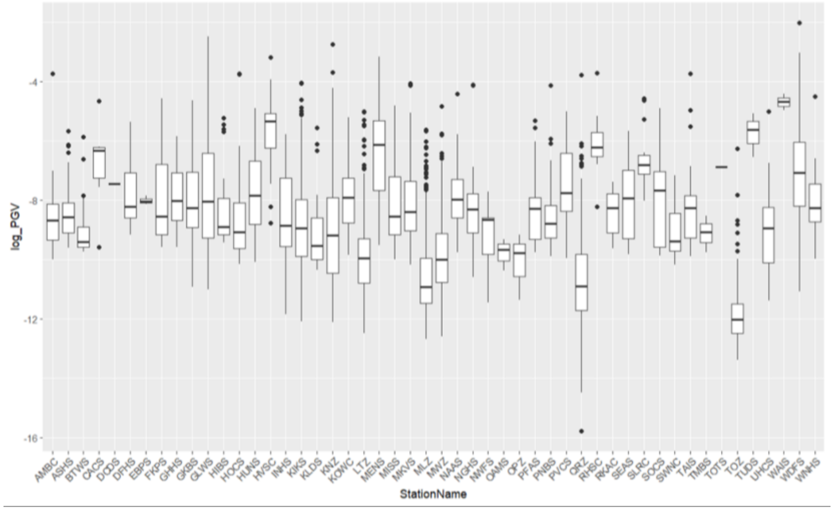


Figure 5.10: Boxplot of log_PGV by stationName

This plot displays the distribution of log-transformed PGV values across selected stations (not all 300+ stations in the dataset are shown, as including all would make the graph unclear). There is significant variability in the median ‘log(PGV)’ values across different stations. Some stations shows wider Interquartile range (IQR) and some have narrow IQR. The variability in median values of 'log(PGV)' across different stations can be attributed to several factors. Firstly, stations have differing geotechnical properties, which influence the recording of seismic waves. secondly, stations are situated in diverse locations; some frequently experience high earthquake activity, while others do not. This visualization suggest that station wise classification is an important factor influencing the log(PGV) and predictor variable relationship in earthquake records.

5.9 Association between log(PGV) and Year Index

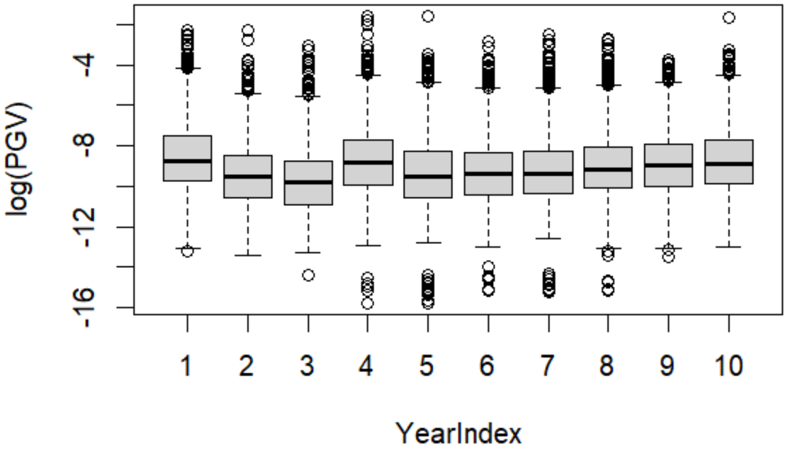


Figure 5.11: Boxplot of log_PGV by Year Index

The boxplot illustrates the distribution of log(PGV) across different Year Index groups. The x-axis represents the Year Index categories, while the y-axis shows the log(PGV) values. This boxplot does not suggest a consistent distribution of log(PGV) across the Year Index groups; there is some variation in the median and spread of values within each group, especially in the first 4 years. This graph suggests that year-wise classification can influence the log(PGV) and predictor variable relationship in earthquake records.

5.10 Distribution of the response variable – log (PGV)

The histogram and box plot provided offer a comprehensive view of the distribution of the logarithm of PGV (log (PGV)) in the earthquake records dataset.

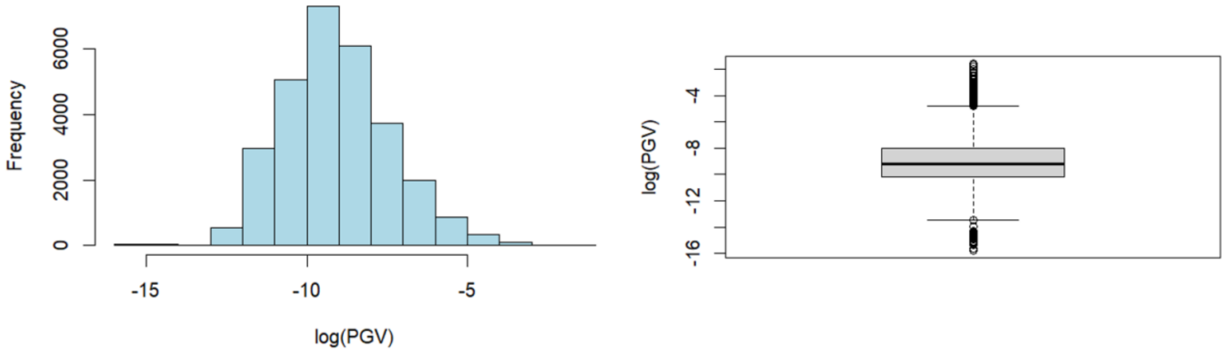


Figure 5.12: Distribution of log(PGV): Histogram and Box plot

The histogram of log (PGV) shows a roughly symmetric distribution centered around -10, with the majority of values falling between -12 and -8. This indicates that the log transformed PGV values are predominantly clustered in this range. The histogram suggests a normal-like distribution with some degree of skewness towards lower values, as indicated by the smaller but noticeable tail extending towards -15. The box plot whiskers extend to about -13 and -7, capturing the bulk of the data while highlighting several outliers beyond these bounds, both lower and higher. These outliers suggest that there are a few records with significantly different PGV values compared to the majority of the dataset. Together, these plots illustrate that while the log(PGV) values are generally centered around -10 with a fairly tight distribution, there is some variability, particularly with a few extreme values that extend the range of the dataset.

6. Advanced Analysis

This chapter explores the application of advanced statistical methods, specifically the Linear Mixed-Effects Model, to predict PGV values of S-waves. Insights from preliminary analysis are utilized to refine the modelling approach. Initially, the best predictor for estimating S-wave intensity is selected, followed by the construction of a simple linear regression model. After discussing the limitations of classical linear regression, the analysis advances to Linear Mixed-Effects Modelling. This model accounts for both fixed effects, representing overall population parameters, and random effects, capturing the variability within and between groups. The chapter details the model selection process, evaluation of model fit, and interpretation of results. By leveraging the nested structure of the data, the aim is to achieve more accurate and robust predictions, enhancing the reliability and applicability of the findings.

6.1 Selection of predictor variable for estimating PGV

This analysis evaluates the selection of a suitable predictor variable from Peak Ground Acceleration (Pa), Peak Ground Velocity (Pv), and Peak Ground Displacement (Pd) for estimating the response variable PGV. Due to the high correlation among Pa, Pv, and Pd, stemming from their integrative relationships (with Pd being derived from integrating Pv, and Pv being derived from integrating Pa), only one of these parameters should be used as a predictor in the model to avoid multicollinearity issues. Both visual and numerical evidence of multicollinearity among the predictor variables is presented in section 5.4.

To determine the most appropriate predictor variable, simple linear regression models were constructed with log-transformed PGV (log(PGV)) as the response variable and log-transformed Pa, Pv, and Pd as individual predictor variables. The performance of these models was assessed using R-squared, AIC, BIC on the training set, and RMSE and MAPE on the test set. The results are summarized in the table 6.1.

Table 6.1: Summary of model performances for different predictors

Model	R-squared	AIC	BIC	RMSE	MAPE
log(PGV)~ log(Pa)	0.7453	57960.13	57984.30	0.8321	7.700 %
log(PGV)~ log(Pv)	0.7538	57167.69	57191.85	0.8305	7.518%
log(PGV)~ log(Pd)	0.6385	66100.02	66124.18	1.0131	9.198%

The model log(PGV) ~ log(Pv) demonstrates superior performance with the highest R-squared value (0.7538), the lowest AIC (57167.69) and BIC (57191.85) values, as well as the lowest RMSE (0.8305) and MAPE (7.518%) values on the test set. These metrics provide robust evidence that log(Pv) is the most effective predictor among the three.

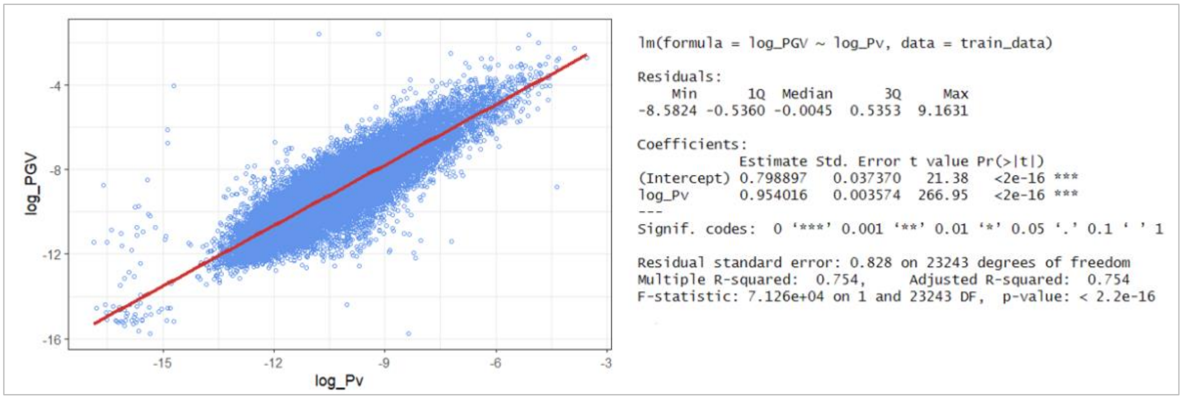
Furthermore, literature supports the selection of Pv as a predictor variable for estimating S-wave intensity, with numerous studies validating its efficacy. In conclusion, based on statistical evidence and existing research, log(Pv)is selected as the predictor variable for estimating log(PGV) over log(Pa) and log(Pd).

6.2 Simple linear regression

First, a simple linear regression model was constructed to regress the response variable, log(PGV), on the predictor variable.

```
model_ols <- lm(log_PGV ~ log_Pv,  data = train_data)
```

Below is the scatter plot of the data along with the output of the regression model.



```
lm(formula = log_PGV ~ log_Pv, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5824 -0.5360 -0.0045  0.5353  9.1631

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.798897   0.037370   21.38   <2e-16 ***
log_Pv      0.954016   0.003574  266.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.828 on 23243 degrees of freedom
Multiple R-squared:  0.754,     Adjusted R-squared:  0.754
F-statistic: 7.126e+04 on 1 and 23243 DF,  p-value: < 2.2e-16
```

Figure 6.1:Scatter plot of log_PGV Vs log_Pv and R output

The resulting model suggests that for every one unit increase in log(Pv), there is, on average, a 0.954 increase in log(PGV). The coefficient for the log(Pv) variable has a standard error of 0.0036, and the model's R-squared value indicates that log(Pv) explains approximately 75.4% of the variance in log(PGV).

At first glance, this appears to be a robust model. However, it violates the assumption of data independence inherent to simple linear regression, as the observations are not independent. Our earthquake dataset includes multiple records for the same earthquake event from different stations, resulting in correlated data. This correlation can lead to biased standard errors and unreliable inferences. Addressing this issue is essential. To illustrate this, a visual representation for a few stations is provided below.

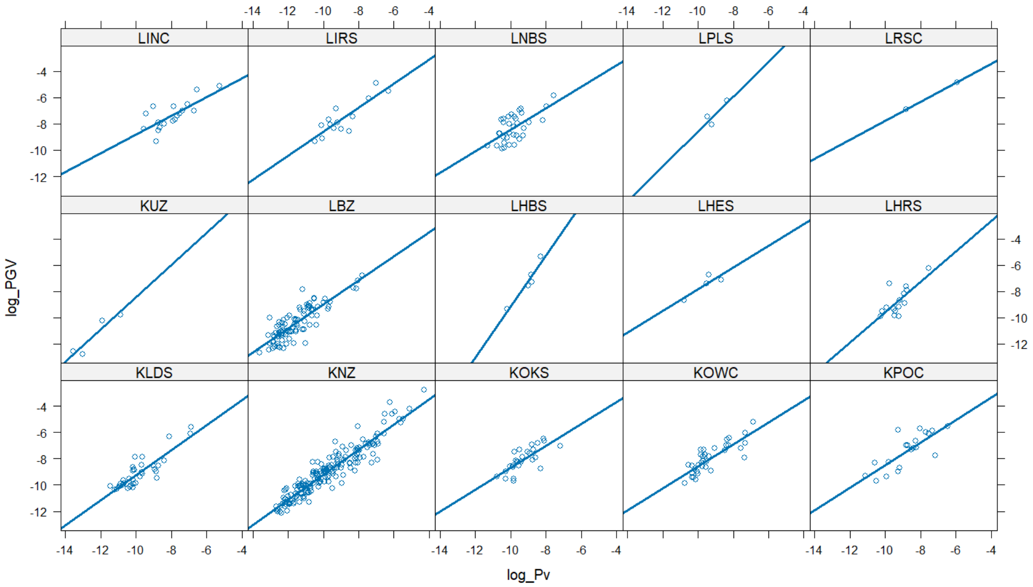


Figure 6.2: Relationship between log(Pv) and log(PGV) for different Stations

The panel of plots indicates that individual earthquake records positively affect higher values of log(PGV) where their log(Pv) is higher, but to varying extents for each station. Among the 15 stations, the intercepts and slopes of the linear relationships differ. Additionally, the amount of data varies significantly between stations, with some having extensive data and others having only a few records.

To further clarify, Figure 6.3 combines the data from all stations into a single plot. The colors represent the station-wise relationship between log(Pv) and log(PGV). The dataset includes 293 stations, and the chart displays 293 regression lines, each corresponding to a specific station. From this figure, the variation in intercepts and slopes among the stations is evident.

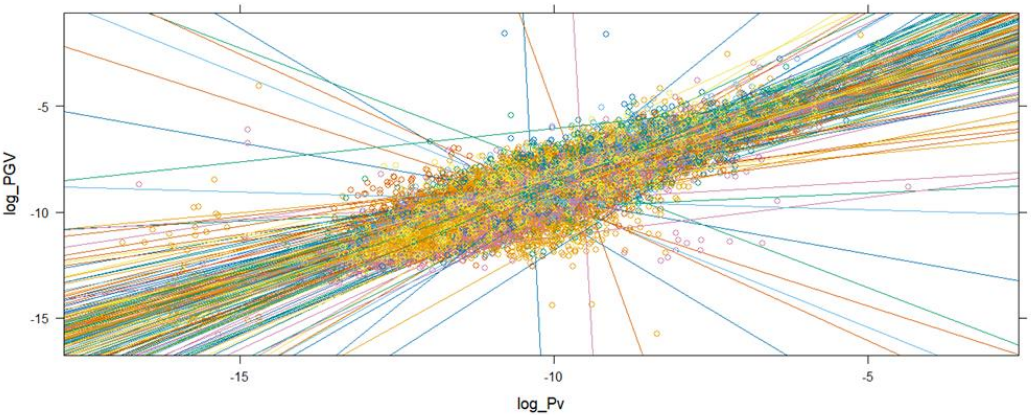


Figure 6.3: Relationship between log(Pv) and log(PGV) for all stations

If a typical regression were used to analyze this data, the different intercepts and slopes would be ignored. Such a model would poorly estimate relationships among these variables. If data are clustered (i.e., in multi-level data), the independent assumption in classical linear regression is violated. Single-station data models are impractical due to insufficient data points for some stations and the complexity of building and fine-tuning models for over 300 stations. In this case, new regression modelling techniques are required.

A linear mixed effect modelling (LMM) is an ideal choice here as it allows the use of all available data, whether the sample size is large or small, and accounts for correlations between data from different stations, site classes, earthquakes, and time periods. LMM incorporate random effects, providing more accurate standard errors and reliable statistical inferences. They effectively handle unbalanced data, accommodating all available data without discarding incomplete cases, ensuring a better model fit and more accurate predictions.

For instance, some stations have many records while others have few, leading to unbalanced data. By using linear mixed effect modelling, repeated measures on the same individuals are accounted for, making the analyses robust and the results more generalizable.

Stations may have inherent characteristics (such as geology or proximity to fault lines) influencing the relationship between log(PGV) and log(Pv) beyond the general effect captured by the fixed effect (log(Pv)). Similarly, earthquakes might display unique characteristics (such as focal depth, magnitude, or slip mechanism), and earthquakes occurring in different years might exhibit unique properties affecting this relationship. By incorporating random effects for stations, earthquake IDs, and years, a LMM accounts for these potential sources of variation, resulting in more accurate estimates and reliable conclusions. This approach captures the variability within and between these groups effectively.

6.3 Assessing the necessity of linear mixed effect modelling

Now that the limitations of the simple linear regression model and the advantages of applying linear mixed-effect modelling to this estimation are understood, it is necessary to check the Intraclass Correlation Coefficient (ICC) and design effect values to ensure that LMM is appropriate for our analysis. The ICC indicates how much of the total variance in the response variable, PGV, can be attributed to differences between groups (stations, earthquake events, and Year Index), with a high ICC justifying the use of mixed effect modelling. The design effect measures how clustering affects the precision of the estimates, with a higher value suggesting a reduced effective sample size due to clustering. Generally, an ICC greater than 0.05 and a design effect exceeding 2 indicate that LMM is suitable for handling correlated unbalanced data.

To determine the necessity of LMM, The ICC value and design effect value, which should be assessed in the null model(unconditional mean model), which serves as our baseline model in LMM with no predictor variables. This model estimates the overall mean of the dependent variable PGV and allows for random intercepts at each grouping level.

Fitting the unconditional means model(null model):

```
model_null<- lmer(`log(PGV)` ~ 1 + (1|StationName) + (1| EarthquakeID) + (1| Year_Index)
```

Below mentioned the resulted ICC value and design effect values assessed on unconditional mean model.

Table 6.2: ICC value and design effect against each group

	ICC Value	Design Effect
Station Name	0.4586	46.02
Earthquake ID	0.6102	2.32
Year Index	0.1621	36.29

The ICC values underscore significant variability in PGV attributed to differences between stations and earthquake events. A low ICC value for Year Index suggests it does not explain a significant amount of the variance in PGV compared to the variation within years. However, Year Index is considered in the LMM because it slightly surpasses the required ICC value margin of 0.05.

The markedly high design effects, especially for stations and year indices, indicate pronounced clustering effects that substantially inflate standard errors. Such inflation can potentially lead to erroneous conclusions if not appropriately addressed in statistical modelling. The design effect value for Earthquake ID also slightly surpasses the required margin.

Incorporating random effects into the LMM is essential for accounting for the nested structure of the data. Including Station Name, Earthquake ID, and Year Index as random effects allows the model to capture unobserved variability within these groups. The high ICC values for Station Name (0.4586) and Earthquake ID (0.6102) indicate significant variability between these groups, suggesting that their inclusion can greatly improve the model's ability to explain the variation in S-wave intensity (log(PGV)). This approach enables the model to capture the unique characteristics of each station and earthquake event, leading to more accurate estimates and reducing bias. Although the ICC for Year Index is lower (0.1621), including it helps address temporal clustering in the data. Therefore, it can be concluded that all three random effects are significant.

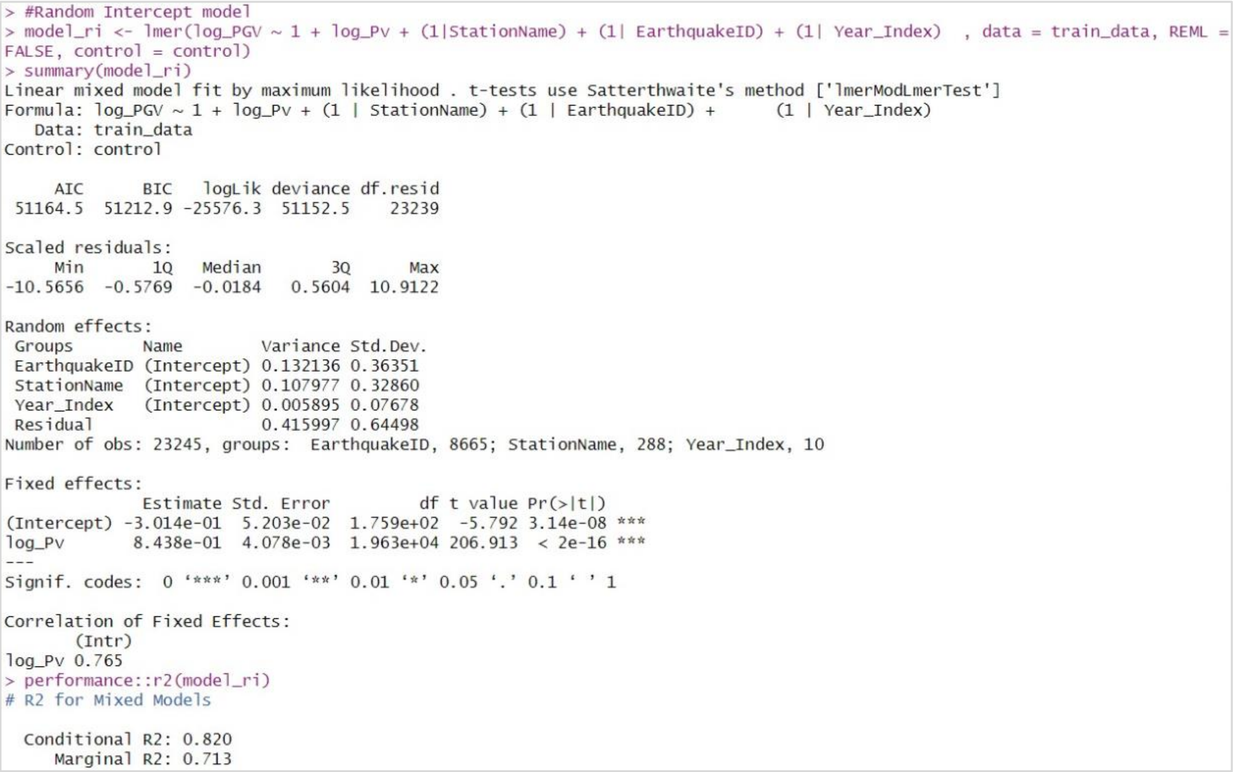
There is now enough evidence to use LMM to estimate log(PGV). The above-mentioned null model will be extended to reflect a simple linear regression model, but with random effects specified for each subject (Station Name, Earthquake ID, Year Index). In this extension, the only variation allowed is the individual subject's intercept for S-wave intensity (log(PGV)).

Extending the null model, an independent variable (predictor variable) will be added to investigate the predictability of the independent variables on the dependent variable. In this model, a fixed-effects term log(Pv) is added to estimate the relationship between the predictor variable x (log(Pv)) and the response variable y (log(PGV)).

6.4 Random intercept model

This section discusses the random intercept model and its results. The model is presented below.

model_ri<- lmer(log_PGV ~ log_Pv + (1|StationName) + (1| EarthquakeID) + (1| Year_Index)



```
> #Random Intercept model
> model_ri <- lmer(log_PGV ~ 1 + log_Pv + (1|StationName) + (1| EarthquakeID) + (1| Year_Index)  , data = train_data, REML =
FALSE, control = control)
> summary(model_ri)
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: log_PGV ~ 1 + log_Pv + (1 | StationName) + (1 | EarthquakeID) +      (1 | Year_Index)
   Data: train_data
Control: control

     AIC      BIC   logLik deviance df.resid
 51164.5  51212.9 -25576.3  51152.5    23239

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.5656  -0.5769  -0.0184   0.5604  10.9122

Random effects:
 Groups       Name        Variance Std.Dev.
 EarthquakeID (Intercept) 0.132136 0.36351
 StationName  (Intercept) 0.107977 0.32860
 Year_Index   (Intercept) 0.005895 0.07678
 Residual                 0.415997 0.64498
Number of obs: 23245, groups:  EarthquakeID, 8665; StationName, 288; Year_Index, 10

Fixed effects:
              Estimate Std. Error        df t value Pr(>|t|)
(Intercept) -3.014e-01  5.203e-02  1.759e+02  -5.792 3.14e-08 ***
log_Pv       8.438e-01  4.078e-03  1.963e+04 206.913  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
       (Intr)
log_Pv 0.765
> performance::r2(model_ri)
# R2 for Mixed Models

  Conditional R2: 0.820
     Marginal R2: 0.713
```

Figure 6.4: Summary and performance of model_ri

Based on the R outputs, the Random Intercept model (model_ri) offers a significantly better fit for explaining log(PGV) compared to the simple linear regression model (model_ols). Model_ri shows a substantial improvement in fit statistics, with an AIC of 51164.5 (57167.69 lower) and a

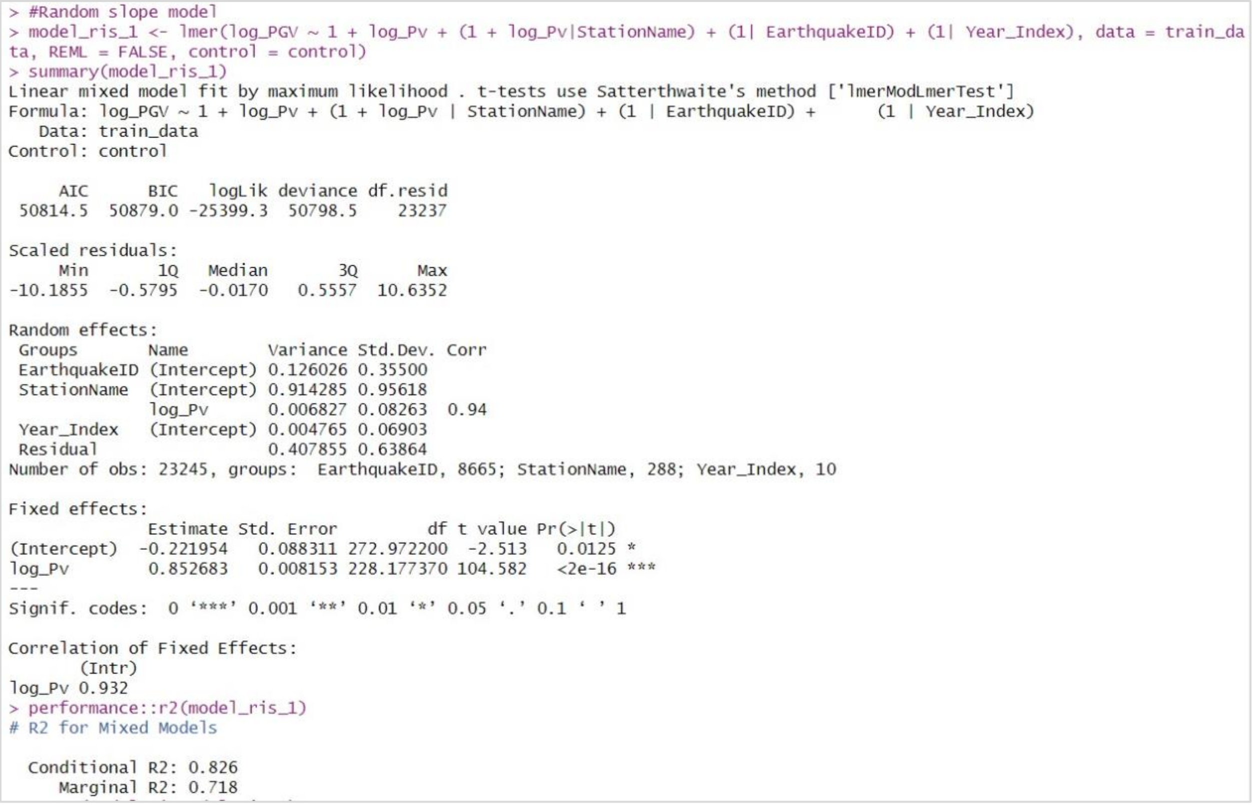
BIC of 51212.9 (57191.85 lower) than the simple linear regression model (model_ols). This suggests a better fit for the data while handling model complexity. By addressing data dependency issues, model_ri is likely to provide more accurate standard errors for the fixed effect coefficient (log_Pv), resulting in a more reliable estimate of the significance of the relationship between log_Pv and log_PGV. Here, the fixed-effects component shows that log_Pv is a significant predictor (t = 206.913, p-value < 2e-16). As log_Pv value increases by 1 point, the log_PGV increases by 0.844 points. Additionally, the Conditional R-squared for model_ri (0.820), which represents variance explained by both fixed and random effects, is higher than the R-squared for cr_model1 (0.754), indicating that model_ri explains a larger proportion of the total variance in log_PGV. Therefore, model_ri offers a statistically superior fit compared to the simple linear regression model, accounting for random variation within EarthquakeID, StationName, and YearIndex, and providing a more accurate representation of the data.

6.5 Check for random slope models

Previously, the coefficient for log_Pv was assumed to be the same across all subjects(Stations, Earthquake ID, Year Index). This assumption was based on the model allowing different intercepts for each subject while maintaining a constant slope. However, analysis of individual station plots indicates that the slopes may also vary between stations. This refined model will better capture the unique characteristics of each station, enhancing the accuracy of our S-wave intensity estimations.

Add random slope for station:

```
model_ris_1<- lmer( log_PGV ~  log_Pv + (1+ log_Pv |StationName) + (1| EarthquakeID) +
(1|  Year_Index)
```

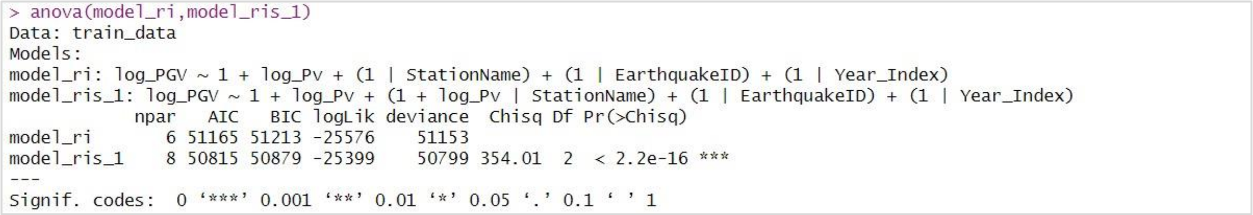


```
> #Random slope model
> model_ris_1 <- lmer(log_PGV ~ 1 + log_Pv + (1 + log_Pv|StationName) + (1| EarthquakeID) + (1| Year_Index), data = train_da
ta, REML = FALSE, control = control)
> summary(model_ris_1)
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: log_PGV ~ 1 + log_Pv + (1 + log_Pv | StationName) + (1 | EarthquakeID) +      (1 | Year_Index)
   Data: train_data
Control: control

     AIC      BIC   logLik deviance df.resid
 50814.5  50879.0 -25399.3  50798.5    23237

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.1855  -0.5795  -0.0170   0.5557  10.6352

Random effects:
 Groups       Name        Variance Std.Dev. Corr
 EarthquakeID (Intercept) 0.126026 0.35500
 StationName  (Intercept) 0.914285 0.95618
              log_Pv      0.006827 0.08263  0.94
 Year_Index   (Intercept) 0.004765 0.06903
 Residual                 0.407855 0.63864
Number of obs: 23245, groups:  EarthquakeID, 8665; StationName, 288; Year_Index, 10

Fixed effects:
             Estimate Std. Error         df t value Pr(>|t|)
(Intercept) -0.221954   0.088311 272.972200  -2.513   0.0125 *
log_Pv       0.852683   0.008153 228.177370 104.582   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
       (Intr)
log_Pv 0.932
> performance::r2(model_ris_1)
# R2 for Mixed Models

  Conditional R2: 0.826
     Marginal R2: 0.718
```

Figure 6.5: Summary and performance of model_ris_1

Based on the resulting R outputs, the AIC and BIC metrics favor model_ris_1, with significantly lower values (model_ri: AIC = 51164.5, BIC = 51212.9; model_ris_1: AIC = 50814.5, BIC = 50879.0). Lower values indicate a better fit while handling model complexity. model_ris_1 shows a slightly higher Conditional R-squared (0.826) compared to model_ri (0.820), suggesting it explains a marginally larger portion of the total variance in log_PGV. Additionally, the logLik statistic increased from -25576.3 in model_ri to -25399.3 in model_ris_1. From the fixed-effects component, it is evident that log_Pv remains a significant predictor of Course (t = 104.582, p-value < 2e-16). As the Written score increases by 1 point, the Course score increases by 0.853 points.

model_ri and model_ris_1 can also be compared using analysis of variance and analysis of deviance tables for the two model objects.



```
> anova(model_ri,model_ris_1)
Data: train_data
Models:
model_ri: log_PGV ~ 1 + log_Pv + (1 | StationName) + (1 | EarthquakeID) + (1 | Year_Index)
model_ris_1: log_PGV ~ 1 + log_Pv + (1 + log_Pv | StationName) + (1 | EarthquakeID) + (1 | Year_Index)
            npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
model_ri       6 51165 51213 -25576    51153
model_ris_1    8 50815 50879 -25399    50799 354.01  2  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6.6: Model comparison between model_ri and model_ris_1

This likelihood ratio χ2-test gives a significant model comparison (p-value < 2.2e-16), which indicates that the random-slope model model_ris_1 is more preferred to fit this data. This finding is consistent with the graphical presentation in figure 6.2, where the slopes are quite different from each station.

By adding a random slope effect for the Station factor, random slope effects for EarthquakeID and YearIndex factors can also be included. Their model summaries and performance can then be examined. Below are those models:

model_ris_2<- lmer(log_PGV ~ log_Pv + (1+log_Pv|StationName) +(1+log_Pv|
Year_Index)+ (1| EarthquakeID)

model_ris_3<- lmer(log_PGV ~ log_Pv + (1+log_Pv|StationName) + (1+log_Pv|
EarthquakeID) + (1| Year_Index)

While these models slightly reduce AIC and BIC values and slightly improve the conditional R² value, they fail to converge due to their complexity. Therefore, a simpler model that provides a better fit should be chosen. A more detailed model comparison will be discussed next.

6.6 Process for selecting the best model

Five models were compared to predict log_PGV, evaluating their fit and complexity . The null model (model_null) served as a baseline with the highest AIC (72542) and BIC (72590) values, indicating a poor fit compared to models with random effects.

Table 6.3: Summary of model selection

Model	AIC	BIC	Loglik	Chi-Sq	Pr(>Chisq)	R^2 Value	Convergence

model_null	72542	72590	-36265		-	marginal: 0.000 conditional: 0.752	converged
model_ri	51165	51213	-25576	21377.1	-	marginal: 0.713 conditional: 0.819	converged
model_ris_1	50815	50879	-25399	354.0	< 2.2e-16 ***	marginal: 0.718 conditional: 0.826	converged
model_ris_2	50796	50877	-25388	22.1	1.582e-05 ***	marginal: 0.718 conditional: 0.826	not converged
model_ris_3	50211	50292	-25096	585.067	-	marginal: 0.698 conditional: 0.832	not converged

Significant improvement was observed with model_ri (random intercept), achieving a substantial decrease in AIC (51165) and BIC (51213). This suggests that accounting for random variation within groups (stations, years or earthquake events) improves the model's ability to explain the data.

Further refinement came with model_ris_1 (random slope and intercept). It displayed a lower AIC (50815) and BIC (50879) compared to model_ri. Additionally, a highly significant Chi-squared test (p-value < 2.2e-16) confirmed that model_ris_1 offers a statistically better fit. This suggests that allowing the effect of the predictor (log_Pv) to vary across stations(random slope) captures additional variation in log_PGV compared to a fixed effect.

While model_ris_2 and model_ris_3 are achieved slightly lower AIC, BIC values respectively (50796 & 50877) and (50211, 50292), but those crucially did not converge during the fitting process. This raises concerns about the model's reliability and prevents its further consideration.

Therefore, based on both information criteria and convergence, model_ris_1 appears to be the most suitable choice. It demonstrates a statistically significant improvement over the random intercept model, captures additional variance by allowing random slopes, and successfully converged during the fitting process.

However, further investigation is needed to check model performances on test data set using RMSE(Root Mean Squared Error), MSE(Mean Squared Error) and MAPE(Mean Absolute Percentage Error) metrics.

Table 6.4: Model comparison on MSE, RMSE, MAPE

Model	**MSE**	**RMSE**	**MAPE**
model_ols	0.6854910	0.8279439	7.424481
model_null	1.0874969	1.0428312	9.454425
model_ri	0.4811100	0.6936209	6.188053
model_ris_1	0.4666977	0.6831528	6.137407
model_ris_2	0.4654574	0.6822444	6.128359
model_ris_3	0.4654574	0.6777532	6.087332

Error measures were used to assess the effectiveness of six models for predicting log_PGV on a test set. The error measures included RMSE, MSE, and MAPE. Lower values across these indicators suggest a model's greater generalizability to new data.

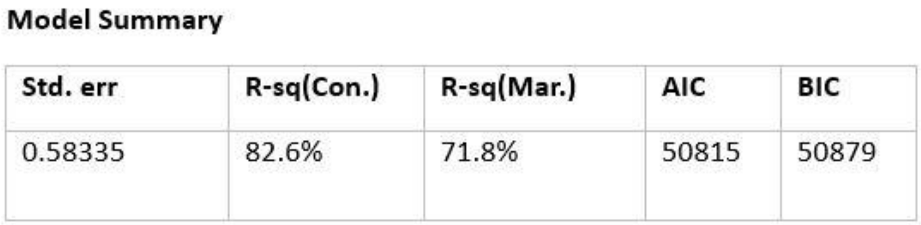
Based on the overall trend of reducing errors throughout the models, model_ris_1 or model_ris_2/3 may be recognized the best models. But model_ris_2/3 not successfully converged throughout the fitting phase (information given in the table 6.3), so they may not be feasible models because they are complex models and take higher computing time.

Finally, there is enough evidence to say that model_ris_1 is the best and feasible model for estimate log_PGV.

6.7 Summary of the selected model

The model equation and summary of the selected model, identified as the best model, are displayed below.

model_ris_1<- lmer(log_PGV ~ log_Pv + (1+ log_Pv |StationName) + (1| EarthquakeID) + (1| Year_Index)

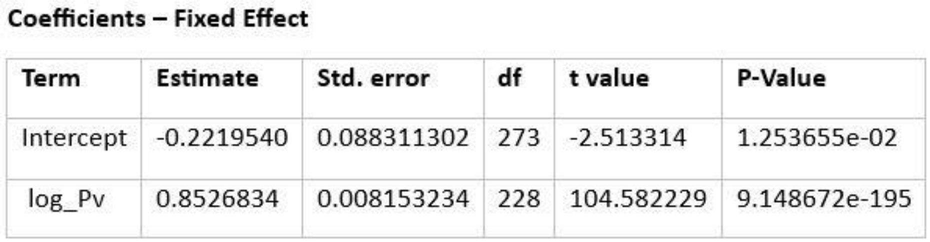


Model Summary

Std. err	R-sq(Con.)	R-sq(Mar.)	AIC	BIC
0.58335	82.6%	71.8%	50815	50879

Figure 6.7: Model summary of model_ris_1

The selected linear mixed-effects model demonstrates a standard error of 0.58335, indicating the average deviation of the observed log_PGV values from the predicted values, compared to the classical regression model's (model_ols) higher residual standard error of 0.828. The conditional R-squared (R-sq (Con.)) for the mixed model is 82.6%, and the marginal R-squared (R-sq (Mar.)) is 71.8%, suggesting the model explains a significant portion of the variance in the data. In contrast, the classical regression model(model_ols) has an R-squared value of 75.4%, explaining a slightly lower proportion of the variability in log_PGV. While the classical model shows a strong relationship between log_Pv and log_PGV, the higher residual standard error indicates substantial unexplained variability. The mixed-effects model, by accounting for non-independence in the data due to repeated measurements, provides a more accurate and reliable representation of the factors influencing PGV.

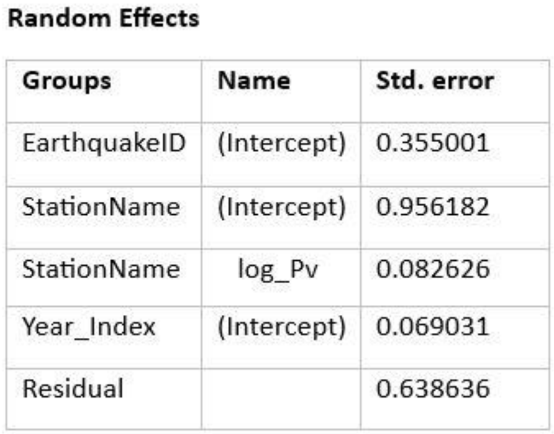


Coefficients – Fixed Effect

Term	Estimate	Std. error	df	t value	P-Value
Intercept	-0.2219540	0.088311302	273	-2.513314	1.253655e-02
log_Pv	0.8526834	0.008153234	228	104.582229	9.148672e-195

Figure 6.8: Coefficient table of model_ris_1

Similar to the classical linear regression model, for every 1 unit increase in log_Pv, there is an average increase of 0.8527 units in log_PGV (S-wave intensity). The intercept is -0.221954 with a p-value of 0.0125, indicating it is significantly different from zero. The coefficient for log_Pv is 0.852683, with a very small standard error of 0.008153234 and a very high t-value of 104.582229, suggesting a strong and highly significant relationship (p-value = 9.148672e-195) between log_Pv and log_PGV.



Random Effects

Groups	Name	Std. error
EarthquakeID	(Intercept)	0.355001
StationName	(Intercept)	0.956182
StationName	log_Pv	0.082626
Year_Index	(Intercept)	0.069031
Residual		0.638636

Figure 6.9: Random effect variation

The random effects table for the chosen linear mixed-effects model indicates the variability attributable to different grouping factors. The standard error for EarthquakeID (intercept) is 0.355001, for StationName (intercept) is 0.956182, and for StationName (log_Pv) is 0.082626, reflecting the variability across different earthquakes and stations. The Year_Index (intercept) has a standard error of 0.069031, indicating variability across different years. The reduced residual standard error is 0.638636, showing the variability not explained by the fixed or random effects. These random effects capture additional variability, enhancing the model's accuracy and generalizability.

The behavior of each individual within the model can be explicitly examined by analyzing the random effect coefficients (extracted using ranef()), which indicate the differences between the individual’s model parameters (intercept and slope) and the overall model parameters (intercept and slope). Alternatively, the linear equation for each participant can be derived using coef().

6.8 Prediction using the selected model

Making predictions with a mixed model differs from classical regression. In LMM models, there is not only the linear regression component (often referred to as fixed effects) but also a component that accounts for specific individuals. Additionally, extracting confidence or prediction intervals is more complex compared to classical regression.

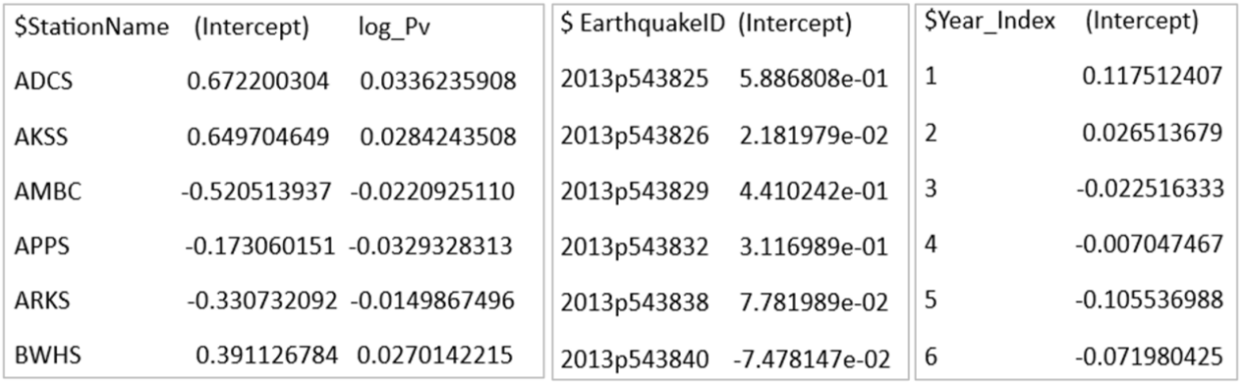
The pooled estimation method can be used in this LMM modelling. Pooled estimation involves estimating the fixed effects (β) using data from all groups combined, while the random effects (bi) are estimated for each group individually. This approach considers the pooled data to capture the overall variability.

6.8.1 Predict new data points in an existing group

To predict random effects for new data points within an existing group, the Best Linear Unbiased Predictor (BLUP) can be applied, utilizing the variance components estimated from the model. The following equation can be used to predict data points within the existing group. The random intercept coefficients ($\hat{b}_0$) and random slope coefficients ($\hat{b}_1$) for the station, Earthquake ID, and Year Index can be extracted accordingly.

$$\log\_\hat{PGV}_{new} = -0.221954 + 0.8527\, log_{P}v_{new} + \hat{b}_{0,StationName} + \hat{b}_{1,StationName} \cdot log_{P}v_{new} + \hat{b}_{0,EarthquakeID} + \hat{b}_{0,Year\_Index}$$

The below figure shows the snippets of the random effect coefficients for a few data points.



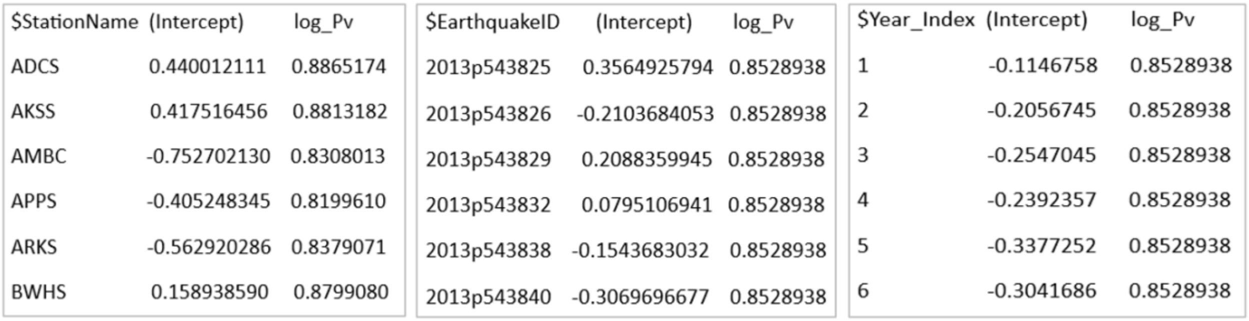
```
$StationName  (Intercept)      log_Pv
ADCS           0.672200304   0.0336235908
AKSS           0.649704649   0.0284243508
AMBC          -0.520513937  -0.0220925110
APPS          -0.173060151  -0.0329328313
ARKS          -0.330732092  -0.0149867496
BWHS           0.391126784   0.0270142215
```

```
$ EarthquakeID  (Intercept)
2013p543825     5.886808e-01
2013p543826     2.181979e-02
2013p543829     4.410242e-01
2013p543832     3.116989e-01
2013p543838     7.781989e-02
2013p543840    -7.478147e-02
```

```
$Year_Index   (Intercept)
1             0.117512407
2             0.026513679
3            -0.022516333
4            -0.007047467
5            -0.105536988
6            -0.071980425
```

Figure 6.10: Snippets of a few random coefficients

Alternatively, the individual model equations for each group can be obtained using the `coef()` function, as illustrated below. The following snippets (figure 6.11) of model coefficients for a few data points, considering select groups.



$StationName	(Intercept)	log_Pv
ADCS	0.440012111	0.8865174
AKSS	0.417516456	0.8813182
AMBC	-0.752702130	0.8308013
APPS	-0.405248345	0.8199610
ARKS	-0.562920286	0.8379071
BWHS	0.158938590	0.8799080

$EarthquakeID	(Intercept)	log_Pv
2013p543825	0.3564925794	0.8528938
2013p543826	-0.2103684053	0.8528938
2013p543829	0.2088359945	0.8528938
2013p543832	0.0795106941	0.8528938
2013p543838	-0.1543683032	0.8528938
2013p543840	-0.3069696677	0.8528938

$Year_Index	(Intercept)	log_Pv
1	-0.1146758	0.8528938
2	-0.2056745	0.8528938
3	-0.2547045	0.8528938
4	-0.2392357	0.8528938
5	-0.3377252	0.8528938
6	-0.3041686	0.8528938

Figure 6.11: Snippets of model equations in group level

6.8.2 Predict new data points in a new group:

Given that linear mixed models typically assume a normal distribution for random effects(as will be demonstrated in Section 6.10.6, "Normality of Random Effects"), this assumption can be leveraged to predict random effects for new groups. The variance-covariance matrix of the random effects can be obtained using `VarCorr(model_ris_1)`. With this matrix, new random effects can be assigned by generating random values from a multivariate normal distribution with mean zero and the extracted variance-covariance matrix. By adding these simulated random effects to the fixed effects predictions for the new group, predictions for data points in new groups can be obtained.

In this specific case, the new random groups are limited to Earthquake ID and Year Index only, as there is no intention to predict intensity at new stations. The goal is to estimate S-wave intensity prediction models for existing stations. Therefore, it is necessary to simulate new random effects only for Year Index and Earthquake IDs. Additionally, these random effects are not correlated, as shown in the model summary (Figure 6.5).

6.9 Comparison of S-wave intensity (log_PGV) prediction: classical regression model, fixed effet and random effect.

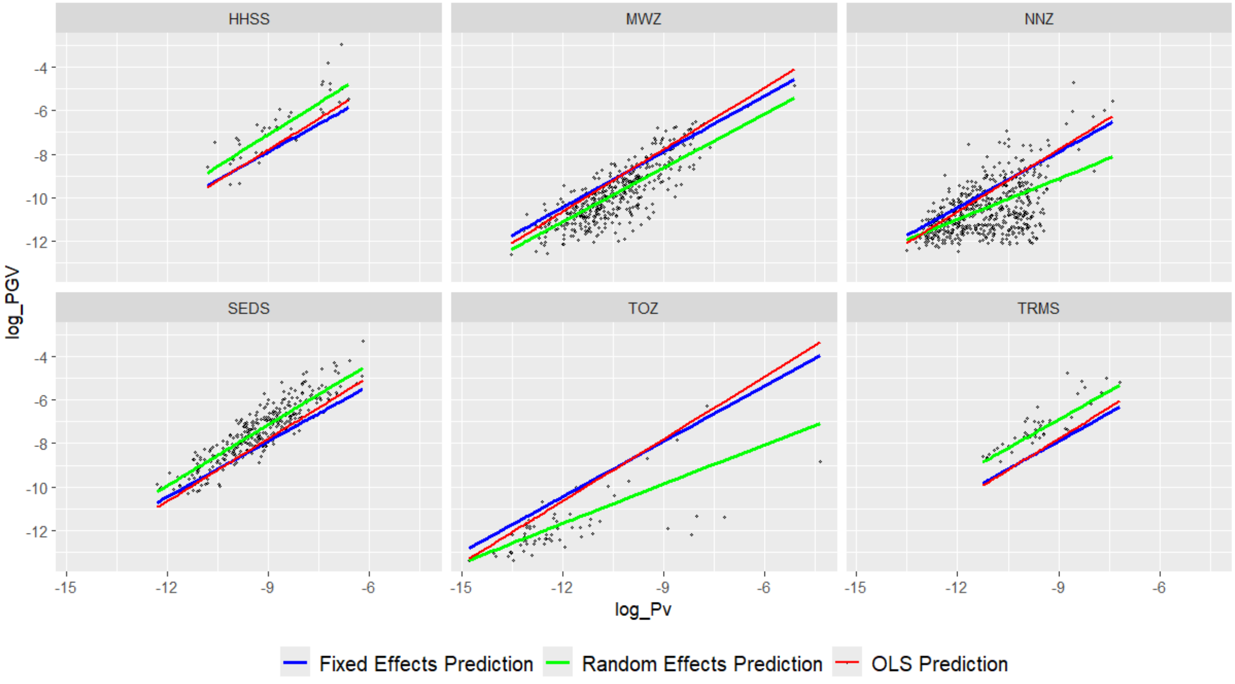


Figure 6.12: Comparison of log_PGV predictions: Fixed effects, random effects and OLS prediction

The figure 6.12 displays scatterplots of log_PGV against log_Pv for six seismic stations, with predictions from Ordinary Least Squares (OLS), Fixed Effects, and Random Effects models. The Random Effects model, represented by the green line, consistently outperforms the OLS (red line) and Fixed Effects (blue line) models by accounting for station-specific variations. This is particularly evident in stations like NNZ and TOZ, where there is more variability in the data, and the Random Effects model adjusts for these differences, resulting in a better fit. In contrast, the OLS model, which does not account for such variations, shows larger deviations, particularly at extreme values of log_Pv. By incorporating random effects, the model can capture between-station differences, improving the accuracy of S-wave intensity predictions across different seismic stations. This highlights the importance of considering station-specific characteristics when modeling earthquake intensity, as it leads to more robust and reliable predictions.

Figure 6.13 illustrates the individual models that can be derived for all the stations in New Zealand.

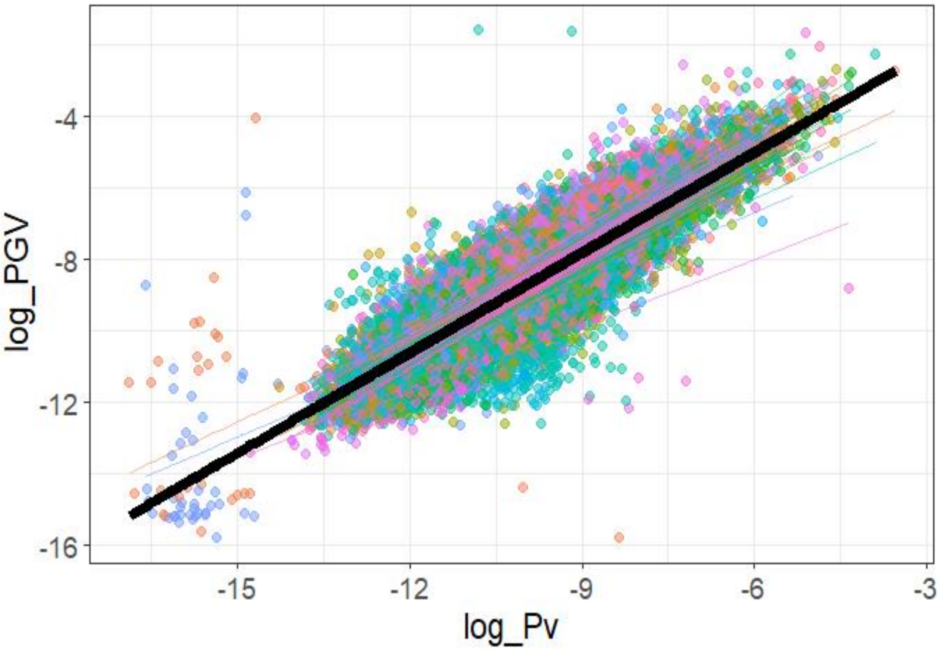


Figure 6.13: Scatter plot with independent linear mixed effect model fits for all stations: Fixed and random effect by station: fixed effect (heavy black line) and random effect (thin lines)

The graph presents a scatterplot of log_PGV against log_Pv, illustrating the relationship between these two variables across 293 stations. The data points are color-coded to represent different stations, with each color corresponding to a specific cluster. The thick black line represents the overall trend or the best fit across all data, showing a positive linear relationship between log_Pv and log_PGV. While the majority of the data follows this trend, some outliers can be observed, particularly at extreme values of log_Pv, which suggests variability in the data. The inclusion of this thick trend line highlights the general linear relationship, while thinner, more colored lines seem to indicate individual regressions for different stations, demonstrating variability in slopes and intercepts. This visualization reinforces the need to account for random effects, as individual stations have distinct behaviors that differ from the overall trend.

6.10 Assessment of LMM model assumptions

6.10.1 Checking linearity and independence of residuals

The residual plot in figure 6.14, shows that nearly all the residuals are randomly scattered around the 0-axis, forming a horizontal band. The green reference line is flat and horizontal, indicating that the assumption of linearity of residuals is satisfied.

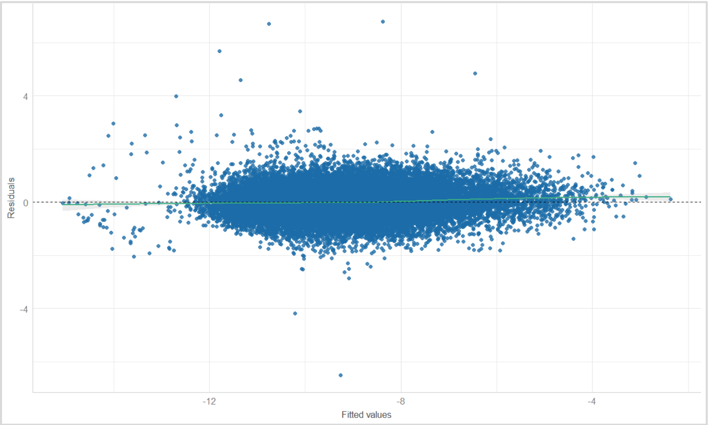


Figure 6.14: Residual vs fitted values plot

6.10.2 Normality of residuals – Q-Q plot for residuals

The normal Q-Q plot in figure 6.15, can be used to get an idea of how the residuals will align with the quantiles of the standard normal distribution. While the majority of the points fall closely along the fitted line, indicating that the residuals largely follow a normal distribution, there are some deviations at the extremes, suggesting potential minor violations of the normality assumption.

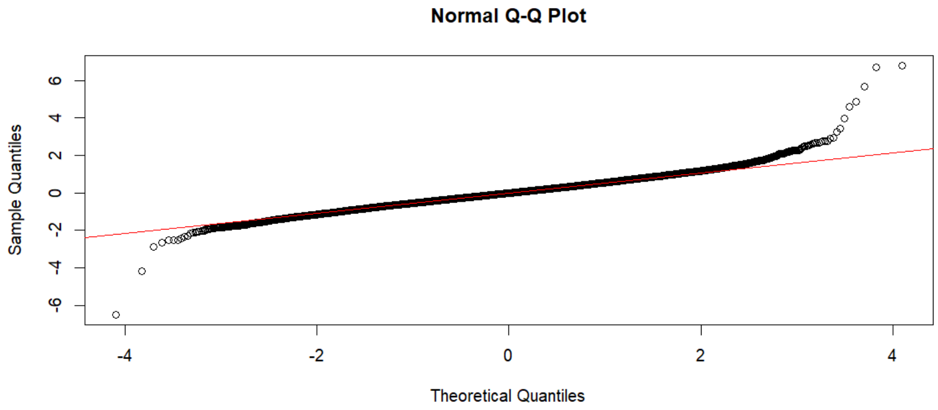


Figure 6.15: The residuals of the model "model_ris_1" plotted against the quantiles of a standard normal distribution.

6.10.3 Histogram of residuals

The histogram in figure 6.16, provides further evidence that the residuals are normally distributed.

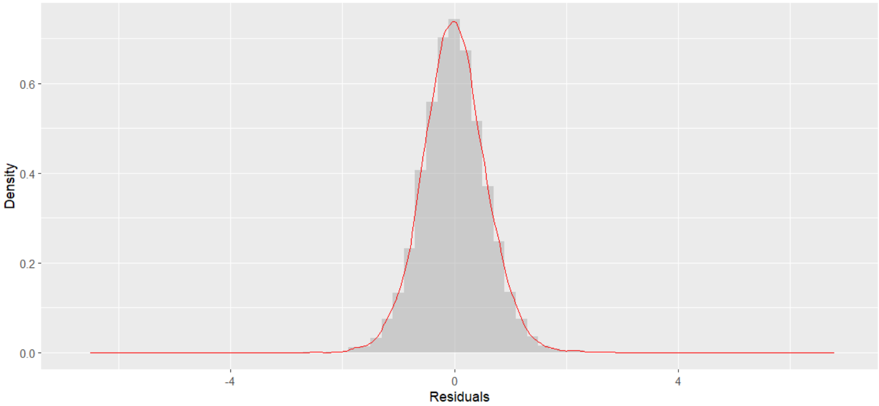


Figure 6.16: Histogram: Distribution of residuals

6.10.4 Independence of residuals

The graph in figure 6.17, indicates no systematic pattern in the residuals over the order of observations. The residuals appear to be randomly scattered around the zero line, suggesting that the assumption of independence is satisfied.

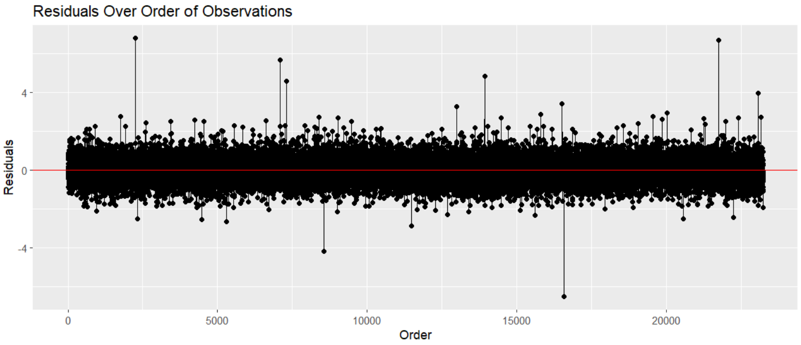


Figure 6.17: Residuals over order of observation

6.10.5 Homoscedasticity

The graph in figure 6.18, shows that the spread of the residuals is consistent across the range of fitted values, with no clear pattern like a funnel shape or trend indicating increasing or decreasing spread. Additionally, the reference line is approximately flat and horizontal, suggesting that the variance of the residuals is constant. Based on this plot, the assumption of homoscedasticity appears to be satisfied.

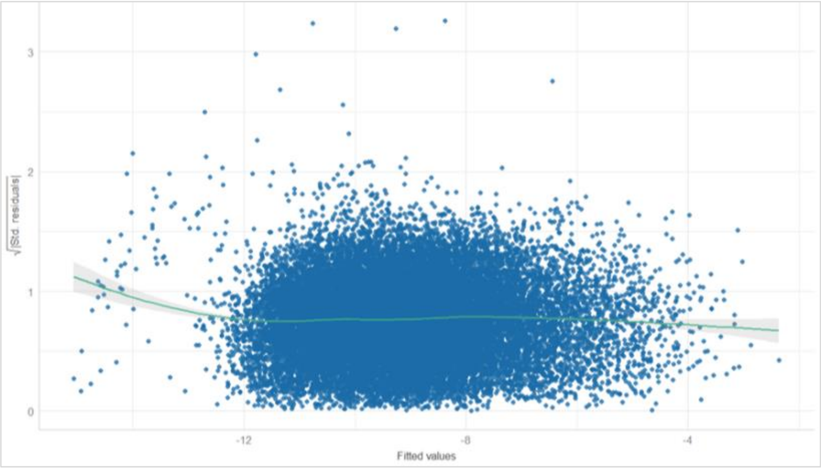


Figure 6.18: Square root of standard residuals vs fitted values

6.10.6 Normality of random effects

The two graphs in figure 6.19 illustrate how a Q-Q plot can be used to assess the normality of random effects: Station (Intercept and Slope). In the intercept plot, most points align closely with the diagonal line, indicating that the random intercept effects for the majority of stations approximately follow a normal distribution. However, some deviations at the extremes suggest minor violations of the normality assumption. Similarly, the random slope effects for most stations generally follow a normal distribution, with slight deviations at the tails.

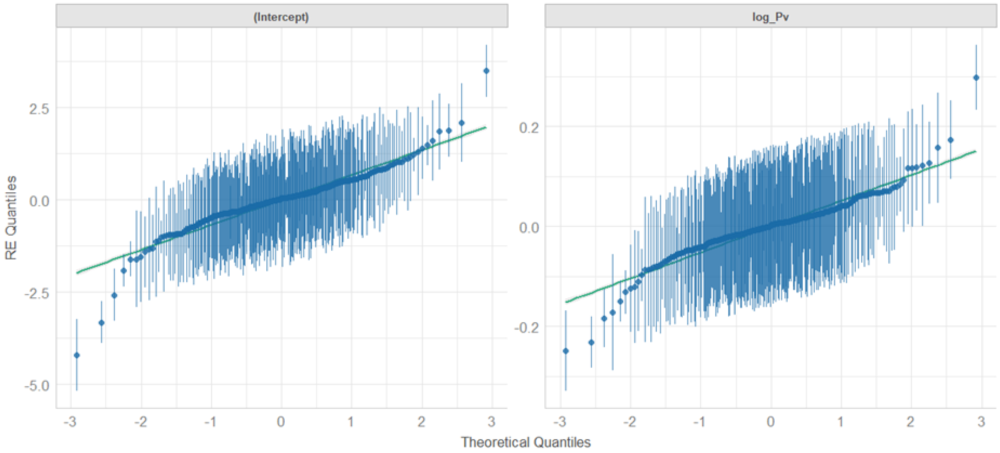


Figure 6.19: Q-Q plot for station effect (intercept and slope)

According to figure 6.20, the Q-Q plot for random effect: EarthquakeID (intercept) shows that the random intercept effects for most earthquake events are approximately normally distributed, though there are some minor deviations at the extremes.

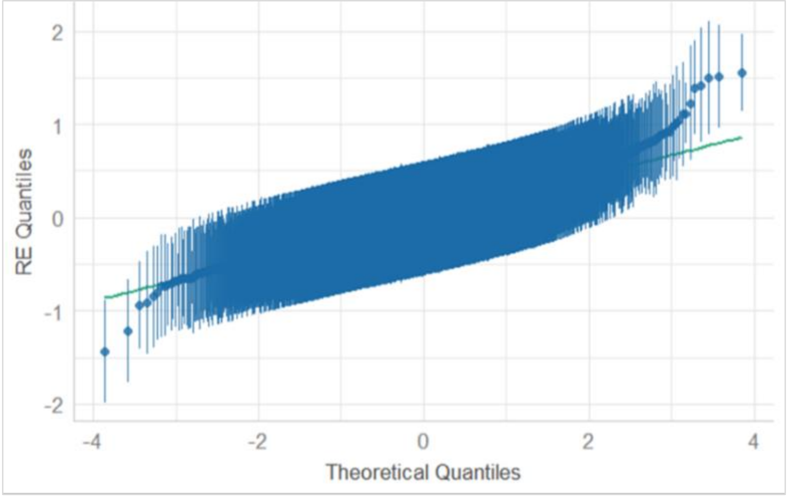


Figure 6.20: Q-Q plot for earthquake-id effect (intercept)

According to the Q-Q plot (figure 6.21) for random effect: Year_Index (intercept), the points closely follow the diagonal line, indicating that the random effects for Year_Index are approximately normally distributed.

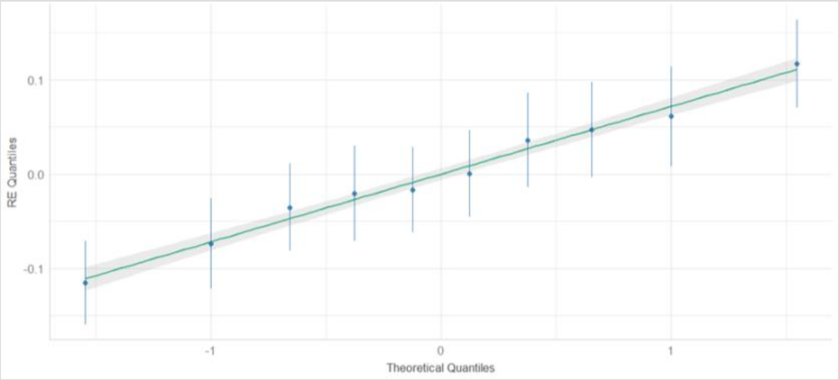


Figure 6.21: Q-Q plot for year index effect (intercept)

6.10.7 Posterior predictive check

The graph in Figure 6.22, observed during the posterior predictive check of the model (model_ris_1), indicates that the model-predicted data closely follows the observed data in the test set. This suggests a good fit of the model for predicting S-wave intensity (log_PGV).

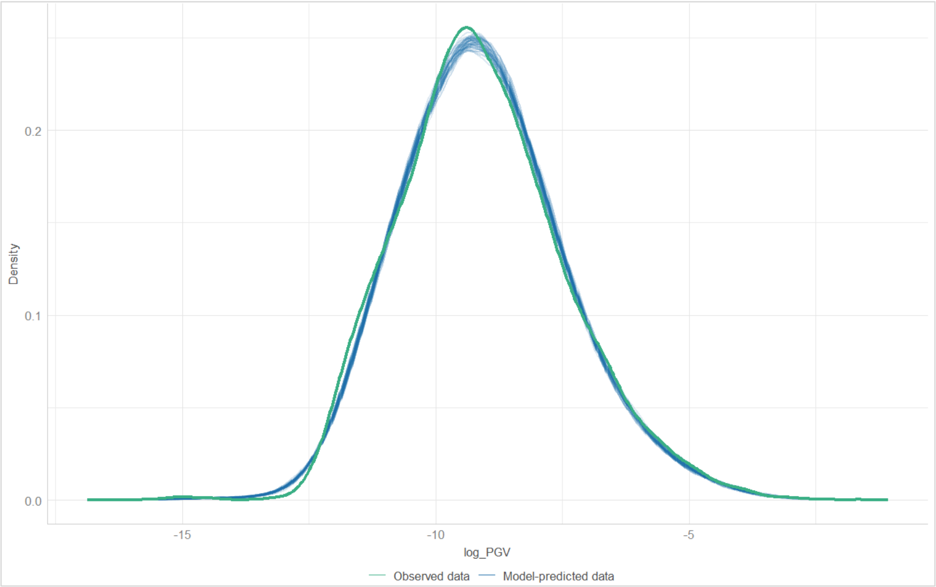


Figure 6.22: Observed data vs model predicted data

7. Discussion and Conclusions

7.1 Discussion

This analysis aimed to lay the groundwork for developing station-specific S-wave intensity estimation models for New Zealand using historical earthquake data. The initial task involved gaining domain knowledge in seismology, particularly in earthquake science, through an extensive literature review. This review also covered global techniques for estimating S-wave intensity and identified research gaps.

The primary task involved creating the necessary dataset and calculating required parameters, including Peak Ground Acceleration (PGA), Peak Ground Velocity (PGV), and Peak Ground Displacement (PGD) for S-waves, as well as Peak Ground Acceleration (Pa), Peak Ground Velocity (Pv), and Peak Ground Displacement (Pd) for P-waves, using seismic waveform data from GeoNet. This process included filtering out missing values by excluding data entries lacking either P or S wave pick times.

After developing the dataset, the most suitable measurement for S-wave intensity (response variable) was identified as PGV for New Zealand's data. This decision was based on the literature review and correlation analysis between S-wave intensity measurements (PGA, PGV, PGD) and P-wave parameters (Pa, Pv, Pd). During data preprocessing, records within a 100km epicentral distance limit were selected to ensure precise estimation of S-wave intensity values based on P-wave parameters. This filtering process resulted in the final selection of 9,206 earthquake events, producing 29,058 earthquake ground motion records for comprehensive analysis. Categorical variables in the dataset were transformed into factors, and year indexes were created from the P-wave pick time variable for future analysis.

In the preliminary analysis, potential outliers were identified, and one outlier was removed after careful consideration. The remaining outliers were deemed actual values. S-wave intensity parameter (PGV) and P-wave parameters (Pa, Pv, Pd) were transformed into logarithmic values to standardize the highly varying raw data, ensuring proper training on estimation models. This transformation effectively compressed data with a wide range of values, reducing the impact of extreme outliers and enhancing the robustness of the relationships. The dataset contained multiple records for the same earthquake event from different stations (repeated measures), resulting in correlated data. The non-independence of this data was confirmed through the necessary visualizations. The behavior of the response variable (log_PGV) was also examined, suggesting a normal-like distribution with some outliers.

In the advanced analysis, the best predictor variable among the three log-transformed P-wave parameters (log_Pa, log_Pv, log_Pd) was identified by investigating the relationship with the response variable, S-wave intensity (log_PGV). R^2 values, RMSE, and MAPE were used to assess the performance of the modeled relationships, with log(Pv) being selected as the predictor variable for estimating log(PGV).

A simple linear regression model was constructed to regress the response variable log(PGV) on the selected predictor variable log(Pv). The results of this model discussed and highlighted violations of model assumptions due to data non-independence, indicating the necessity of a Linear Mixed Effects Model (LMM). The necessity of LMM was assessed using ICC and Design Effect values on the unconditional mean model, with Station Name, Earthquake ID, and Year Index as random effects. Several random intercept and slope models were built to find a feasible model, and the best model was chosen based on AIC, BIC, Log-likelihood, Chi-squared, and p-values, along with model complexity and convergence. The results were cross-validated using MSE, RMSE, and MAPE values on the test set. The overall model was then constructed based on the chosen best model coefficients, and station-specific models were derived by retrieving coefficients of the random effects for estimation on both existing and new data. Finally, LMM model assumptions were checked to validate the model estimations.

7.2 Conclusion

- The analysis identified log(Pv) as the most suitable predictor for estimating S-wave intensity (log(PGV)).
- The classical linear regression model showed statistical significance (F-statistic: 7.126e+04, p-value < 2.2e-16) and explained 75.4% of the variance in log(PGV). However, this model failed to account for data non-independence and the nested data structure.

- To address the limitations of the classical regression model, Linear Mixed Models (LMM) were employed. The unconditional mean model of LMM revealed significant variance between stations (ICC = 0.459) and earthquakes (ICC = 0.610), indicating data clustering.

- The model with random intercepts and slopes for stations, and random intercepts for earthquake-id and year-index significantly improved the model fit (p-value < 2.2e-16), achieving higher R-squared values (conditional: 0.826, marginal: 0.718). This model was chosen as the best for the analysis due to having the lowest AIC (50814.5) and BIC (50879.0). Comparing the classical regression model to the chosen LMM showed an improvement in the R-squared value by 7%. Cross-validation with the test set demonstrated that the LMM significantly reduced RMSE and MAPE values compared to the ordinary regression model.

- The chosen LMM provides a more robust and accurate approach for estimating log(PGV) by accounting for correlated errors and station-specific variations. By incorporating random effects for station, earthquake-id, and year-index, the model captures variations specific to each station and earthquake event, leading to more accurate estimations and credible intervals. This improved accuracy helps in understanding baseline variations in S-wave intensity associated with different earthquakes, stations, and years. This model's station-level application enhance the accuracy of ground-shaking intensity predictions, potentially improving the reliability of warning systems and public safety.

7.3 Suggestions for future work

- Investigation of advanced machine learning models: Future research can explore the application of more complex machine learning models, such as deep learning and ensemble methods, to enhance the accuracy of earthquake intensity estimations. By leveraging advanced algorithms and larger datasets, the predictive performance and reliability of the intensity estimation models are aimed to be improved.

- Real-time inference adaptation: It is essential to investigate methods for adapting the developed models to provide real-time inferences and outputs. This capability will facilitate the seamless integration of the models with existing EEW systems, enabling timely and accurate warnings that can enhance disaster response and mitigation efforts.

8. References

Abdalzaher, M. S., Soliman, M. S., & El-Hady, S. M. (2023a). Seismic Intensity Estimation for Earthquake Early Warning Using Optimized Machine Learning Model. *IEEE Transactions on Geoscience and Remote Sensing*, *61*. https://doi.org/10.1109/TGRS.2023.3296520

Abdalzaher, M. S., Soliman, M. S., & El-Hady, S. M. (2023b). Seismic Intensity Estimation for Earthquake Early Warning Using Optimized Machine Learning Model. *IEEE Transactions on Geoscience and Remote Sensing*, *61*. https://doi.org/10.1109/TGRS.2023.3296520

Abdalzaher, M. S., Soliman, M. S., Krichen, M., Alamro, M. A., & Fouda, M. M. (2024). Employing Machine Learning for Seismic Intensity Estimation Using a Single Station for Earthquake Early Warning. *Remote Sensing*, *16*(12). https://doi.org/10.3390/rs16122159

Ahmadzadeh, S., Doloei, G. J., & Zafarani, H. (2020). Ground Motion to Intensity Conversion Equations for Iran. *Pure and Applied Geophysics*, *177*(11). https://doi.org/10.1007/s00024-020-02586-x

Akkar, S., & Özen, Ö. (2005). Effect of peak ground velocity on deformation demands for SDOF systems. *Earthquake Engineering and Structural Dynamics*, *34*(13). https://doi.org/10.1002/eqe.492

Allen, R. M., & Melgar, D. (2019a). Earthquake early warning: Advances, scientific challenges, and societal needs. In *Annual Review of Earth and Planetary Sciences* (Vol. 47). https://doi.org/10.1146/annurev-earth-053018-060457

Allen, R. M., & Melgar, D. (2019b). Earthquake early warning: Advances, scientific challenges, and societal needs. In *Annual Review of Earth and Planetary Sciences* (Vol. 47). https://doi.org/10.1146/annurev-earth-053018-060457

Baltagi, B. H., & Liu, L. (2020). Forecasting with unbalanced panel data. *Journal of Forecasting*, *39*(5). https://doi.org/10.1002/for.2646

Becker, J. S., Potter, S. H., Prasanna, R., Tan, M. L., Payne, B. A., Holden, C., Horspool, N., Smith, R., & Johnston, D. M. (2020a). Scoping the potential for earthquake early warning in Aotearoa New Zealand: A sectoral analysis of perceived benefits and challenges. *International Journal of Disaster Risk Reduction*, *51*. https://doi.org/10.1016/j.ijdrr.2020.101765

Becker, J. S., Potter, S. H., Prasanna, R., Tan, M. L., Payne, B. A., Holden, C., Horspool, N., Smith, R., & Johnston, D. M. (2020b). Scoping the potential for earthquake early warning in Aotearoa New Zealand: A sectoral analysis of perceived benefits and challenges. *International Journal of Disaster Risk Reduction*, *51*. https://doi.org/10.1016/j.ijdrr.2020.101765

Böse, M., Wenzel, F., & Erdik, M. (2008). PreSEIS: A neural network-based approach to earthquake early warning for finite faults. *Bulletin of the Seismological Society of America*, *98*(1). https://doi.org/10.1785/0120070002

Bradley, B. A., Cubrinovski, M., MacRae, G. A., & Dhakal, R. P. (2009). Ground-motion prediction equation for SI based on spectral acceleration equations. *Bulletin of the Seismological Society of America*, *99*(1). https://doi.org/10.1785/0120080044

Carle, C. F., James, A. C., & Maddess, T. (2013). The pupillary response to color and luminance variant multifocal stimuli. *Investigative Ophthalmology and Visual Science*, *54*(1). https://doi.org/10.1167/iovs.12-10829

Caruso, A., Colombelli, S., Elia, L., Picozzi, M., & Zollo, A. (2017). An on-site alert level early warning system for Italy. *Journal of Geophysical Research: Solid Earth*, *122*(3), 2106–2118. https://doi.org/10.1002/2016JB013403

Chandrakumar, C., Tan, M. L., Holden, C., Stephens, M., & Punchihewa, A. (2024). *Estimating S-wave Amplitude for Earthquake Early Warning in New Zealand: Leveraging the First 3 Seconds of P-Wave*. https://doi.org/10.21203/rs.3.rs-4475416/v1

Chiang, Y. J., Chin, T. L., & Chen, D. Y. (2022). Neural Network-Based Strong Motion Prediction for On-Site Earthquake Early Warning. *Sensors*, *22*(3). https://doi.org/10.3390/s22030704

Fayaz, J., & Galasso, C. (2023). A deep neural network framework for real-time on-site estimation of acceleration response spectra of seismic ground motions. *Computer-Aided Civil and Infrastructure Engineering*, *38*(1). https://doi.org/10.1111/mice.12830

Festa, G., Zollo, A., Picozzi, M., Colombelli, S., Elia, L., & Caruso, A. (2022). Earthquake Early Warning Systems: Methodologies, Strategies, and Future Challenges. *Advances in Science, Technology and Innovation*, 193–196. https://doi.org/10.1007/978-3-030-73026-0_44

Franco, C., Kausar, S., Silva, M. F. B., Guedes, R. C., Falcao, A. O., & Brito, M. A. (2022). Multi-Targeting Approach in Glioblastoma Using Computer-Assisted Drug Discovery Tools to Overcome the Blood–Brain Barrier and Target EGFR/PI3Kp110β Signaling. *Cancers*, *14*(14). https://doi.org/10.3390/cancers14143506

GeoNet. (n.d.). *GeoNet FDSN Web Service*. Retrieved September 16, 2024, from https://www.geonet.org.nz/data/access/FDSN

GeoNet Recent Quakes. (n.d.).

GNS Science. (n.d.). *Natural Hazards and Risks*.

Grandin, R., Borges, J. F., Bezzeghoud, M., Caldeira, B., & Carrilho, F. (2007). Simulations of strong ground motion in SW Iberia for the 1969 February 28 (Ms = 8.0) and the 1755 November 1 (M ~ 8.5) earthquakes - II. Strong ground motion simulations. *Geophysical Journal International*, *171*(2). https://doi.org/10.1111/j.1365-246X.2007.03571.x

Hou, B., Li, S., & Song, J. (n.d.). Support Vector Machine-Based On-Site Prediction for China Seismic Instrumental Intensity from P-Wave Features. *Pure and Applied Geophysics*, *180*. https://doi.org/10.1007/s00024

Hsu, T. Y., & Huang, C. W. (2021a). Onsite Early Prediction of PGA Using CNN With Multi-Scale and Multi-Domain P-Waves as Input. *Frontiers in Earth Science*, *9*. https://doi.org/10.3389/feart.2021.626908

Hsu, T. Y., & Huang, C. W. (2021b). Onsite Early Prediction of PGA Using CNN With Multi-Scale and Multi-Domain P-Waves as Input. *Frontiers in Earth Science*, *9*. https://doi.org/10.3389/feart.2021.626908

Hsu, T. Y., & Nieh, C. P. (2020a). On-site earthquake early warning using smartphones. *Sensors (Switzerland)*, *20*(10). https://doi.org/10.3390/s20102928

Hsu, T. Y., & Nieh, C. P. (2020b). On-site earthquake early warning using smartphones. *Sensors (Switzerland)*, *20*(10). https://doi.org/10.3390/s20102928

Hsu, T. Y., & Pratomo, A. (2022). Early Peak Ground Acceleration Prediction for On-Site Earthquake Early Warning Using LSTM Neural Network. *Frontiers in Earth Science*, *10*. https://doi.org/10.3389/feart.2022.911947

Huang, J. Z., Chen, M., Maadooliat, M., & Pourahmadi, M. (2012). A cautionary note on generalized linear models for covariance of unbalanced longitudinal data. *Journal of Statistical Planning and Inference*, *142*(3). https://doi.org/10.1016/j.jspi.2011.09.011

Idha, R., Sari, E. P., Humaidi, S., Simanjuntak, A. V. H., & Muksin, U. (2023). Response of Geologic Units to The Ground Parameters of Tarutung Earthquake 2022 Mw 5.8: A Preliminary Study. *IOP Conference Series: Earth and Environmental Science*, *1288*(1). https://doi.org/10.1088/1755-1315/1288/1/012032

J., C.-C., Lin, P.-Y., Chang, T.-M., Lin, T.-K., Weng, Y.-T., Chang, K.-C., & Tsai, K.-C. (2012). Development of On-Site Earthquake Early Warning System for Taiwan. In *Earthquake Research and Analysis - New Frontiers in Seismology*. https://doi.org/10.5772/28056

Jayaram, N., & Baker, J. W. (2009). Correlation model for spatially distributed ground-motion intensities. *Earthquake Engineering and Structural Dynamics*, *38*(15). https://doi.org/10.1002/eqe.922

Johnson, C. W., Kilb, D., Baltay, A., & Vernon, F. (2020). Peak Ground Velocity Spatial Variability Revealed by Dense Seismic Array in Southern California. *Journal of Geophysical Research: Solid Earth*, *125*(6). https://doi.org/10.1029/2019JB019157

Kanamori, H., Hauksson, E., & Heaton, T. (1997). Real-time seismology and earthquake hazard mitigation. In *Nature* (Vol. 390, Issue 6659). https://doi.org/10.1038/37280

Kubokawa, T. (2010). Corrected empirical Bayes confidence intervals in nested error regression models. *Journal of the Korean Statistical Society*, *39*(2). https://doi.org/10.1016/j.jkss.2009.08.001

Kuehn, N. M., & Scherbaum, F. (2015). Ground-motion prediction model building: a multilevel approach. *Bulletin of Earthquake Engineering*, *13*(9). https://doi.org/10.1007/s10518-015-9732-3

Lara, P., Bletery, Q., Ampuero, J., Inza, A., & Tavera, H. (2023). Earthquake Early Warning Starting From 3 s of Records on a Single Station With Machine Learning. *Journal of Geophysical Research: Solid Earth*, *128*(11). https://doi.org/10.1029/2023JB026575

Learn NZ. (n.d.). *Earthquakes in New Zealand*.

Liang, H. J., Li, J. L., Di, Y. L., Zhang, A. S., & Zhu, F. X. (2015). Logarithmic Transformation is Essential for Statistical Analysis of Fungicide EC50 Values. *Journal of Phytopathology*, *163*(6). https://doi.org/10.1111/jph.12342

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020a). Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-17591-w

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020b). Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-17591-w

Nakayachi, K., Becker, J. S., Potter, S. H., & Dixon, M. (2019). Residents' Reactions to Earthquake Early Warnings in Japan. *Risk Analysis*, *39*(8). https://doi.org/10.1111/risa.13306

National Emerency Management Agency. (n.d.). *Canterbury Earthquake*.

Peng, C., Jiang, P., Chen, Q., Ma, Q., & Yang, J. (2019). Performance evaluation of a dense MEMS-based seismic sensor array deployed in the Sichuan-Yunnan border region for earthquake early warning. *Micromachines*, *10*(11). https://doi.org/10.3390/mi10110735

RUSU, V. D., & ROMAN, A. (2018). THE FINANCING BEHAVIOUR OF THE CEE FIRMS UNDER THE IMPACT OF FINANCIAL CRISIS: A PANEL REGRESSION APPROACH. *European Journal of Sustainable Development*, *7*(1). https://doi.org/10.14207/ejsd.2018.v7n1p329

Sarkar, S., Kumar, S., Roy, A., & Das, B. (2023). Multilayer Perceptron Based Early On-Site Estimation of PGA During an Earthquake. *Lecture Notes in Electrical Engineering*, *947*, 313–326. https://doi.org/10.1007/978-981-19-5936-3_29

Sarkar, S., Roy, A., Das, B., & Kumar, S. (2023). Neuroevolution-Based Earthquake Intensity Classification for Onsite Earthquake Early Warning. *Lecture Notes in Electrical Engineering*, *946*, 345–356. https://doi.org/10.1007/978-981-19-5868-7_26

Sarkar, S., Roy, A., Kumar, S., & Das, B. (2022a). Seismic Intensity Estimation Using Multilayer Perceptron for Onsite Earthquake Early Warning. *IEEE Sensors Journal*, *22*(3). https://doi.org/10.1109/JSEN.2021.3137395

Sarkar, S., Roy, A., Kumar, S., & Das, B. (2022b). Seismic Intensity Estimation Using Multilayer Perceptron for Onsite Earthquake Early Warning. *IEEE Sensors Journal*, *22*(3), 2553–2563. https://doi.org/10.1109/JSEN.2021.3137395

Shearer, P. M. (2009). Introduction to seismology, second edition. *Cambridge University Press*, *4*(3).

Stoppa, F., & Berti, C. (2013a). Reducing seismic risk by understanding its cultural roots: Inference from an Italian case history. *Natural Science*, *05*(08). https://doi.org/10.4236/ns.2013.58a1010

Stoppa, F., & Berti, C. (2013b). Reducing seismic risk by understanding its cultural roots: Inference from an Italian case history. *Natural Science*, *05*(08). https://doi.org/10.4236/ns.2013.58a1010

Strauss, J. A., & Allen, R. M. (2016). Benefits and costs of earthquake early warning. *Seismological Research Letters*, *87*(3). https://doi.org/10.1785/0220150149

Vinnell, L. J., Tan, M. L., Prasanna, R., & Becker, J. S. (2023). Knowledge, perceptions, and behavioral responses to earthquake early warning in Aotearoa New Zealand. *Frontiers in Communication*, *8*. https://doi.org/10.3389/fcomm.2023.1229247

Wald, D. J., Quitoriano, V., Heaton, T. H., & Kanamori, H. (1999). Relationships between peak ground acceleration, peak ground velocity, and modified mercalli intensity in California. *Earthquake Spectra*, *15*(3). https://doi.org/10.1193/1.1586058

Wu, Y. M., Hsiao, N. C., & Teng, T. L. (2004). Relationships between strong ground motion peak values and seismic loss during 1999 Chi-Chi, Taiwan Earthquake. *Natural Hazards*, *32*(3). https://doi.org/10.1023/B:NHAZ.0000035550.36929.d0

Wu, Y. M., Kanamori, H., Allen, R. M., & Hauksson, E. (2007). Determination of earthquake early warning parameters, τc and Pd, for southern California. *Geophysical Journal International*, *170*(2). https://doi.org/10.1111/j.1365-246X.2007.03430.x

Wu, Y. M., & Mittal, H. (2021). A review on the development of earthquake warning system using low-cost sensors in taiwan. In *Sensors* (Vol. 21, Issue 22). https://doi.org/10.3390/s21227649

Wu, Y. M., Mittal, H., Huang, T. C., Yang, B. M., Jan, J. C., & Chen, S. K. (2019). Performance of a low-cost earthquake early warning system (P-ALErt) and shake map production during the 2018 M w 6.4 Hualien, Taiwan, earthquake. In *Seismological Research Letters* (Vol. 90, Issue 1). https://doi.org/10.1785/0220180170

Wu, Y. M., Teng, T. liang, Shin, T. C., & Hsiao, N. C. (2003a). Relationship between peak ground acceleration, peak ground velocity, and intensity in Taiwan. *Bulletin of the Seismological Society of America*, *93*(1). https://doi.org/10.1785/0120020097

Wu, Y. M., Teng, T. liang, Shin, T. C., & Hsiao, N. C. (2003b). Relationship between peak ground acceleration, peak ground velocity, and intensity in Taiwan. *Bulletin of the Seismological Society of America*, *93*(1). https://doi.org/10.1785/0120020097

www.building.govt.nz. (n.d.). *New seismic risk guidance released*.

Xia, J., Li, Y., Cheng, Y., Li, J., & Tian, S. (2021). Research on compressive sensing of strong earthquake signals for earthquake early warning. *Geomatics, Natural Hazards and Risk*, *12*(1). https://doi.org/10.1080/19475705.2021.1889689

Yang, X., Chen, Y., Teng, S., & Chen, G. (2021). A novel method for predicting local site amplification factors using 1-D convolutional neural networks. *Applied Sciences (Switzerland)*, *11*(24). https://doi.org/10.3390/app112411650

Yuniarti, D., Rosadi, D., & Abdurakhman. (2022). Application of Groupwise Principal Sensitivity Components on Unbalanced Panel Data Regression Model for Gross Regional Domestic Product in Kalimantan. *Pertanika Journal of Science and Technology*, *30*(4). https://doi.org/10.47836/pjst.30.4.01

Zollo, A., Amoroso, O., Lancieri, M., Wu, Y. M., & Kanamori, H. (2010). A threshold-based earthquake early warning using dense accelerometer networks. *Geophysical Journal International*, *183*(2). https://doi.org/10.1111/j.1365-246X.2010.04765.x

Zollo, A., Colombelli, S., Caruso, A., & Elia, L. (2023). An Evolutionary Shaking-Forecast-Based Earthquake Early Warning Method. *Earth and Space Science*, *10*(4). https://doi.org/10.1029/2022EA002657

9. Apendices

9.1 Appendix 1 – R Script

```
install.packages("lme4")
install.packages("lmerTest")
install.packages("dplyr")
install.packages("lattice")
install.packages("car")
install.packages("nlme")
install.packages("performance")
install.packages("sjPlot")
install.packages("glmmTMB")
install.packages("flexplot")
install.packages("devtools")
install.packages("emmeans")

library(lme4)
library(lmerTest)
library(dplyr)
library(lattice)
library(car)
library(nlme)
library(performance)
library(sjPlot)
library(glmmTMB)
library(ggplot2)
library(sjPlot)
library(devtools)
library(emmeans)
library(flexplot)

setwd("C:/Users/thanu/Desktop/MSc Final Research/Rscript")
eqdata<-read.csv('eqdata.csv',check.names=FALSE)
head(eqdata)
str(eqdata)

eqdata$EarthquakeID<-factor(eqdata$EarthquakeID)
eqdata$EarthquakeKey<-factor(eqdata$EarthquakeKey)
eqdata$StationName<-factor(eqdata$StationName)
eqdata$Year_Index<-factor(eqdata$Year_Index)
eqdata$Month_Index<-factor(eqdata$Month_Index)

# Split data into training and testing sets
set.seed(123)
train_index <- sample(seq_len(nrow(eqdata)), size = 0.8 * nrow(eqdata))
train_data <- eqdata[train_index, ]
test_data <- eqdata[-train_index, ]
```

```
# ordinary least-squares (OLS) multiple regression / Classical regression
models
model_ols<-lm(log_PGV ~ log_Pv, data = train_data)
summary(model_ols)
plot(model_ols)
AIC(model_ols)
BIC(model_ols)

# Linear mixed effect modeling optimizer
control <- lmerControl(optimizer = "bobyqa")

#Null model
model_null <- lmer(log_PGV ~ 1 + (1|StationName) + (1| EarthquakeID) + (1|
Year_Index), data = train_data, REML = FALSE, control = control)
summary(model_null)
anova(model_null)
performance::r2(model_null)
dotplot(ranef(model_null,condVar=TRUE))

# Extract variance components
variance_components       <- as.data.frame(VarCorr(null_model))
var_between_station       <- variance_components$vcov[variance_components$grp
== "StationName"]
var_between_earthquake_id <- variance_components$vcov[variance_components$grp
== "EarthquakeID"]
var_between_year_index    <- variance_components$vcov[variance_components$grp
== "Year_Index"]
var_within                <- attr(VarCorr(null_model), "sc")^2  # Residual
variance

# Calculate ICC for each
ICC_station               <- var_between_station / (var_between_station +
var_within)
ICC_earthquake_id         <- var_between_earthquake_id /
(var_between_earthquake_id + var_within)
ICC_year_index            <- var_between_year_index / (var_between_year_index
+ var_within)

# Print ICC values
print(paste("ICC for station:", ICC_station))
print(paste("ICC for earthquake ID:", ICC_earthquake_id))
print(paste("ICC for year index:", ICC_year_index))

# Calculate the average cluster size for each grouping variable
average_cluster_size_station       <- mean(tapply(eqdata$log_PGV,
eqdata$StationName, length))
average_cluster_size_earthquake_id <- mean(tapply(eqdata$log_PGV,
eqdata$EarthquakeID, length))
average_cluster_size_year_index    <- mean(tapply(eqdata$log_PGV,
eqdata$Year_Index, length))
```

```
# Calculate the design effect for each grouping variable
design_effect_station       <- 1 + (average_cluster_size_station - 1) *
ICC_station
design_effect_earthquake_id <- 1 + (average_cluster_size_earthquake_id - 1) *
ICC_earthquake_id
design_effect_year_index    <- 1 + (average_cluster_size_year_index - 1) *
ICC_year_index

# Print design effects
print(paste("Design effect for station:", design_effect_station))
print(paste("Design effect for site class:", design_effect_site_class))
print(paste("Design effect for earthquake ID:", design_effect_earthquake_id))
print(paste("Design effect for year index:", design_effect_year_index))

#Random Intercept model
model_ri <- lmer(log_PGV ~ 1 + log_Pv + (1|StationName) + (1| EarthquakeID) +
(1| Year_Index)  , data = train_data, REML = FALSE, control = control)
summary(model_ri)
anova(model_ri)
performance::r2(model_ri)
dotplot(ranef(model_ri,condVar=TRUE))
plot(model_ri)


#Random slope model
model_ris_1 <- lmer(log_PGV ~ 1 + log_Pv + (1 + log_Pv|StationName) + (1|
EarthquakeID) + (1| Year_Index), data = train_data, REML = FALSE, control =
control)
summary(model_ris_1)
anova(model_ris_1)
performance::r2(model_ris_1)
dotplot(ranef(model_ris_1$StationName,condVar=TRUE))

#Extract fixed effects of ris1
fixed_effects_model_ris_1  <- summary(model_ris_1)$coefficients
print(fixed_effects_model_ris_1)

# Extract random effects variance components of ris1
ranef_var_model_ris_1 <- VarCorr(model_ris_1)
print(ranef_var_model_ris_1)

# standard error of residuals
residuals_ris_1 <- resid(model_ris_1)
sd_error_ris_1 <- sd(residuals_ris_1)
print(sd_error_ris_1)

# model_ri and ris_1 comparison
anova(model_ri,model_ris_1)

# second random slop model: ris_2
model_ris_2 <- lmer(log_PGV ~ 1 + log_Pv + (1 + log_Pv|StationName) +(1 +
log_Pv| Year_Index)+ (1| EarthquakeID), data = train_data, REML = FALSE,
control = control)
```

```r
summary(model_ris_2)
anova(model_ris_2)
performance::r2(model_ris_2)
dotplot(ranef(model_ris_2,condVar=TRUE))

# third random slop model : ris_3
model_ris_3 <- lmer( log_PGV ~ 1 + log_Pv + (1 + log_Pv|StationName) + (1 +
log_Pv| EarthquakeID) + (1 | Year_Index), data = train_data, REML = FALSE,
control = control)
summary(model_ris_3)
anova(model_ris_3)
performance::r2(model_ris_3)
dotplot(ranef(model_ris_3,condVar=TRUE))

#model selection
anova(model_null,model_ri,model_ris_1,model_ris_2,model_ris_3)

# Make predictions on the test set
predicted_values_model_ols   <- predict(model_ols, newdata = test_data)
predicted_values_model_null  <- predict(model_null, newdata = test_data,
allow.new.levels = TRUE)
predicted_values_model_ri    <- predict(model_ri, newdata = test_data,
allow.new.levels = TRUE)
predicted_values_model_ris_1 <- predict(model_ris_1, newdata = test_data,
allow.new.levels = TRUE)
predicted_values_model_ris_2 <- predict(model_ris_2, newdata = test_data,
allow.new.levels = TRUE)
predicted_values_model_ris_3 <- predict(model_ris_3, newdata = test_data,
allow.new.levels = TRUE)

# Calculate RMSE
rmse__model_ols   <- sqrt(mean(((test_data$log_PGV -
predicted_values_model_ols)^2)))
rmse__model_null  <- sqrt(mean(((test_data$log_PGV -
predicted_values_model_null)^2)))
rmse__model_ri    <- sqrt(mean(((test_data$log_PGV -
predicted_values_model_ri)^2)))
rmse__model_ris_1 <- sqrt(mean(((test_data$log_PGV -
predicted_values_model_ris_1)^2)))
rmse__model_ris_2 <- sqrt(mean(((test_data$log_PGV -
predicted_values_model_ris_2)^2)))
rmse__model_ris_3 <- sqrt(mean(((test_data$log_PGV -
predicted_values_model_ris_3)^2)))

print(paste("RMSE:", rmse__model_ols))
print(paste("RMSE:", rmse__model_null))
print(paste("RMSE:", rmse__model_ri))
print(paste("RMSE:", rmse__model_ris_1))
print(paste("RMSE:", rmse__model_ris_2))
print(paste("RMSE:", rmse__model_ris_3))

# Calculate MAPE
```

```
mape__model_ols    <- mean(abs((test_data$log_PGV -
predicted_values_model_ols) / (test_data$log_PGV))) * 100
mape__model_null  <- mean(abs((test_data$log_PGV -
predicted_values_model_null) / (test_data$log_PGV))) * 100
mape__model_ri     <- mean(abs((test_data$log_PGV - predicted_values_model_ri)
/ (test_data$log_PGV))) * 100
mape__model_ris_1 <- mean(abs((test_data$log_PGV -
predicted_values_model_ris_1) / (test_data$log_PGV))) * 100
mape__model_ris_2 <- mean(abs((test_data$log_PGV -
predicted_values_model_ris_2) / (test_data$log_PGV))) * 100
mape__model_ris_3 <- mean(abs((test_data$log_PGV -
predicted_values_model_ris_3) / (test_data$log_PGV))) * 100

print(paste("MAPE:", mape__model_ols, "%"))
print(paste("MAPE:", mape__model_null, "%"))
print(paste("MAPE:", mape__model_ri, "%"))
print(paste("MAPE:", mape__model_ris_1, "%"))
print(paste("MAPE:", mape__model_ris_2, "%"))
print(paste("MAPE:", mape__model_ris_3, "%"))

## Testings ##
simple_model <- lmer(log_PGV ~ 1 + log_Pv + (1 + log_Pv|StationName), data =
train_data, REML = FALSE, control = control)
visualize(simple_model, plot ="model", sample=288, pch = 8)
p + theme(legend.text = element_text(size =2))
visualize(simple_model, plot ="model",formula =log_PGV~log_Pv | StationName,
sample=5, legend(pch = 8) )
visualize(simple_model, plot ="model",formula =log_PGV~log_Pv + StationName,
sample=5,legend(pch = 2) )

simple_model2 <- lmer(log_PGV ~ 1 + log_Pv + (1 + log_Pv|StationName) + (1 +
log_Pv| EarthquakeID) + (1| Year_Index) , data = train_data, REML = FALSE,
control = control)
visualize(simple_model2, plot ="model", sample=20 )
visualize(simple_model2,formula =log_PGV~log_Pv |Year_Index ,  sample=10 )
visualize(simple_model2,formula =log_PGV~log_Pv + StationName,  sample=15)

# Subset stations
subsch = sort(unique(train_data$StationName))[95:109]
# Get the data for these stations and name it as "dat"
dat = train_data[train_data$StationName %in%
subsch,c("StationName","log_Pv","log_PGV")]
dat$StationName = as.factor(dat$StationName)
xyplot(log_PGV~log_Pv|as.factor(Year_Index),
       group=StationName, type=c("p","r"), test_data)

# Call xyplot and add regression lines
xyplot(log_PGV~log_Pv|StationName,type=c("p","r"), lwd=3, dat)

par(mar = c(4, 4, .1, .1))
xyplot(log_PGV~log_Pv, group=StationName, type=c("p","r"),
       auto.key = list(columns = nlevels(dat$StationName)), dat)
```

```
par(mar = c(4, 4, .1, .1))
xyplot(log_PGV~log_Pv, group=StationName, type=c("p","r"), train_data)

# Sort the data for better plotting
dSci= eqdata[order(train_data$StationName, train_data$log_Pv),]
# Make the plot
xyplot(log_PGV~log_Pv, group=StationName, type=c("p","r"), dSci)

predicted_values_model_ols   <- predict(model_ols, newdata = test_data)
predicted_values_model_ris_1 <- predict(model_ris_1, newdata = test_data,
allow.new.levels = TRUE)
#Step 5: Define the number of stations to analyze (modify as needed)
num_stations <- 15
par(mfrow = c(3, 5))

# Step 6: Loop through stations and create plots
for (i in 1:num_stations) {
  # Extract station ID for the current iteration (assuming order reflects
stations)
  current_station_id <- eqdata$StationName[i]

  # Filter data for the current station
  station_subset <- subset(eqdata, StationName == current_station_id)
  actual_intensity <- test_data$log_PGV  # Assuming "s_wave_intensity" is the
outcome variable

  # Extract predictions from fitted models (assuming appropriate variable
names)
  simple_predictions_subset <- predict(model_ols, newdata = test_data)
  multilevel_predictions_subset <- predict(model_ris_1, newdata = test_data,
allow.new.levels = TRUE)

  # Create the scatter plot
  plot(actual_intensity, simple_predictions_subset,
       pch = 16, col = "blue",
       xlab = "Actual S-wave Intensity",
       ylab = "Predicted S-wave Intensity (Simple Model)",
       main = paste("Station", current_station_id))

  # Add points for the multilevel model predictions
  points(actual_intensity, multilevel_predictions_subset,
         pch = 16, col = "red")

  # Add legend
  legend("topright", legend = c("Simple Model", "Multilevel Model"),
         pch = c(16, 16), col = c("blue", "red"))

  # Add regression line (optional)
  abline(lm(actual_intensity ~ simple_predictions_subset))
  abline(lm(actual_intensity ~ multilevel_predictions_subset), col = "red")
}
## End Testings ##
```

```
# Make predictions using fixed effects only and then using both fixed and
random effects
train_data <- train_data %>%
  mutate(pred_fixef = predict(model_ris_1, newdata = ., re.form = NA),
         pred_ranef = predict(model_ris_1, newdata = ., re.form = ~(1 +
log_Pv | StationName)),
         abs_diff   = abs(pred_fixef - pred_ranef))

train_data <- train_data %>%
  mutate(pred_ols = predict(model_ols, newdata = .))

# Summarize the average absolute differences by station
station_diffs <- train_data %>%
  group_by(StationName) %>%
  summarize(avg_abs_diff = mean(abs_diff, na.rm = TRUE)) %>%
  arrange(desc(avg_abs_diff))

# Select the top 6 stations with the largest average absolute differences
top_stations <- station_diffs %>%
  top_n(6, avg_abs_diff) %>%
  pull(StationName)

# Filter the dataset to include only the selected stations
filtered_data <- train_data %>%
  filter(StationName %in% top_stations)

# Make predictions using fixed effect only and then random effects and plot
the results
filtered_data %>%
  mutate(pred_fixef = predict(model_ris_1, newdata = ., re.form = NA),
         pred_ranef = predict(model_ris_1, newdata = ., re.form =
~(1+log_Pv|StationName)),
         pred_ols = predict(model_ols, newdata = .)) %>%
  ggplot(aes(x = log_Pv, y = log_PGV)) +
  geom_point(shape = 21,
             size = 0.6,
             color = "black",
             fill = "grey",
             show.legend = TRUE) +
  geom_line(aes(y = pred_fixef, color = "Fixed Effects Prediction"),
            size = 1) +
  geom_line(aes(y = pred_ranef, color = "Random Effects Prediction"),
            size = 1) +
  geom_line(aes(y = pred_ols, color = "OLS Prediction"),
            size = 0.8) +
  facet_wrap(~StationName) +
  labs(x = "log_Pv",
       y = "log_PGV") +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 12),
        legend.title = element_blank()) +
  scale_color_manual(values = c("Fixed Effects Prediction" = "blue",
                                "Random Effects Prediction" = "green",
                                "OLS Prediction" = "red"),
                     breaks = c( "Fixed Effects Prediction", "Random Effects
Prediction", "OLS Prediction")
  )

# Function to calculate RMSE on test set
calc_rmse <- function(model, data) {
  predictions <- predict(model, newdata = data, allow.new.levels = TRUE)
  sqrt(mean((data$log_PGV - predictions)^2))
}

# Function to calculate MSE on test set
calc_mse <- function(model, data) {
  predictions <- predict(model, newdata = data,allow.new.levels = TRUE)
  mean((data$log_PGV - predictions)^2)
}

# Function to calculate MAPE on test set
calc_mape <- function(model, data) {
  predictions <- predict(model, newdata = data, allow.new.levels = TRUE)
  mean(abs((data$log_PGV - predictions) / data$log_PGV)) * 100
}

# Function to calculate R² for linear mixed models
calc_r2 <- function(model) {
  performance::r2(model)
}

# Function to check model convergence
check_convergence <- function(model) {
  if (is.null(model@optinfo$conv$opt)) {
    return("converged")
  } else {
    return("not converged")
  }
}

## Checking model assumptions
# Extract residuals and fitted values
residuals <- residuals(model_ris_1)
fitted <- fitted(model_ris_1)

# 1. Linearity
# Plot observed vs fitted values
ggplot(data.frame(fitted, log_PGV = train_data$log_PGV), aes(x = fitted, y =
log_PGV)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, col = 'red') +
  labs(x = "Fitted Values", y = "Observed Values")

# Residuals vs fitted values plot
ggplot(data.frame(fitted, residuals), aes(x = fitted, y = residuals)) +
  geom_point() +
```

```r
  geom_hline(yintercept = 0, col = 'red') +
  labs(x = "Fitted Values", y = "Residuals")

# 2. Normality of residuals
# Q-Q plot of the residuals
qqnorm(residuals)
qqline(residuals, col = 'red')

# Histogram of the residuals
ggplot(data.frame(residuals), aes(x = residuals)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.2, fill = 'darkgray',
alpha = 0.5) +
  geom_density(col = 'red') +
  labs(x = "Residuals", y = "Density")

plot(model_ris_1, refit = TRUE)  # Residuals vs fitted values
plot(model_ris_1, cookwd = TRUE)

# 4. Independence of residuals
# Plot residuals over time or order of observations (if time/order data
available)
# Here, assuming the order of observations is in the row numbers
ggplot(data.frame(order = 1:length(residuals), residuals), aes(x = order, y =
residuals)) +
  geom_point() +
  geom_line() +
  geom_hline(yintercept = 0, col = 'red') +
  labs(title = "Residuals Over Order of Observations", x = "Order", y =
"Residuals")

plot(model_ris_1, residuals = TRUE)

# Extract random effects and convert to data frame
random_effects <- ranef(model_ris_1)
station_random_effects <- as.data.frame(random_effects$StationName) %>%
mutate(grp = rownames(.))
earthquake_random_effects <- as.data.frame(random_effects$EarthquakeID) %>%
mutate(grp = rownames(.))
year_random_effects <- as.data.frame(random_effects$Year_Index) %>%
mutate(grp = rownames(.))

# Plot random effects for StationName using ggplot2
ggplot(station_random_effects, aes(x = grp, y = `(Intercept)`)) +
  geom_point() +
  geom_errorbar(aes(ymin = `(Intercept)` - 1.96 * `(Intercept)`, ymax =
`(Intercept)` + 1.96 * `(Intercept)`), width = 0.1) +
  labs(title = "Random Effects for StationName", x = "Station Name", y =
"Random Effect") +
  theme(axis.text.x = element_text(angle = 90))

qqnorm(station_random_effects$`(Intercept)`)
qqline(station_random_effects$`(Intercept)`, col = "red")
```

```
# Plot random effects for EarthquakeID using ggplot2
ggplot(earthquake_random_effects, aes(x = grp, y = `(Intercept)`)) +
  geom_point() +
  geom_errorbar(aes(ymin = `(Intercept)` - 1.96 * `(Intercept)`, ymax =
`(Intercept)` + 1.96 * `(Intercept)`), width = 0.1) +
  labs(title = "Random Effects for EarthquakeID", x = "Earthquake ID", y =
"Random Effect") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Plot random effects for Year_Index using ggplot2
ggplot(year_random_effects, aes(x = grp, y = `(Intercept)`)) +
  geom_point() +
  geom_errorbar(aes(ymin = `(Intercept)` - 1.96 * `(Intercept)`, ymax =
`(Intercept)` + 1.96 * `(Intercept)`), width = 0.1) +
  labs(title = "Random Effects for Year_Index", x = "Year Index", y = "Random
Effect") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

check_model(model_ris_1)
diagnostic_plots <- plot(check_model(model_ris_1, panel = FALSE))
diagnostic_plots[[1]]
diagnostic_plots[[2]]
diagnostic_plots[[3]]
diagnostic_plots[[4]]
diagnostic_plots[[5]]
diagnostic_plots[[6]]
diagnostic_plots[[7]]
diagnostic_plots[[8]]


## Preliminary Analysis plot
# Time-Based Plots
ggplot(eqdata, aes(x = Year_Index, y = log_PGV)) +
  geom_boxplot() +
  labs(title = "Time Series Plot of log_PGV", x = "Year Index", y =
"log_PGV")

ggplot(eqdata, aes(x = Year_Index, y = log_Pv)) +
  geom_boxplot() +
  labs(title = "Time Series Plot of log_Pv", x = "Year Index", y = "log_Pv")


# Sample 25 stations
sampled_stations <- eqdata %>%
  distinct(StationName) %>%
  sample_n(50)

# Filter the data to include only the sampled stations
sampled_data <- eqdata %>%
  filter(StationName %in% sampled_stations$StationName)

# Group-Based Plots
ggplot(sampled_data, aes(x = StationName, y = log_PGV)) +
  geom_boxplot() +
  labs(title = "Boxplot of log_PGV by StationName", x = "StationName", y =
"log_PGV") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Adjust for
readability


# Outlier investigation in preliminary analysis

boxplot(eqdata_raw$Pa)
boxplot(eqdata_raw$Pv)
boxplot(eqdata_raw$Pd)
boxplot(eqdata_raw$PGV)

Q1 <- quantile(eqdata_raw$Pa, 0.25)
Q3 <- quantile(eqdata_raw$Pa, 0.75)
IQR <- Q3 - Q1

lower_whisker <- Q1 - 1.5 * IQR
upper_whisker <- Q3 + 1.5 * IQR

outliers <- which(eqdata_raw$Pa > upper_whisker | eqdata_raw$Pa <
lower_whisker)

outlier_data <- eqdata_raw[outliers, ]
cat("Number of outliers:", length(outliers), "\n")
print(outlier_data)

write.xlsx(outlier_data, file =
"C:/Users/thanu/Desktop/IP_WORK/OutlierData_Pa.csv"
, rowNames = FALSE)

# Removing outliers from data considering Pa
cleaned_data <- eqdata_raw[eqdata_raw$Pa != 9.043736232, ]
boxplot(cleaned_data$Pa)

# Removing outliers considering PGV
Q_1 <- quantile(eqdata_raw$PGV, 0.25)
Q_3 <- quantile(eqdata_raw$PGV, 0.75)
IQR1 <- Q_3 - Q_1

lower_whisker1 <- Q_1 - 1.5 * IQR1
upper_whisker1 <- Q_3 + 1.5 * IQR1

outliers1 <- which(eqdata_raw$PGV > upper_whisker1 | eqdata_raw$PGV <
lower_whisker1)

outlier_data1 <- eqdata_raw[outliers1, ]
cat("Number of outliers:", length(outliers1), "\n")
print(outlier_data1)

write.xlsx(outlier_data1, file =
"C:/Users/thanu/Desktop/IP_WORK/OutlierData_PGV.csv"
```

```
           , rowNames = FALSE)


# Analysis on selecting predictor variable

eqdata_raw <- read.csv("C:/Users/thanu/Desktop/IP_WORK/Data_Final.csv",
header = TRUE)
head(eqdata_raw)
summary(eqdata_raw)

cleaned_data <- eqdata_raw[eqdata_raw$Pa != 9.043736232, ]
summary(cleaned_data)

model1x2 <- lm(PGV ~ Pv, data = cleaned_data)
summary(model1x2)

par(mfrow=c(1,1))
with(cleaned_data, {
  plot(Pv, PGV)
  abline(model1x2)
})

# Convert the Character variables to factors
eqdata_raw$StationName <- factor(eqdata_raw$StationName)
eqdata_raw$SiteClass <- factor(eqdata_raw$SiteClass)
levels(eqdata_raw$StationName)

#Removing outliers
eqdata <- eqdata_raw[eqdata_raw$Pa != 9.043736232, ]
summary(eqdata)

#Dividing data into Train set & Test set
set.seed(123)
trainIndex <- createDataPartition(eqdata$PGV, p = .8, list = FALSE, times =
1)
data_train <- eqdata[trainIndex,]
data_test  <- eqdata[-trainIndex,]


###Choose best simple linear regression model
modelAx1 <- lm(log(PGV)~ log(Pa), data = data_train)
modelAx2 <- lm(log(PGV) ~ log(Pv), data = data_train)
modelAx3 <- lm(log(PGV) ~ log(Pd), data = data_train)
modelBx1 <- lm(log(PGA) ~ log(Pa), data = data_train)
modelBx2 <- lm(log(PGA) ~ log(Pv), data = data_train)
modelBx3 <- lm(log(PGA) ~ log(Pd), data = data_train)

summary(modelAx1)
summary(modelAx2)
summary(modelAx3)
summary(modelBx1)
summary(modelBx2)
summary(modelBx3)
```

```
AIC(modelAx1,modelAx2,modelAx3,modelBx1,modelBx2,modelBx3)
BIC(modelAx1,modelAx2,modelAx3,modelBx1,modelBx2,modelBx3)
# Make predictions on the test set
predicted_values_modelAx1 <- predict(modelAx1, newdata = data_test)
predicted_values_modelAx2 <- predict(modelAx2, newdata = data_test)
predicted_values_modelAx3 <- predict(modelAx3, newdata = data_test)
predicted_values_modelBx1 <- predict(modelBx1, newdata = data_test)
predicted_values_modelBx2 <- predict(modelBx2, newdata = data_test)
predicted_values_modelBx3 <- predict(modelBx3, newdata = data_test)

# Calculate RMSE
rmse__modelAx1 <- sqrt(mean(((log(data_test$PGV)) -
predicted_values_modelAx1)^2))
rmse__modelAx2 <- sqrt(mean((log(data_test$PGV) -
predicted_values_modelAx2)^2))
rmse__modelAx3 <- sqrt(mean((log(data_test$PGV) -
predicted_values_modelAx3)^2))
rmse__modelBx1 <- sqrt(mean(((log(data_test$PGA)) -
predicted_values_modelBx1)^2))
rmse__modelBx2 <- sqrt(mean((log(data_test$PGA) -
predicted_values_modelBx2)^2))
rmse__modelBx3 <- sqrt(mean((log(data_test$PGA) -
predicted_values_modelBx3)^2))

print(paste("RMSE:", rmse__modelAx1))
print(paste("RMSE:", rmse__modelAx2))
print(paste("RMSE:", rmse__modelAx3))
print(paste("RMSE:", rmse__modelBx1))
print(paste("RMSE:", rmse__modelBx2))
print(paste("RMSE:", rmse__modelBx3))

# Calculate MAPE
mape__modelAx1 <- mean(abs((log(data_test$PGV) - predicted_values_modelAx1) /
(log(data_test$PGV)))) * 100
mape__modelAx2 <- mean(abs((log(data_test$PGV) - predicted_values_modelAx2) /
(log(data_test$PGV)))) * 100
mape__modelAx3 <- mean(abs((log(data_test$PGV) - predicted_values_modelAx3) /
(log(data_test$PGV)))) * 100

print(paste("MAPE:", mape__modelAx1, "%"))
print(paste("MAPE:", mape__modelAx2, "%"))
print(paste("MAPE:", mape__modelAx3, "%"))

# Choose best regression model using Two variables
modelBx1x2 <- lm(log(PGV) ~ log(Pa) + log(Pv), data = data_train)
modelBx1x3 <- lm(log(PGV) ~ log(Pa) + log(Pd), data = data_train)
modelBx2x3 <- lm(log(PGV) ~ log(Pv) + log(Pd), data = data_train)

summary(modelBx1x2)
summary(modelBx1x3)
summary(modelBx2x3)
```

```
AIC(modelBx1x2, modelBx1x3, modelBx2x3)
BIC(modelBx1x2, modelBx1x3, modelBx2x3)

# Make predictions on the test set
predicted_values_modelBx1x2 <- predict(modelBx1x2, newdata = data_test)
predicted_values_modelBx1x3 <- predict(modelBx1x3, newdata = data_test)
predicted_values_modelBx2x3 <- predict(modelBx2x3, newdata = data_test)

# Calculate RMSE
rmse__modelBx1x2 <- sqrt(mean((log(data_test$PGV) -
predicted_values_modelBx1x2)^2))
rmse__modelBx1x3 <- sqrt(mean((log(data_test$PGV) -
predicted_values_modelBx1x3)^2))
rmse__modelBx2x3 <- sqrt(mean((log(data_test$PGV) -
predicted_values_modelBx2x3)^2))


print(paste("RMSE:", rmse__modelBx1x2))
print(paste("RMSE:", rmse__modelBx1x3))
print(paste("RMSE:", rmse__modelBx2x3))

# Calculate MAPE
mape__modelBx1x2 <- mean(abs((log(data_test$PGV) -
predicted_values_modelBx1x2) / (log(data_test$PGV)))) * 100
mape__modelBx1x3 <- mean(abs((log(data_test$PGV) -
predicted_values_modelBx1x3) / (log(data_test$PGV)))) * 100
mape__modelBx2x3 <- mean(abs((log(data_test$PGV) -
predicted_values_modelBx2x3) / (log(data_test$PGV)))) * 100

print(paste("MAPE:", mape__modelBx1x2, "%"))
print(paste("MAPE:", mape__modelBx1x3, "%"))
print(paste("MAPE:", mape__modelBx2x3, "%"))

# Choose best regression model using all three variables
modelCx1x2x3 <- lm(log(PGV) ~ log(Pa) + log(Pv) + log(Pd), data = data_train)
summary(modelCx1x2x3)
AIC(modelCx1x2x3)
BIC(modelCx1x2x3)
predicted_values_modelCx1x2x3 <- predict(modelCx1x2x3, newdata = data_test)
rmse__modelCx1x2x3 <- sqrt(mean((log(data_test$PGV) -
predicted_values_modelCx1x2x3)^2))
print(paste("RMSE:", rmse__modelCx1x2x3))
mape__modelCx1x2x3 <- mean(abs((log(data_test$PGV) -
predicted_values_modelCx1x2x3) / (log(data_test$PGV)))) * 100
print(paste("MAPE:", mape__modelCx1x2x3, "%"))


# Ridge Regression
x = as.matrix(log(data_train[, c("Pa", "Pv", "Pd")]))
y_train = log(data_train$PGV)
y_train <- matrix(y_train, ncol = 1)

x_test = as.matrix(log(data_test[, c("Pa", "Pv", "Pd")]))
```

```
y_test = log(data_test$PGV)
y_test <- matrix(y_test, ncol = 1)

lambdas <- 10^seq(2, -3, by = -.1)
ridge_reg = glmnet(x, y_train, nlambda = 25, alpha = 0, family = 'gaussian',
lambda = lambdas)

summary(ridge_reg)

cv_ridge <- cv.glmnet(x, y_train, alpha = 0, lambda = lambdas)
optimal_lambda <- cv_ridge$lambda.min
optimal_lambda

# Compute R^2 from true and predicted values
eval_results <- function(true, predicted, df) {
  SSE <- sum((predicted - true)^2)
  SST <- sum((true - mean(true))^2)
  R_square <- 1 - SSE / SST
  RMSE = sqrt(SSE/nrow(df))

  # Model performance metrics
  data.frame(
    RMSE = RMSE,
    Rsquare = R_square
  )
}

# Prediction and evaluation on train data
predictions_train <- predict(ridge_reg, s = optimal_lambda, newx = x)
eval_results(y_train, predictions_train, data_train)

# Prediction and evaluation on test data
predictions_test <- predict(ridge_reg, s = optimal_lambda, newx = x_test)
eval_results(y_test, predictions_test, data_test)

# Lasso Regression
lambdas <- 10^seq(2, -3, by = -.1)

# Setting alpha = 1 implements lasso regression
lasso_reg <- cv.glmnet(x, y_train, alpha = 1, lambda = lambdas, standardize =
TRUE, nfolds = 5)
plot(lasso_reg)

# Best
lambda_best <- lasso_reg$lambda.min
lambda_best

lasso_model <- glmnet(x, y_train, alpha = 1, lambda = lambda_best,
standardize = TRUE)

predictions_train <- predict(lasso_model, s = lambda_best, newx = x)
eval_results(y_train, predictions_train, data_train)
```

```
predictions_test <- predict(lasso_model, s = lambda_best, newx = x_test)
eval_results(y_test, predictions_test, data_test)

# multiple linear regression- Pv and site class as predivctor variables
modelDx2<- lm(log(PGV) ~ log(Pv) + SiteClass, data = data_train)
summary(modelDx2)
AIC(modelDx2)
BIC(modelDx2)
predicted_values_modelDx2 <- predict(modelDx2, newdata = data_test)
rmse__modelDx2 <- sqrt(mean((log(data_test$PGV) -
predicted_values_modelDx2)^2))
print(paste("RMSE:", rmse__modelDx2))
mape__modelDx2 <- mean(abs((log(data_test$PGV) - predicted_values_modelDx2) /
(log(data_test$PGV)))) * 100
print(paste("MAPE:", mape__modelDx2, "%"))
```

9.2 Appendix 2: Data retrieving process from GeoNet FDSN server (python)

```
import obspy
from obspy import UTCDateTime
from obspy.clients.fdsn import Client
import matplotlib
import os
import csv
from obspy import read_events

# Create FDSN client
c = Client("GEONET")
start_time = UTCDateTime(2013, 1, 1, 0, 0, 0)
end_time = UTCDateTime(2024, 3, 28, 23, 59, 59)
min_magnitude = 3.0
max_magnitude = 3.1
min_lat = -47.5617
max_lat = -34.2192
min_lon = 165.8271
max_lon = 179.6050


catelog_dir = "D:\WORK\GeoNet Data\catalogue_2013_2024"

# Download catelogs
while max_magnitude <= 8:
    try:
        catalog = c.get_events(starttime=start_time, endtime=end_time,
minmagnitude=min_magnitude, maxmagnitude=max_magnitude, minlatitude=min_lat,
maxlatitude=max_lat, minlongitude=min_lon, maxlongitude=max_lon)

##write the catalog to a file called catalogue
        catalog.write("D:\WORK\GeoNet Data\catalogue_2013_2024\\" +
"catalogue_over_S_"+str(min_magnitude), format="QUAKEML")
        print(catalog)
        min_magnitude = round((min_magnitude + 0.1),1)
        max_magnitude = round((max_magnitude + 0.1),1)
    except:
        min_magnitude = round((min_magnitude + 0.1),1)
        max_magnitude = round((max_magnitude + 0.1),1)


# Download full waveforms
    for line in csv_reader:
        try:
            earthquakeID = line[0].split('/')[-1] #or earthquakeID = line[0]
if there is no /
            start_time = line[2]
            station_code = line[1]
            starttime = UTCDateTime(start_time) -30
            endtime = UTCDateTime(start_time) + 60
            for output_i in output:
                st = c.get_waveforms(network=network, station=station_code,
                                     location=location, channel=channel,
                                     starttime=starttime, endtime=endtime,
attach_response=True)

    ## Peak value calculation of each waveform
    #P_Wave
    PGA_P = [max(np.abs(channel1_p[0].data)),
max(np.abs(channel2_p[0].data)), max(np.abs(channel3_p[0].data))]
    PGV_P = [max(np.abs(channel1_v[0].data)),
max(np.abs(channel2_v[0].data)), max(np.abs(channel3_v[0].data))]
    Pd_P = [max(np.abs(channel1_d[0].data)), max(np.abs(channel2_d[0].data)),
max(np.abs(channel3_d[0].data))]

    # S_Wave
    PGA_S = max_root_square_sum(Schannel2_p[0].data,Schannel3_p[0].data)
    PGV_S = max_root_square_sum(Schannel2_v[0].data,Schannel3_v[0].data)
    Pd_S = max_root_square_sum(Schannel2_d[0].data, Schannel3_d[0].data)
```