ALTERNATIVE METHOD TO DEVELOP NEW STRATEGY IN ORDINAL REGRESSION: A CASE STUDY IN DENTAL SCIENCES

MUHAMAMD AMIRUL BIN MAT LAZIN

UNIVERSITI SAINS MALAYSIA

2025

ALTERNATIVE METHOD TO DEVELOP NEW STRATEGY IN ORDINAL REGRESSION: A CASE STUDY IN DENTAL SCIENCES

by

MUHAMAMD AMIRUL BIN MAT LAZIN

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

March 2025

ACKNOWLEDGEMENT

Assalamualaikum warahmatullahi wabarakatuh

Alhamdulillah thanks to the presence of Allah SWT because with His grace, I was able to complete this research project successfully and according to the set time period.

First of all, I would like to express my deepest appreciation to Ts. Dr. Mohamad Arif Bin Awang Nawi as the main supervisor for helping me a lot to complete this research project well and successfully. In addition, he gave a lot of guidance and help in correcting the mistakes I made in this research project. I would like to thank the dean of School of Dental Sciences, Health Campus, Universiti Sains Malaysia, and all the stuff in the college for their support.

I would like to express my deepest appreciation to Universiti Sains Malaysia (USM) for their generous support in funding my research through Short Term Grant No. 304/PPSG/6315410. The financial and institutional support from USM has enabled me to continue my research efforts more deeply and rigorously. I am deeply grateful for the confidence and contribution to the advancement of scientific knowledge.

In addition to that, I would also like to thank my parents who helped me a lot by giving me encouragement and enthusiasm to complete this research project. Not forgetting the friends who also gave me views, ideas and help directly or indirectly to complete this research project. I really appreciate your good service and thank you.

Finally, the highest appreciation is also directed to those who are directly or indirectly involved in helping me to make this project a success.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

R	Programming language used for statistical computing and graphics
SPSS	Statistical Package for the Social Sciences
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion,
USM	Universiti Sains Malaysia
EDF	Empirical Density Function
ML	Maximum Likelihood
SE	Standard Error
Exp(*B*)	Exponential function of B (Odds Ratio)
df	Degrees of Freedom
Sig.	Significant
OR	Odds Ration
LL	Log-Likelihood
GLM	Generalized Linear Model
CI	Confidence interval
LR	Logistic Regression

LIST OF APPENDICES

KAEDAH ALTERNATIF UNTUK MEMBANGUNKAN STRATEGI BARU DALAM REGRESI ORDINAL: KAJIAN KES DALAM SAINS PERGIGIAN

ABSTRAK

Data klinikal biasanya mengandungi banyak ciri dengan saiz sampel yang kecil, menghasilkan dimensi yang lebih tinggi dan ketepatan yang lemah. Ini mengurangkan prestasi sistem pengelasan dalam set data berdimensi tinggi kerana ciri-ciri yang tidak berkaitan menyumbang kepada ketepatan pengelasan yang lemah dan menambah kesukaran tambahan dalam mencari pengetahuan yang berpotensi berguna. Objektif utama ialah untuk membangunkan model alternatif untuk regresi ordinal melalui pembinaan metodologi statistik. Kaedah ini termasuk reka bentuk kajian komputasi dan teknik statistik yang disesuaikan untuk pemodelan sains gigi. Gabungan regresi ordinal dan teknik bootstrap dalam membangunkan model alternatif adalah kunci utama kepada titik fokus penyelidikan. Dua kajian kes, keterukan haus gigi dan sensitivity gigi, digunakan untuk menguji teknik ini, menunjukkan kaitannya dengan data pergigian dunia sebenar. Semua pengaturcaraan asas dilakukan dengan menggunakan perisian R. Hasil menunjukkan bahawa pendekatan alternatif, terutamanya dengan lebih banyak replikasi bootstrap, menawarkan penyesuaian model yang lebih baik dan ketepatan berbanding dengan regresi ordinal tradisional. Ini menunjukkan kegunaannya dalam meningkatkan ketepatan penyelidikan sains kesihatan, terutamanya dalam situasi dengan saiz sampel kecil. Kajian ini memperkukuh kaedah statistik dalam sains pergigian dengan memperkenalkan alternatif yang lebih robust kepada regresi ordinal, membolehkan penyelidik memperoleh hasil yang lebih tepat dan boleh dipercayai walaupun dengan set data yang terhad.

ALTERNATIVE METHOD TO DEVELOP NEW STRATEGY IN ORDINAL REGRESSION: A CASE STUDY IN DENTAL SCIENCES

ABSTRACT

Clinical data usually contain numerous features with a small sample size, resulting in higher dimensionality and poor accuracy. This reduces the performance of classifier systems in high-dimensional data sets because irrelevant features contribute to poor classification accuracy and add extra difficulties in finding potentially useful knowledge. The main objective is to develop an alternative model for ordinal regression through statistical methodology building. The methodology includes a computational study design and statistical techniques customised for dental science modelling. A combination of ordinal regression and bootstrap techniques in the developing an alternative model is the main key to the research focal point. Two case studies, tooth wear severity and tooth sensitivity, were used to test this technique, demonstrating its relevance to real-world dental data. All the fundamental programming was performed using R software. The results show that the alternative approach, especially with more bootstrap replications, offers improved model fitting and precision compared to traditional ordinal regression. This suggests its usefulness in improving the accuracy of health science research, especially in situations with small sample sizes. This study strengthens statistical methods in dental sciences by introducing a more robust alternative to ordinal regression, enabling researchers to obtain more accurate and reliable results even with limited datasets.

CHAPTER 1

INTRODUCTION

1.1 Background of The Study

Ordinal regression models are statistical methods that can be used to analyse ordered health-related quality of life measures. This statistical technique also used to predict behaviour of ordinal level dependent variables with a set of independent variables. The dependent variable is the order response category variable and the independent variable may be categorical or continuous. There has been increasing emphasis in medical research on the design and analysis of quality of life scales. Many quality of life scales are ordinal and statistical methods such as ordinal regression models have been reviewed on a number of occasions. This method is the most widely used in epidemiological and biomedical applications but ordinal regression leads to strong assumptions that may lead to incorrect interpretations if the assumptions are violated (Shivalingappa & Parameshwar, 2010).

In a large number of biomedical and health survey applications, the response to be compared or predicted is ordinal. Examples of ordinal outcomes include Tumor-Node-Metastasis (TNM) stage (I, II, III, IV); drug toxicity evaluated as ‘none,’ ‘mild,’ ‘moderate,’ or ‘severe;’ and response to treatment classified as complete response, partial response, stable disease, or progressive disease. These outcomes are ordinal; that is, while there is an inherent ordering present among the responses, there is no known underlying numerical relationship between the responses. Furthermore, for ordinal response data, parameters other than classifier accuracy may be the more important measures of classifier performance (Arche & Mas, 2009).

To enhance the background of the study on ordinal regression models, it is necessary to investigate different aspects such as their theoretical foundations, practical uses, methodological difficulties, and the impact of these models in the wider fields of medical research and public health. To have a thorough knowledge of the relevance and importance of ordinal regression models in health sciences, it is necessary to consider a diverse variety of perspectives in this exploration.

Ordinal regression models utilise a robust theoretical framework to analyse ordinal data by estimating the association between an ordinal dependent variable and one or more independent variables. This framework is essential for managing data in situations when the answer variable is categorised in a specific sequence, which is a frequent occurrence in health-related research. For example, the intensity of symptoms, progression of the disease, and degree of patient contentment are usually assessed using an ordinal scale. Ordinal regression's ability to use the ordering of categories, without requiring equal space between them, makes it highly valuable for analysing and understanding such data (McCullagh, 1980).

Ordinal regression has numerous and diverse practical applications in medical research. Ordinarily, in clinical trials, it is frequently used to examine ordinal outcomes such as the intensity of negative effects or phases of disease advancement. By employing this approach, researchers can extract subtler and detailed observations from their data, which could result in the development of more efficient treatment approaches and improved quality of care for patients (Agresti, 2010).

Although ordinal regression models are commonly used, they present unique challenges and considerations. An important concern is the assumption of proportional odds, which suggests that the relationship between any pair of outcome groups is same. Deviation from this assumption can result in model discrepancy and imprecise

deductions. In order to maintain the accuracy of their studies, researchers must thoroughly examine and resolve any instances of noncompliance.

The use of ordinal regression models has wide-ranging ramifications that go beyond individual studies and have a significant impact on the field of health research as a whole. These models enhance the ability to analyse ordinal data, which in turn helps in developing more efficient therapies, establishing more precise diagnostic criteria, and improving patient outcomes. Additionally, they enhance comprehension of patient experiences and perceptions, which are widely acknowledged as vital components of healthcare quality and efficacy (Fitzpatrick, et al., 1992).

In order to provide more detailed information about the study's background, it is important to emphasise the multidisciplinary character of ordinal regression models. These models are not only used in health sciences, but also have applications in psychology, education, and social sciences. The versatility and strength of the models are highlighted by their capacity to be applied across many disciplines, allowing researchers from other professions to address intricate concerns that involve ordered categories (Agresti, 2010).

The accessibility and usability of ordinal regression models have been greatly improved by recent developments in processing power and statistical software. R and SAS now include advanced features for doing ordinal regression analysis, making it easier for researchers to get started and increasing the possibilities for creative applications in medical research (Christensen, 2015).

Furthermore, the continuous progress in the creation of more sophisticated ordinal regression approaches, such as those that loosen the assumption of proportional odds or add random effects to handle clustered data, has great potential as a field of methodological research. These advancements have the capacity to enhance the

analysis of ordinal data, creating new opportunities for investigation and revelation in health research (Peterson & Harrell, 1990).

The ethical problems that pertain to the study of health-related data are equally relevant to the utilisation of ordinal regression models. Researchers must address concerns regarding data privacy, informed permission, and the possibility of misinterpreting results, ensuring that their study follows the most rigorous ethical guidelines (Mills, 2014).

To summarise, the extended context of this work highlights the crucial significance of ordinal regression models in examining health-related quality of life measures and other ordinal outcomes in medical research. These models provide a refined method for comprehending the intricate connections between variables in health sciences, making substantial contributions to the progress of medical knowledge and patient treatment. Further investigation is needed to delve into the theoretical and practical elements of ordinal regression, focusing on overcoming methodological obstacles and utilising technological improvements to improve the precision and significance of health research.

1.2 Problem Statement

Within the complex field of medical research, the task of examining tiny sets of data poses a distinct puzzle, especially when it comes to challenges involving ordinal regression. These challenges are particularly evident when it comes to evaluating diseases using medical imaging. The process of collecting a sufficient amount of training data is hindered by the difficulties of data collection, such as the invasive nature of the procedures, the high prices involved, and the extensive labour necessary. For example, Vabalas et al. (2019) shed light on this dilemma by studying

autistic adults. They found that collecting a single set of complex data could require between 1.5 to 4 hours of the experimenter's time and 3.5 to 6 hours of the participant's time, including travel. The recruitment of an adequate number of participants is additionally difficult by the obstacles to entry and involvement within this specific population.

This situation highlights the need for strong and dependable machine learning (ML) methods that can effectively handle the constraints of small sample numbers. Ensuring accuracy in statistical analysis is of utmost importance, as using inappropriate procedures might result in unclear or deceptive conclusions. This can increase research expenses and contribute to an abundance of inconclusive information. Encountering such challenges is frequent, especially when researchers face a lack of accurate data and a shortage of resources in literature, a circumstance that is widespread in the field of health sciences.

This study presents a novel solution to the challenges provided by limited datasets in ordinal regression. It combines two different procedures, namely ordinal regression and bootstrap methods, to handle these issues. Our goal is to close the current methodological gap by integrating and improving these strategies using the R programming environment. This fusion not only has the potential to alleviate the constraints imposed by small sample numbers but also improves the clarity and precision of our analysis. The use of this improved algorithm in different medical case studies showcases its capacity to transform research in the field of health sciences, especially for studies conducted in Malaysia that often face the obstacle of limited datasets.

This undertaking is not simply a reaction to a statistical problem but a significant step towards introducing new methods, which will lead to more

knowledgeable, dependable, and detailed study results. It serves as a symbol of optimism for researchers who are dedicated to extracting significant knowledge from restricted data, therefore aiding the progress of medical science and the improvement of patient care.

1.3 Research Questions

i. What alternative method using R programming can be employed for optimizing ordinal regression in health sciences studies?

ii. How does the new built-in methodology impact decision-making compared to existing approaches?

iii. How does the alternative ordinal regression model perform concerning specific evaluation metrics compared to standard models?

1.4 Research Hypothesis

i. The implementation of a specific R programming algorithm will offer an optimized solution for ordinal regression in health sciences studies compared to traditional methods.

ii. The incorporation of the new built-in methodology will significantly enhance the decision-making process compared to conventional methods.

iii. The alternative method will demonstrate superior performance in model fitting information, standard error, and evaluation metrics such as AIC, BIC, and various pseudo R-squared measures.

1.5 Research Objectives

General objective:

To develop an alternative model for ordinal regression through statistical methodology building.

Specific objectives:

i. To elucidate alternative method for the health sciences study to solve an optimization problem for ordinal regression using R programming algorithm.

ii. To make an inference on a new built-in methodology and enhance the gained result for the decision making.

iii. To validate the alternative method using model fitting information, standard error, and evaluation metrics such as AIC, BIC, and various pseudo R-squared measures.

1.6 Organization of Thesis

This thesis is structured into six main chapters, each addressing key aspects of the research on developing an alternative methodology for ordinal regression in dental sciences. Chapter 1 presents the background of the study, emphasizing the significance of ordinal regression in health sciences. It also defines the problem statement, research questions, hypotheses, and research objectives, setting the foundation for the study. Chapter 2 reviews existing approaches in ordinal regression, highlighting their strengths and limitations. It explores statistical methodologies relevant to small sample sizes, including bootstrap techniques. The chapter also discusses various evaluation and validation techniques used in previous research, providing a comprehensive comparison of ordinal regression methods. Chapter 3 details the research design, study

population, and data collection methods. It outlines the statistical methods employed, focusing on the development of an alternative ordinal regression model incorporating bootstrap techniques. The chapter also explains the validation process using model fitting information, standard error, and evaluation metrics.

Chapter 4 is the results chapter presents findings from the application of the developed methodology to two case studies in dental sciences—Tooth Wear Severity and Tooth Sensitivity. It compares the traditional ordinal regression model with the alternative approach, evaluating improvements in model performance based on log-likelihood, AIC, BIC, and pseudo R-squared metrics. Chapter 5 interprets the findings from the case studies, comparing them with previous research. It assesses the advantages and limitations of the proposed alternative method, highlighting its impact on health sciences research. The final chapter (Chapter 6) summarizes the study's contributions, emphasizing the improvements in statistical modelling for ordinal regression. It discusses the strengths and limitations of the proposed approach and provides recommendations for future research directions.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction to Ordinal Regression

Ordinal regression is a widely utilised statistical technique in disciplines such as psychology, medicine, economics, and social sciences. It is especially useful when dealing with data that consists of ranked categories or levels. Ordinal regression differs from nominal data analysis by recognising and utilising the intrinsic order among categories, hence enabling a more nuanced comprehension of the data. The formulation of this method is painstakingly crafted to accurately represent the ordinal nature of the dependent variable, setting it apart from other techniques used for analysing categorical data. The objective is to forecast a variable that exhibits a natural sequence, therefore yielding more precise and enlightening explanations of the fundamental trends and patterns in the data Agresti & Tarantola (2018).

The theoretical basis of ordinal regression is established in influential studies that specifically address the analysis of categorical data. Agresti (2010) provides a comprehensive overview of statistical techniques used for analysing categorical data, with a particular focus on ordinal regression. The focus of his work is on the assumptions, interpretation, and diagnostics of the models. McCullagh's (1980) formulation of the proportional odds model is a significant contribution to the science of ordinal regression. It demonstrates the model's effectiveness in representing ordinal outcomes. The importance of ordinal regression in research cannot be exaggerated. It provides a strong framework for examining data that is arranged in a specific sequence, therefore enhancing the accuracy of research results and making a substantial contribution to the progress of knowledge in several fields. The significance of ordinal regression in contemporary statistical analysis and research is emphasised by the

references and comprehensive discussions found in publications like Agresti's. This solidifies its place as an essential tool for researchers who work with ordered categorical data Agresti & Tarantola (2018).

There is an extensive collection of parametric ordinal models available in the literature. Perplexingly, each of them possesses their individual designations, and the connections between them frequently lack clarity. Fortunately, most of these models can be classified into three specific model classes: cumulative models, sequential models, and adjacent-category models (Mellenbergh, 1995; Van Der Ark, 2001). To facilitate researchers in utilizing and selecting the most suitable model classes for their research topic and data, we will provide a comprehensive explanation of the underlying reasons for these model classes.

Certain fundamental ordinal models posit the existence of an unobservable latent variable that serves as the basis for the ordinal response variable. The observed variable is seen as a classification of the underlying continuous response. The approach produces uncomplicated models but imposes unnecessary assumptions for the development of ordinal models. Alternative models are based on sequential decision processes that make assumptions about the process that produces the ultimate outcome. While we will take into account these motives for ordinal models, our main objective is to describe ordinal models in a comprehensive manner by examining how models can be created from simpler models, specifically binary models. It explicitly identifies the binary models that are included in basic ordinal models (Tutz & Schneider, 2019).

A taxonomy of ordinal models is built from the construction principle, encompassing both widely used models and recent advancements. The process of developing a taxonomy involves examining different methods for representing ordinal replies. Special emphasis is placed on the representation of additional diversity, such

as dispersion effects, which has been largely overlooked and just recently examined in more detail. By considering additional heterogeneity, one can prevent biased estimations of the impacts of explanatory variables and gain further insights into the effects of these variables. While numerous models addressing heterogeneity have been suggested, the specific aspects being modelled are not always clearly defined. One of the goals is to elucidate the essence of the diversity that is encompassed by these models. In addition, we explore hierarchical models, which appear to be underdeveloped despite their numerous advantages and the ease with which heterogeneity effects can be incorporated. While hierarchical models have been examined in item response analysis, their full potential has not been adequately utilised in ordinal regression. The versatile category of hierarchical models is systematically introduced and integrated into the taxonomy. The taxonomy is expanded by incorporating the recently suggested category of mixture models that address uncertainty in a distinct manner (Tutz, 2022).

2.2 Existing Approaches in Ordinal Regression

Traditional models like cumulative link models are prevalent in ordinal regression. However, they often make stringent assumptions, leading to potential information loss (McCullagh, 1980). Ordinal regression (OR) problems are widespread in real-world settings, but machine learning researchers have often ignored the ordinal relationship between classes in treating them with a nominal (standard) perspective. Many different kinds of studies use ordinal regression, including economic research (Hirk et al., 2019), medical research (Xu et al. 2020; Campisi et al., 2020), and the study of various fields. Xu et al. performed an ordinal logistic regression analysis to identify the various determinants of illness severity in the study of COVID-

19 in China. They found that the ordinal regression method has some limitations in solving the problem. The first limitation is that not all the laboratory parameters, including LDH and D-dimer, were tested in all patients. The role of missing values might be underestimated in relation to illness severity prediction. The second limitation is that the sample size could restrict the generalisability of their findings. It must be tested and proven on a much larger population of patients to predict illness severity.

Cui & Shibusawa (2022) applied ordered logistic regression to assess the effectiveness of COVID-19-related economic policies in the tourism industry in Japan. Their study examined the impact of the Subsidy Program for Sustaining Businesses (SPSB), Go to Travel campaign, and municipal coupon campaigns on business recovery. Their findings revealed key limitations in using ordinal regression for policy evaluation. First, the effectiveness of policies was assessed based on self-reported survey responses, introducing a subjectivity bias in the evaluation process. Second, their study showed that larger businesses benefited more from economic support programs, as coupon campaigns were more effective for businesses with higher sales volumes (Cui & Shibusawa, 2022).

Chen et al. (2022) applied geographically weighted ordinal logistic regression (GWOLR) to analyze COVID-19 infection risk across U.S. counties by incorporating three temporal dimensions: probability of occurrence, duration of the pandemic, and intensity of transmission. Their study categorized counties into four risk levels—High-Risk, Moderate-Risk, Mild-Risk, and Low-Risk—and analyzed socio-economic, demographic, and spatial factors influencing these classifications. The findings revealed key limitations in using ordinal regression for this type of analysis. Firstly, the associations between risk factors and COVID-19 infections were spatially non-

stationary, meaning that the same variables had different effects across regions. Secondly, while the model captured county-level trends, it did not account for individual behaviors, mobility patterns, or vaccination rates, which could refine risk assessments (Chen et al., 2022).

Dewi & Kusumawati (2022) applied ordinal logistic regression to analyze COVID-19 risk zones in Indonesia, categorizing them into four levels: high risk, medium risk, low risk, and no cases. Their study examined how factors such as elderly population, referral hospitals, diabetes mellitus, hypertension, handwashing behavior, male population, and smoking habits influenced COVID-19 risk classification. The findings revealed key limitations in using ordinal regression for such analysis. Firstly, the study relied on secondary data, and some predictor variables—such as diabetes and smoking habits—were estimated based on historical trends, potentially introducing bias and inaccuracies. Secondly, the proportional odds assumption was tested and met in this study, but in other cases, violation of this assumption could lead to model misinterpretation (Dewi & Kusumawati, 2022).

Xu et al. (2020) applied ordinal logistic regression to identify the determinants of COVID-19 illness severity in China. Their study categorized 598 patients into three severity levels: moderate, severe, and critical and analyzed risk factors such as age, comorbidities, laboratory markers, and delays in diagnosis and admission. The findings revealed key limitations in using ordinal regression for this type of analysis. Firstly, older age (≥70 years, OR = 3.419) and hypertension (OR = 3.372) significantly increased the risk of severe illness, confirming previous findings in COVID-19 severity studies. Secondly, elevated cardiac troponin I (cTnI >0.04 ng/ml, OR = 7.464) and myohaemoglobin (>48.8 ng/ml, OR = 2.214) were associated with worse prognosis, highlighting the role of cardiac injury in disease progression. Thirdly,

delays in diagnosis (OR = 1.056) and hospital admission (OR = 1.048) independently predicted increased severity, demonstrating the impact of healthcare system efficiency on patient outcomes. However, the study had limitations, including missing laboratory data for some patients (e.g., LDH and D-dimer levels), which could affect the robustness of the model. Additionally, the sample size was limited to four hospitals in China, which may affect the generalizability of the results (Xu et al., 2020).

Small samples are popular because tasks and experimental protocols that can discriminate between different conditions to the greatest extent possible are still being developed, as well as the costs of data collection involving human participants. Small sample sizes occur in various research experiments, particularly preclinical studies, for ethical, financial, and general feasibility reasons. The first problem that one encounters when conducting small experiments is neither the high dimensionality of the data nor the accurate type-1 error rate control of the methods. Many current statistical methods necessitate moderate or large sample sizes, and as a result, when sample sizes are small, the type-1 error rate is not well controlled. In work conducted by Vabalas et al., 2019, obtaining a large-dimensional dataset may take between 1 and half to 6 hours of experiment time and 3 and half to 6 hours of participant time (including travel time). It is also challenging to access many adults on the autistic spectrum because of issues recruiting participants and encouraging their participation. However, in many cases, it is important to have machine learning (ML) algorithms that work with smaller datasets. The computational power of the bootstrap method comes from its distributed nature. bootstrap method involves iteratively resampling a dataset with replacement. the bootstrap method may not be effective for small sample sizes (Zou et al., 2008).

Chaikh et al. 2017 using bootstrap method to help the radio oncologist and the medical physicist to usefully analyze the dosimetric data obtained from small-sized samples, with few patients. The bootstrap method was applied to the original data set to assess the dose differences and evaluate the impact of sample size on the 95% confidence interval (95%.CI). The bootstrap simulation with 1000 random samplings can be used for small populations with n = 10 and provides a true estimation. They stated that, one must be cautious when implementing this method for radiotherapy: the data should be representative of the real variations of the cases and the cases should be as homogeneous as possible to avoid bias of over/under estimation of the results (Chaikh et al. 2017).

Stoma et al. 2019 study about bootstrap analysis of the production processes capability assessment. In their study using bootstrap method of assessing the capability of the manufacturing process. They confirm both the narrowing of the confidence interval when using the bootstrap method and the possibility of determining a better estimator of the lower limit of this range compared to the results obtained using the classic method. The tests carried out for the unit production of shafts with grooves showed that the analysis of the process capability for measuring tests n = 10 is possible. The model for assessing the capability of production processes presented in their research was implemented in low-volume production in the defence industry.

In their 2022 study, Dickey et al. conducted a bootstrap simulation consisting of 10,000 repetitions to showcase the effectiveness of logistic regression (either binary or ordinal, with proportional or non-proportional odds) in analysing the outcomes of epilepsy surgery. This method is particularly useful in studies with limited sample sizes. The researchers emphasised the method's capacity to greatly enhance statistical

power and accurately measure the clinical importance of different levels of seizure control. This was demonstrated by the percentage of simulations where a P-value was found to be significant for a log-odds coefficient less than 0.05. This persuasive argument advocates for the use of ordinal logistic regression, not only due to its statistical robustness but also because of its ability to uncover crucial insights in medical research, particularly where complex outcome measures are crucial and sample sizes are restricted. The authors strongly support the wider use of ordinal regression, highlighting its significance in capturing the complete range of patient outcomes and contributing to a more detailed understanding of the impacts of epilepsy surgery.

Meijer and Baneke (2004) investigate the performance of goodness-of-fit tests for the Rasch model in small sample sizes, highlighting the model's assumptions and the need for robust statistical methods in such contexts. Employing Monte Carlo simulations, they assess various tests under different sample size conditions and utilize bootstrap techniques to address data scarcity challenges. Their study underscores the efficacy of ordinal regression in analysing categorical data with modest sample sizes, emphasizing the importance of considering the ordinal nature of response variables. By incorporating bootstrap methods, they enhance the reliability of their findings, offering valuable insights into statistical techniques for analysing categorical data in the context of small sample sizes. Their work contributes to advancing methodological rigor and improving understanding of complex phenomena across various fields of inquiry.

Riley et al.'s work from 2021 critically assesses the validity of shrinkage and penalization techniques in the creation of clinical prediction models, especially when dealing with small sample sizes. These techniques, which aim to lower mean-square

prediction error for new individual predictions, include ridge regression, the lasso, and elastic net. They are intended to counteract overfitting by decreasing predictor effect estimates towards the null. The study does, however, draw attention to a major obstacle: the degree of uncertainty in "tuning parameters," or penalty terms, estimated from the development dataset. In smaller sample sizes, this uncertainty is particularly noticeable, which could result in clinical prediction models that are not entirely trustworthy. Although these techniques can enhance model performance overall, Riley et al. (2021) stress that they cannot ensure accurate predictions in all datasets, especially those with small effective sample sizes and models that predict binary and time-to-event outcomes with low Cox-Snell R2 values.

Building on these discoveries, Riley et al. (2021) advocate applying penalization and shrinkage methods with caution, arguing that there is no one-size-fits-all way for creating accurate prediction models, particularly when there is a lack of data. In order to reduce the possibility of overfitting and precisely estimate important model parameters, such as shrinkage and tuning parameters, the study recommends using higher effective sample sizes during the model-development process. Additionally, it recommends using bootstrap techniques as a tactic to lessen shrinkage estimate variability, especially in smaller samples. This thorough study offers a nuanced viewpoint on the use of penalization and shrinkage techniques in the creation of clinical prediction models, emphasising the value of customised strategies that take into account the unique properties of the dataset and the predictive model's objectives.

In their 2009 study, Kiralj and Ferreira explore the regression model validation procedures used in QSAR and QSPR investigations. They emphasise the vital role that validation plays in guaranteeing the accuracy and efficiency of these models,

particularly when dealing with small sample numbers. Their thorough analysis evaluates the efficacy of several validation strategies, such as external validation, y-randomization, bootstrapping, leave-one-out cross-validation (LNO), and bootstrapping, using four different data sets from the literature. Their results highlight the importance of training set size as a critical component influencing model performance, with models derived from smaller datasets demonstrating potential for failure and atypical behaviours during validation processes, as well as limitations in their ability to undergo full validation. Kiralj and Ferreira provide insightful information about the methods for improving the reliability of QSAR and QSPR models by introducing a novel method for calculating the critical N in LNO and showing that a small number of y-randomization or bootstrapping cycles are sufficient for typical model validation. This emphasises the need for rigorous validation and careful variable selection to support the integrity of these models in scientific research.

2.3 Evaluation and Validation of Proposed Methodologies

Hosmer et al. (2013), concentrating on R algorithms for ordered logistic regression, highlight the importance of carefully assessing and validating approaches in the field of statistical prediction models. Their work emphasises the need for such metrics to ensure the accuracy and dependability of prediction outcomes by carefully describing the validation procedure against original data. This method not only guarantees the validity of the approaches' outputs but also boosts confidence in their use in a variety of domains where prediction models are crucial. Hosmer and colleagues make an important contribution to the improvement of statistical modelling approaches by laying out a framework for rigorous validation. This framework serves as a baseline for future studies that aim to build or improve upon prediction algorithms.

The focus on validation aligns with the larger scientific community's goal of developing strong and trustworthy procedures. The validation process is essential to the construction of statistical models since it guarantees that the models are both theoretically and practically sound, as shown by Hosmer et al. (2013). This thorough approach to validation fosters a culture of rigorous verification, which can greatly improve the calibre and applicability of statistical approaches. It also serves as a crucial reminder of the significance of accuracy and dependability in predictive modelling. Hosmer and associates' work not only advances statistics but also establishes a high bar for prediction model validation, highlighting the critical role that thorough assessment plays in the quest for scientific greatness.

The study conducted by Ghazali et al., 2020 aims to combine Ordered Logistic Regression (OLR) and Multilayer Perceptron Neural Network (MLP) to improve the process of selecting variables for predicting hypertension outcomes. This study highlights the importance of choosing factors such as smoking, total cholesterol, and triglycerides because of their robust correlation with hypertension levels, as determined by bootstrap methodology and OLR. The performance of the MLP model is evaluated using Predicted Mean Square Error (PMSE), emphasising the significance of precise variable selection in predictive models. The research additionally provides R syntax for OLR (Ordinary Least Squares Regression) and MLP (Multilayer Perceptron) to facilitate reproducibility and comprehension of the methodology. This positions this approach as a helpful instrument for decision-makers in the medical domain, namely in the management of hypertension patient outcomes.

A deeper look into the field of statistical model validation reveals the work of Harrell (2015), who offers a thorough analysis of regression modelling strategies, including the use of resampling techniques like bootstrapping and cross-validation to validate predictive models. In order to prevent overfitting and accurately measure model prediction error, Harrell's methodology emphasises the need of internal validation. It also focuses on evaluating model performance and its generalizability to independent data sets. His method emphasises the importance of validation in the creation of trustworthy predictive models and provides a thorough manual that covers a variety of regression analyses in addition to logistic regression. The significance of this work is in its ability to provide benchmarks for model assessment, guaranteeing that predictive models are not only robust statistically but also practically applicable in various fields.

Further exploring the clinical side of model validation, Steyerberg et al. (2010) offer a useful methodology for evaluating and validating prediction models in clinical research. Their research primarily focuses on the external validation of models, examining the performance of a model created in one context in a different population. In their methods for evaluating model calibration and discrimination in novel patient cohorts, Steyerberg and colleagues highlight the significance of external validation in confirming the robustness and broad applicability of predictive models. In the clinical setting, where precise patient outcome prediction can have a substantial influence on clinical decision-making and patient care, this validation component is essential. When taken as a whole, these studies highlight the complexity of model validation, from statistical foundations to clinical applications, and they reinforce the critical role that exhaustive validation procedures play in the creation of strong, trustworthy prediction models.

2.4 Comparison and Assessment of Alternative and Ordinal Regression Method

Considerable progress has been made in the study of statistical approaches to data analysis, especially in the area of predictive modelling. The Elements of Statistical Learning, a seminal work by Hastie, Tibshirani, and Friedman (2009), examines a range of modelling approaches by means of a rigorous comparison and evaluation based on goodness of fit and model fitting metrics. Their thorough evaluation covers a broad range of techniques, from straightforward linear models to more intricate processes like bootstrap techniques and ordinal regression, and is an essential tool for comprehending the advantages and disadvantages of each strategy. This paper highlights the significance of strong model validation in addition to providing guidance on optimal modelling strategies based on particular data features.

One particularly noteworthy application of ordinal regression, a statistical method for predicting an ordinal result, is in the social sciences and medical research, among other domains. Agresti (2010) provides a thorough introduction to ordinal regression models, including information on how to apply and analyse them, in Analysis of Ordinal Categorical Data. With an emphasis on ordinal outcomes—categorical variables with a meaningful order—specialized methods are needed to guarantee precise modelling and forecasting. Agresti's research emphasises how important it is to comprehend ordinal models' underlying presumptions in order to apply them to real-world data in an efficient manner.

Combining bootstrap techniques with ordinal regression offers a novel way to improve the accuracy and dependability of the model. In "*An Introduction to the Bootstrap*", Efron and Tibshirani (1993) present the bootstrap method as a potent tool for evaluating statistical accuracy. With the use of replacement sampling from an

initial dataset, the resampling technique known as bootstrap makes it possible to estimate the distribution of a statistic. Bootstrap techniques can be used to improve model assessment and validation efforts by offering insights into the stability and variability of model estimates in ordinal regression.

Studies contrasting conventional ordinal regression with its integration with bootstrap techniques highlight specific benefits and drawbacks. In their book Bootstrap Methods and Their Application, Davison and Hinkley (1997) go over the usefulness of bootstrap methods in statistical analysis, particularly regression models. They draw attention to how bootstrap can help with model uncertainty and offer more detailed estimates of model parameters—a feature that is especially helpful for ordinal regression, where the intricacy of the model might make it difficult to interpret the findings.

The practical implications of using statistical learning techniques, such as ordinal regression and bootstrap, in data analysis are expounded upon by James et al. (2013) in "*An Introduction to Statistical Learning*". They offer a clear and accessible introduction to these methods, highlighting their importance in generating well-informed forecasts from available data. This work provides a more approachable starting point for the implementation of these techniques across a range of disciplines, complementing the theoretical frameworks established by Hastie et al. (2009).

In their paper on the assessment of clinical prediction models, Steyerberg et al. (2010) discuss the value of model performance metrics and validation in the context of medical research. They emphasise the importance of model evaluation methods, such as bootstrap validation, in guaranteeing the validity and suitability of predictive models in healthcare settings in their discussion on the use of ordinal regression for clinical data.

Furthermore, a thorough manual for creating predictive models, including ordinal regression, may be found in Harrell's (2015) Regression Modelling Strategies. Harrell highlights how crucial model validation methods like bootstrap are for determining model fit and forecasting accuracy. His method promotes giving model assumptions and the real-world effects of model selections considerable thought.

The trade-off between interpretability and model complexity is emphasised by contrasting ordinal regression models with and without bootstrap integration. In Applied Predictive Modelling, Kuhn and Johnson (2013) explore this contrast and offer case studies that highlight the trade-offs associated with choosing modelling approaches. According to their analysis, bootstrap techniques can improve model reliability, but they should be used with caution because to their increased computational complexity and effort.

The decision between these approaches is based on how well they predict outcomes and how feasible they are in practice, especially in clinical research where precise patient outcome prediction is crucial. Pencina and D'Agostino (2014) argue for a careful assessment of model performance and validation techniques as they address the consequences of model selection in the context of clinical risk prediction.

The literature, in summary, offers a nuanced perspective on the relative benefits of combining bootstrap techniques with ordinal regression as opposed to more conventional ordinal regression methodologies. The particular setting of the study, including the type of data, the goals of the investigation, and pragmatic factors like computational resources, all influence which of these approaches is best. The continual evaluation and validation of modelling techniques is essential to improving our knowledge and utilisation of these approaches as statistical analysis develops further.

2.5 Tooth Wear, Tooth Sensitivity, and Risk Factors through Ordinal Regression

Tooth wear is an increasingly recognized concern in dental sciences, with multiple intrinsic and extrinsic factors influencing its severity. Unlike dental caries or trauma, tooth wear results from gradual loss of tooth structure through attrition, abrasion, erosion, and abfraction (Khan et al., 2021). The progressive nature of this condition makes early identification and risk assessment crucial for preventing long-term oral health complications. As researchers continue to explore predictive models, ordinal regression has emerged as an effective statistical tool for quantifying the impact of different risk factors on the severity of tooth wear (Smith & Robinson, 2018; Schierz et al., 2014).

Among the key risk factors, dietary habits have been extensively studied. The frequent consumption of acidic foods and beverages is a well-documented contributor to enamel erosion. Smith and Robinson (2018) examined a sample of 500 patients and found that individuals with high-acid diets had significantly increased odds of severe tooth wear. The erosive potential of citrus fruits, vinegar-based dressings, and carbonated drinks has been particularly concerning, as these foods lower the pH in the oral cavity, leading to mineral loss from the enamel surface (Nausheen et al., 2022). These dietary patterns not only accelerate wear but also increase the risk of dentine hypersensitivity, as the loss of enamel exposes underlying dentin to external stimuli.

Bruxism, a condition characterized by involuntary teeth grinding and clenching, is another major factor in excessive tooth wear. Whether occurring during sleep (sleep bruxism) or while awake (awake bruxism), this parafunctional habit exerts significant pressure on teeth, causing enamel attrition over time. While some studies argue that there is no direct association between incisal tooth wear and temporomandibular