وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم الإحصاء والمعلوماتية

إقتراح خوارزمية حصينة لإختزال متغيرات الإنحدار مع المحاكاة والتطبيق على بيانات مرضى الثلاسيميا في نينوى

رسالة مقدمة
إلى مجلس كلية علوم الحاسوب والرياضيات
في جامعة الموصل كجزء من متطلبات نيل شهادة
ماجستير علوم في الإحصاء

من قبل
فاطمة محمد أحمد حسين

بإشراف
أ.م. د. بشار عبدالعزيز مجيد الطالب

١٤٤٤ هـ
٢٠٢٣ م

المستخلص

تقوم فكرة الرسالة على تقليل أو استبعاد تأثير عدم تحقق فرض التوزيع الطبيعي للبيانات، بسبب وجود أنواع من القيم الشاذة فيها عند الرغبة في إختيار افضل معادلة إنحدار بالطرق الحصينة، وتمَّ تحقيق ذلك من خلال إدخال أوزان من طرق حصينة في التقدير واختبار حصانتها وملاءمتها للنموذج مسبقاً وبعد أكمال عمليات التقدير بالأساليب الموزونة يتم إختيار الأوزان الناتجة من أعلى الطرق الحصينة كفاءة وإدخال هذه الأوزان في كل خطوة من خطوات طرق أختيار أفضل معادلة إنحدار أي أن كل عملية ادخال أو استبعاد لمتغير في مرحلة ما سيكون بواسطة دالة موزونة بوزن حصين ضد القيم الشاذة، فينتج عن ذلك نموذج يحقق صفتين وهما الحصانة وتقليل الأبعاد مع زيادة الكفاءة في ان واحد.

وقد تم استخدام اسلوب المحاكاة على نماذج بأبعاد (عدد المتغيرات المستقلة في الانموذج) مختلفة وأحجام عينات مختلفة ونسب تلويث مختلفة في المتغير المعتمد مرة، وفي المتغيرات المستقلة مرة اخرى وفي الاثنين معاً مرة ثالثة كما تغيير قيمتي الوسط الحسابي والتباين، مع التركيز على دراسة احتمال تأثير وجود القيم الشاذة على المتغيرات التي ستبقى في الانموذج والمتغيرات التي سيتم حذفها وقد نتجت لدينا حالات كانت النماذج النهائية فيها بأبعاد مختلفة عن النماذج النهائية بدون وجود القيم الشاذة(أي أن الانموذج الامثل لم يكن نفسه في كل الحالات).

ولتحقيق فكرة الرسالة تمّتْ مقارنة عدد من طرق التقدير الحصينة ومقارنة النتائج مع طريقة المربعات الصغرى الإعتيادية (Ordinary Least Squares Method) OLS وطريقة (Least Absolute Shrinkage and Selection Operator) LASSO الحصينة المكيفة (Adaptive Robust Lasso) على بيانات تجريبية بأستخدام المحاكاة بتنفيذ السيناريوهات المشار اليها في اعلاه وكذلك على بيانات لعينة من مرضى الثلاسيميا في محافظة نينوى، وتمَّ بعد ذلك استخدام أسلوب مقدرات M لتقدير أنموذج الإنحدار، بالأعتماد على ثلاثة دوال وزن وهي Huber وHampel وBisquare، وتمَّ توظيف مقدرات MM وS في تحصين عملية التقدير، وتمَّ كذلك تغيير مقدرات التباين، حيث تمَّ استخدام مقدر وسيط الأخطاء المطلقة (Mean Absolute Deviation) MAD، وتمَّ تغيير المقدر الإبتدائي ليكون مقدر المربعات الصغرى الاعتيادية OLS مرة، ومقدر المربعات الصغرى المشذبة (Laest Trimmed Squares Method) LTS مرة أخرى، ونتج لدينا اربع وعشرون مقدراً مختلفاً مع مقدري OLS و LASSO المكيفة . وفي النهاية تمَّ استخدام الأوزان التي نتجت عن كلٍّ مقدر لوزن طريقة المربعات الصغرى الإعتيادية للحصول على مقدرات المربعات الصغرى الموزونة . وقد تمّت المقارنة بين النماذج باستخدام بعض معايير المقارنة مثل معامل التحديد R^2 ومعامل التحديد المعدل $\overline{R}^2$ واحصاءة مالو CP وقيمة مجموع مربعات الأخطاء (Residuals Sum of Suquares) RSS، كما تمّت مقارنة نتائج المحاكاة باستخدام معيار جذر متوسط مربعات الخطأ (Root Mean Squares Error) RMSE.

وتم في نهاية الرسالة أنتخاب الطريقة الأكثر كفاءة في تجارب المحاكاة لكل حالة من حالات وجود القيم الشاذةفي البيانات (في المتغير المعتمد أو المتغيرات المستقلة او الاثنين معاً) لتكوين مايشبه الدليل الأسترشادي للباحثين في تحديد طريقة التقدير الملائمة ومن ثم إدخال الطرق التي أثبتت كفاءتها في مراحل أختيار أفضل معادلة انحدار. وقد تم ذلك من خلال خوارزمية تبين خطوات عملية اختيار المتغيرات فضلاً عن مخطط أنسيابي Flowchart يصف الخوارزمية المقترحة.

Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Statistics and Informatics

proposing Robust Algorithm to Reduce Regression Variables with Simulation and Application on Thalassemia Patients Data in Nineveh

A Thesis Submitted to the Council of the College of

Computer Science and Mathematics

University of Mosul as a Partial Fulfillment of Requirements

for the Degree of Master of Science

in

Statistics

By

Fatima Mohamed Ahmed Hussein

Supervised by

Assist. Prof. Dr. Bashar Abdulaziz Majeed Al-Talib

2023 A.D **1444 A.H**

Abstract:

The idea of this paper is based on reducing or excluding the effect of not satisfying the assumption of normal distribution of the data because of the presence of types of outlying values in it when choosing the best regression equation by the robust methods, and this was achieved by using weights from the robust methods in the estimate and testing their robustness and suitability for the model in advance and then Choosing the weights resulting from the highest efficient robust methods and inserting these weights into the selection stages of methods to choose the best regression equation, resulting in a model that achieves two characteristics at the same time, which are robustness and reducing dimensions in return for increasing efficiency.

The simulation was used on different sample sizes and different contamination rates in the dependent, independent, and in both together a third time, with a focus on studying the possible impact of the presence of outliers on the variables that will remain in the model and the variables that will be deleted.

To achieve the idea of the paper, a number of estimation methods were compared and the results were compared, where the ordinary least squares method (OLS) was applied to the data of thalassemia patients in Nineveh Governorate, and then the M estimator method was used to estimate the regression model, and three weighting functions were used, namely Huber, Hampel and Bisquare, and the weight functions were changed Through the use of MM and S estimators, the variance estimators were also changed, so the Median Absolute Deviation estimator (MAD) was used, and the initial estimator was changed to be the least squares estimator once and the least trimmed of squares estimator LTS again, and then we have thirty different estimators.

In the end, the weights that resulted from each estimator were used to weigh the Ordinary Least Squares method to obtain the least squares estimators with different weights. A comparison was made between the models using some comparison criteria such as the Coefficient of Determination R^2 , the Adjusted Coefficient of Determination R^2_{adj} , Mallowss CP statistics, and the Residuals sum of squares RSS, and the simulation results were compared using the standard Root Mean Squares Error RMSE.