Journal of Statistics & Applied Sciences, Issue -10
ISSN 2718-0999

Research Article

Received: date:30.07.2024
Accepted: date:24.11.2024
Published: date:31.12.2024

Residual Modelling as a New Approach for Variable Selection

Asli Nurefşan Kocak[1], Muhammet Furkan Daşdelen[2], Mehmet Kocak[3]*

[1]Istanbul Medipol University, International School of Medicine, Istanbul, Turkey asli.kocak@std.medipol.edu.tr
[2]Istanbul Medipol University, International School of Medicine, Istanbul, Turkey muhammed.dasdelen@std.medipol.edu.tr
[3]Istanbul Medipol University, International School of Medicine, Istanbul, Turkey mehmetkocak@medipol.edu.tr
Orcid:0009-0000-5367-7443[1] Orcid:0000-0003-2251[2]-2093 Orcid:0000-0002-3386-1734[3]
***Correspondence:** mehmetkocak@medipol.edu.tr

Abstract: Variable selection in statistical model building still has challenges to overcome as the depth and breadth of the research data is expanding. To help reduce this challenge, we introduce a new approach in variable selection, called residual modeling, which can be applicable regardless of the number of predictors. We compare the statistical power and type-1 error retainment of the forward, backward, and stepwise variable selection approaches with the proposed modeling strategy controlling for known predictors. In Residual Modeling, each predictor enters the model as a single predictor, whose resulting residuals become the dependent variable for the next predictor, and so on. We compare these models under different scenarios with varying sample sizes and various combinations of significant and insignificant predictors. When there exist known predictors from the literature, in identifying new significant predictors controlling for these known predictors, Residual Modelling shows higher statistical power especially as the number of predictors increases compared to the other variable selection methods used. It also has reduced bias in parameter estimation and reduced standard errors. The Type-1 error was retained at its nominal level for Residual Modelling while forward, backward, and stepwise variable selection approaches had slightly reduced Type-1 Error rates. When dealing with multiple predictors in the presence of known significant predictors, Residual Modelling offers a practical solution without causing loss of statistical power or increased Type-1 Error Rate.

Keywords: variable selection, dimension reduction, forward selection, backward selection, stepwise selection, residual modelling

1. Introduction

In statistical modelling, we often deal with multiple predictors (i.e., independent variables) for a given response variable (i.e., dependent variable). The immediate, and typically the easiest, approach would be to fit a model with all the predictors included in the model and assess the variables that show significant association with the response variable. Then, the model is 'cleaned' by removing the insignificant variables. The primary goal here is to choose the right set of variables for inclusion in a statistical model at the end. There are several approaches to variable selection, each with its own methodology and criteria for selection.

Forward Selection: This approach begins with an intercept-only model (i.e., no predictor in the model) and adds variables one at a time. At each step, the variable that provides the most significant improvement to the model is added, until no significant improvement can be made[1].

Backward Elimination: In contrast to forward selection, backward elimination starts with all candidate variables included in the model. Variables are removed one at a time if they are not statistically significant, with the least significant variables removed first [1,2].

Stepwise Selection (also known as Mixed Selection): A combination of forward selection and backward elimination, stepwise selection involves adding variables as in forward selection but also includes a step where variables can be removed if they no longer provide a significant contribution to the model in the presence of newly entered predictors. That is, a variable can go in and out of the model as the model building progresses [2].

Citation : A. N. Kocak, M. F. Daşdelen, and M. Kocak, "Residual Modelling as a New Approach for Variable Selection," Journal of Statistics and Applied Sciences, no. 10, pp. 86–94, Dec. 2024, doi:10.52693/jsas.1525029.

In the above selection approaches, instead of using the p-values, approaches like the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) can also be used to compare models. These criteria help in selecting a model that balances good fit with simplicity, penalizing models with excessive variables [3].

Each of these methods has its own set of advantages and disadvantages and can be chosen based on the specific requirements of the statistical analysis being performed, and each may affect the stability, unbiasedness, and validity of the final model [2,4]. All these approaches suffer from overfitting (especially with Backward selection approach), existing multicoliearity among predictors, inflated Type-1 Error due to multiple testing.

In this work, we introduce a new variable selection approach, called Residual Modelling. The term "Residual Modelling" has previously been used in the literature mainly within the context of forecasting [5,6] and voice and video editing [7,8]. We use the term specifically within the context of regression model building and variable selection.

2. Materials and Methods

This study has been carried out according to the Helsinki declaration with the Istanbul Medipol University Ethics Committee review and approval on September 30, 2019 (Ethics Committee Application No: 10840098-604.01.01-E.53819).

For a response variable $\boldsymbol{Y}$ with n-observations, the intercept-only regression model can be expressed mathematically as follows:

$$Y_i = \mu + \varepsilon_i, i = 1,2,3, \dots, n$$

If there are 5 predictors (i.e., independent variables) in addition to the known predictors (KP), for example, $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{X_4}$, and $\boldsymbol{X_5}$, the model that contains all predictors is called Full Model, and mathematically expressed as follows: $Y_i = \beta_0 + \boldsymbol{\beta}_{KP} * \boldsymbol{X}_{KPi} + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \beta_3 * X_{3i} + \beta_4 * X_{4i} + \beta_5 * X_{5i} + \varepsilon_i$

where $i = 1,2,3, \dots, n$, and $\boldsymbol{\beta}_{KP}$ vector of parameters represents the effects of the known predictors $\boldsymbol{X}_{KP}$. In the above model, β_0 is the intercept of the model, representing the average value of the response variable $\boldsymbol{Y}$ when all the effects of the predictors are equal to zero. The parameter β_1 expresses the predicted change in $\boldsymbol{Y}$ for each unit of change in $\boldsymbol{X_1}$ affects $\boldsymbol{Y}$ keeping all other predictors fixed, β_2 shows expresses the predicted change in $\boldsymbol{Y}$ for each unit of change in $\boldsymbol{X_2}$ keeping all other predictors fixed, and so on. The error term ε represents random error, which is assumed to be independently and identically distributed Gaussian variates with zero mean and variance σ^2.

Setting entry and stay criteria around p-values, forward, backward, and stepwise variable selection strategies can be employed. Statistical packages also allow the user to set certain variables (e.g., known predictors-KP) to be kept in the model regardless of their significance. This is a critical functionality as we would be more interested in testing whether or not a given candidate predictor is significantly associated with the outcome variable controlling for the known factors and predictors.

2.1. A New Variable Selection Approach: Residual Modelling

In residual modelling, each predictor enters the model as a single predictor; then, the potential effect of this predictor is removed from the response variable by using the residuals of this model as the response variable for the next potential predictors. This process continues until all variables are tested. Table-1 shows an illustration of the Residual Modeling approach with 5 predictors.

Table 1: Residual Modelling Framework with five predictors ($\boldsymbol{\beta}_{NV}$ represents the effects of known predictors $\boldsymbol{X}_{NV}$)

Step 1: X_1 is the only predictor in the model	$Y = \beta_{01} + \boldsymbol{\beta}_{KP} * \boldsymbol{X}_{KP} + \beta_1 * X_1 + \varepsilon_1$
Step 2: The residuals (ε_1) from Step 1 are used as the response variable to assess the effect of $\boldsymbol{X_2}$	$\varepsilon_1 = \beta_{02} + \beta_2 * X_2 + \varepsilon_2$
Step 3: The residuals (ε_2) from Step 2 are used as the response variable to assess the effect of $\boldsymbol{X_3}$	$\varepsilon_2 = \beta_{03} + \beta_3 * X_3 + \varepsilon_3$
Step 4: The residuals (ε_3) from Step 3 are used as the response variable to assess the effect of $\boldsymbol{X_4}$	$\varepsilon_3 = \beta_{04} + \beta_4 * X_4 + \varepsilon_4$
Step 5: The residuals (ε_4) from Step 4 are used as the response variable to assess the effect of $\boldsymbol{X_5}$	$\varepsilon_4 = \beta_{05} + \beta_5 * X_5 + \varepsilon_5$

The order of the variables entering the model selection can be done in a random fashion as well although it should not make a difference under the independence assumption. With Residual Modelling, we hope to answer the following question: Does $\boldsymbol{X_2}$ have significant association with $\boldsymbol{Y}$ after removing the association of $\boldsymbol{X_1}$ with $\boldsymbol{Y}$. Then, we move to the next variable to assess its association with $\boldsymbol{Y}$ after removing the effects of $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$.

Similar to the competitive variable selection approaches, known predictors enter the Residual Modelling strategy in Step-1; that is, the effects of the known predictors (KPs) are removed from the response variable in Step-1 and the residuals from this model becomes the response variable for the next candidate predictor to be tested.

We compared the performance of these four approaches in terms of the statistical power, Type-1 error retention, estimation bias and standard errors through simulations.

Simulation Design:

- Number of predictors: 5, 10, 20 ($X_1 - X_{20}$) of whom two are binary factors (X_6, X_8)
- Response Variable-1:
  $Y1=120+0.2X_1+0.2X_5+0.4X_6+0.4X_8+0.2X_{10}+\varepsilon$
- Response Variable-2: $Y2=120+0.2X_1+0.2X_2+0.2X_3+0.2X_4+0.2X_5+0.4X_6+0.2X_7+0.4X_8+0.2X_9+0.2X_{10}+\varepsilon$
- Assumed Known Predictors: X_1 only, X_{10} only, X_1 and X_5, X_1 and X_5 and X_{10}
- Sample size: 100, 125, 150, 175, 200 (Considering at least 10 records per predictor)
- Number of simulation repeats: 1000

The simulations were summarized as the number of significant runs for each of the predictors, where the percentage of the runs with significant results for insignificant variables will represent the empirical Type-1 Error Rate and the percentage of the runs with significant results for significant variables will represent the empirical Statistical Power. Parameter estimates and their standard errors were also summarized as the averages across the simulation runs.

All analyses were conducted in a parallel manner in SAS® Version 9.4 (SAS Institute, Cary, North Carolina, USA).

3. Results

Empirical statistical power advantage of Residual Modelling was apparent especially as the sample size increases (Figure 1-2, Supplementary Figures A1-2). The performance of the Backward Elimination (BE) approach is the second best overall.

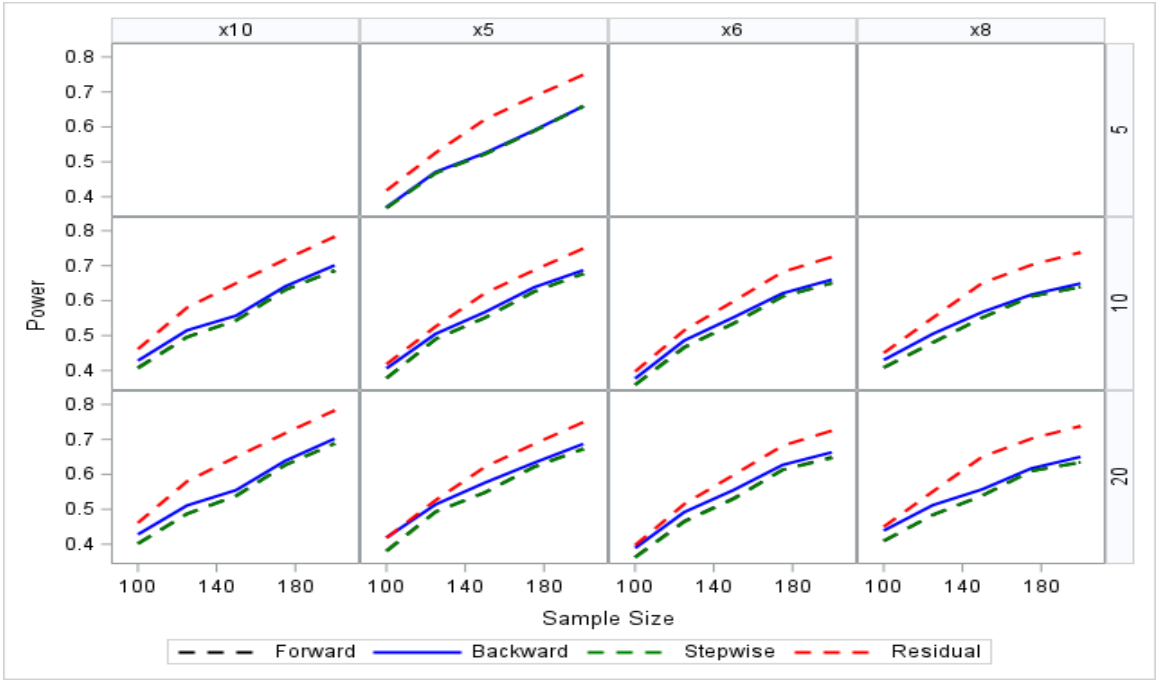


Figure 1: Empirical Power when X1 is the known significant predictor

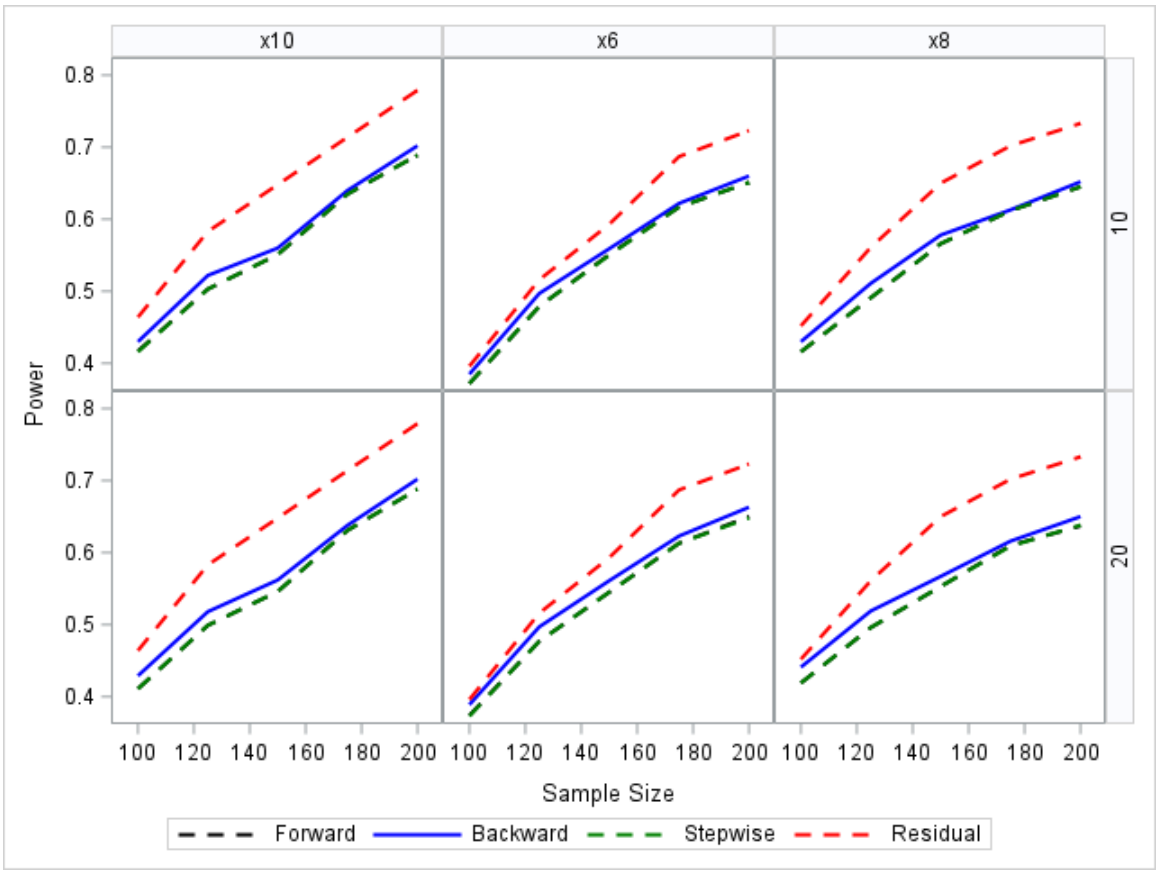


Figure 2: Empirical Power when X_1 and X_5 are considered as the known significant predictors

In Figure 3, we observe that the Type-1 Error Rate is retained by Residual Modelling at around the nominal level of 0.05. The other models underestimate the Type-1 Error rate and it seems to be a function of sample size as well. We observe an overall decline in Type-1 Error as the sample size increases.

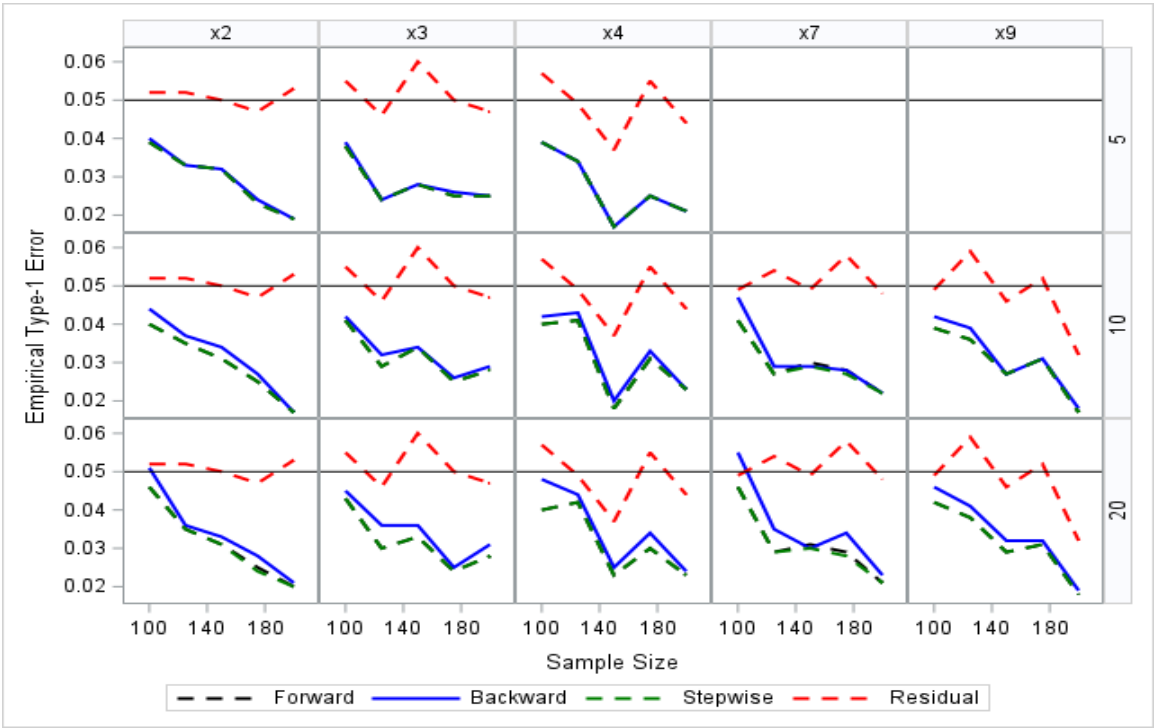


Figure 3: Empirical Type-1 Error rate for representative insignificant variables

Residual Modelling resulted in stable parameter estimation both for the known and new significant predictors; although the parameter estimation of the known predictors by the other three variable selection approaches was stable for the known predictors, they overestimated the parameters for the new significant predictors, which improved as the sample size increases as shown in Figure-4. This situation was the same for the two binary predictors (Supplementary Figure A3).

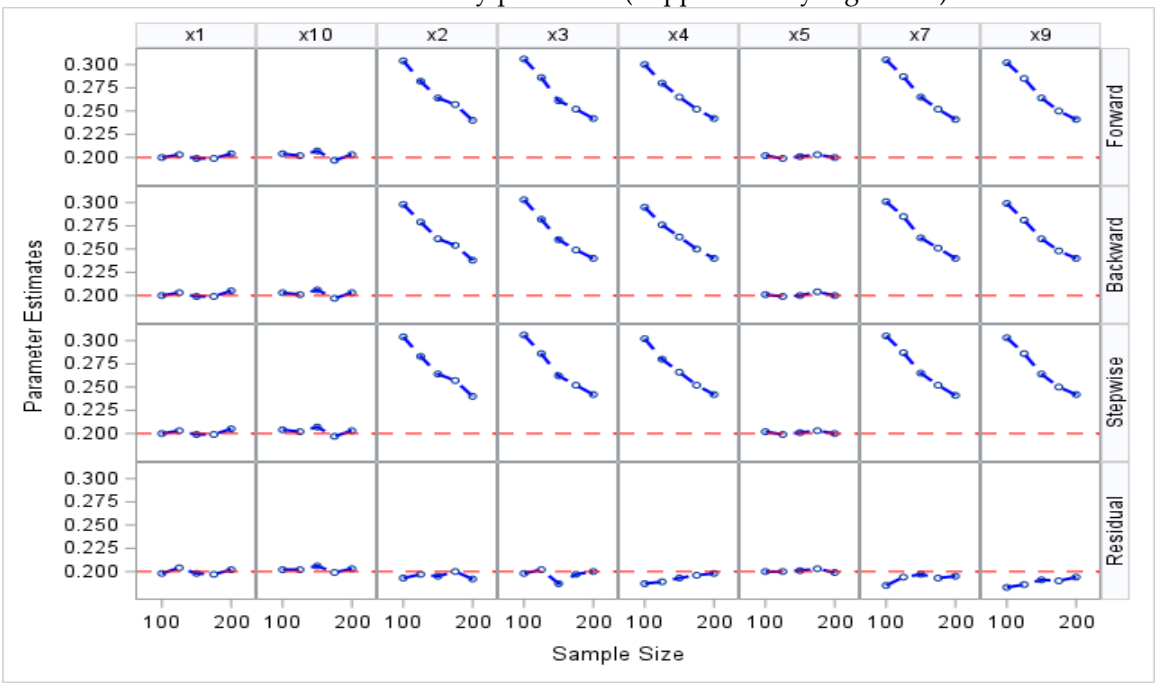


Figure 4: Parameter Estimates for the continuous significant predictors by Variable Selection methods

Standard error was elevated for Residual Modelling especially with higher number of predictors in the model compared the other three approaches as shown in Figure-5. Standard error improves with the increases sample size as anticipated.

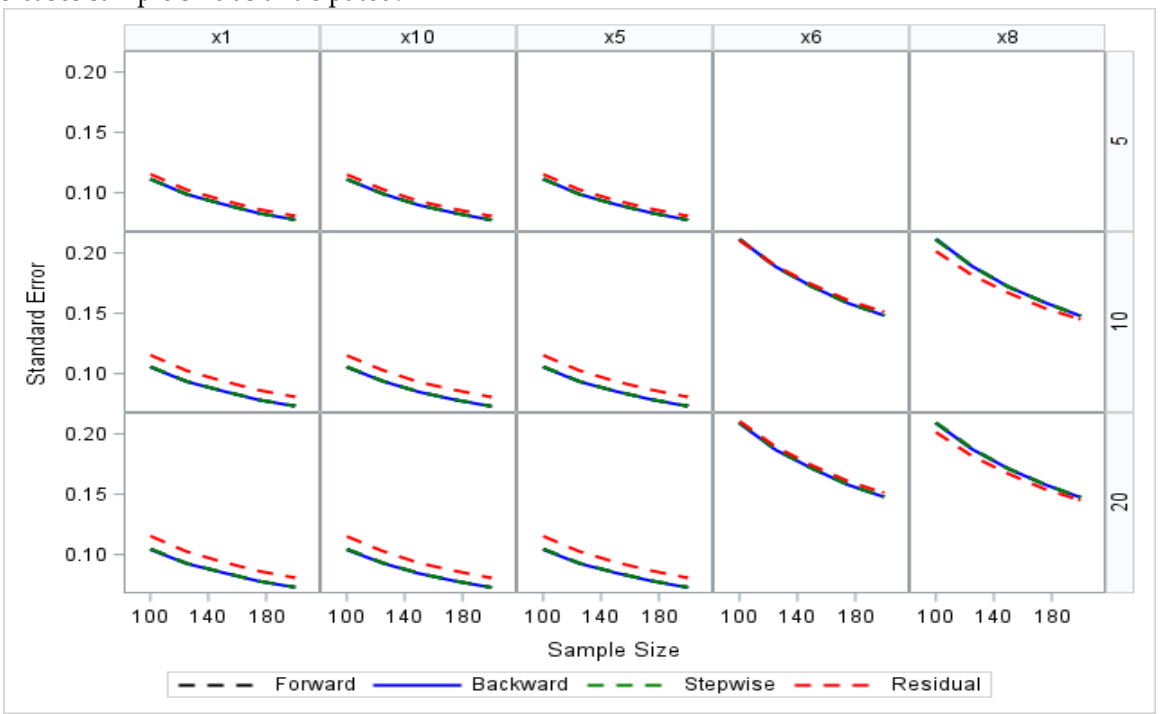


Figure 5: Percentage of simulation repeats showing significance rate of redundant predictors in both models.

The same set of simulations were run under two additional scenarios: Scenario-1: Known predictors were applied in all models and the residuals of these models were used as the response variables for

the forward, backward, and stepwise variable selection approaches. Scenario-2: Candidate variables were shuffled to assess the impact of random order of predictors on the performance. Both scenarios resulted in very similar performances in all models, indicating the robustness of Residual Modelling under random assessment of candidate predictors and the robustness of the other three variable selection approaches under the complete removal of the effects of the known predictors.

3.1. Real-World Application

The lung and throat cancer (TLC) deaths per 100,000 population were obtained for 81 provinces of Turkiye for Year 2019. Ever smoking, ever alcohol consumption, and ever exposure to second-hand smoking at home were obtained as the known factors for the LTC deaths. We wish to investigate if the yearly variation (using the coefficient of variation-CV) of humidity, air pressure, wind speed, SO2 and Particulate Matter-10 are associated with the LTC deaths controlling for the known effects of smoking and alcohol. As a result, all for approaches identified the CV of humidity and windspeed as significant new predictors of the LTC deaths, suggesting that provinces with higher CV of humidity and windspeed reported relatively lower LTC deaths (Supplementary Figure A4).

4. Discussion

We have shown through simulations that the Residual Modelling performs better as the number of predictors increase while Full Modelling seems slightly more statistically powerful with small number of predictors.

In this investigation, we aimed to address a challenge in variable selection and model interpretation. Rather than focusing on statement like "Predictor $\boldsymbol{X_1}$ is significantly associated with the response variable $\boldsymbol{Y}$ after controlling the effect of $\boldsymbol{X_2}$", we feel that a statement like "Predictor $\boldsymbol{X_1}$ is significantly associated with the response variable $\boldsymbol{Y}$ after removing the effect of $\boldsymbol{X_2}$" is better and more interpretable. For example, when studying lung cancer mortality, smoking is a known factor. When both smoking and other candidate factors are entered into the model, smoking may lose its significance just due to some intrinsic association of smoking with other factors and the research may conclude "Smoking was not found to be significantly associated with lung cancer mortality in the presence of such and such predictors." Residual Modelling allows us to remove the effects of all known factors before we investigate the effects of other potential factors with the following question in mind: "Do these additional factors have any significant effect on $\boldsymbol{Y}$ beyond the known effects of smoking, etc."

Another advantage of Residual Modelling is that it does not limit itself with the number of predictors at hand. For example, Residual Modelling approach can be applied to gene expression studies where the number of predictors far exceeds the number of patients or samples at hand, utilizing the full power of the entire sample size for each potential predictors, while Full Modelling requires that the sample size be much higher than the number of predictors. Here, the effects of the known factors are removed from the response variable and then the effect of each gene can be investigated on the residuals. Familywise error control measures can be taken on their p-values as usual.

In model building, the subject level expertise supported by a comprehensive literature search is indispensable to identify the known predictors of the response variable at hand. In Residual Modelling, this becomes much more critical as the effects of these known predictors are removed from the response variable literally before investigating the candidate predictors. This requires extra attention especially when the predictors have inherent correlations among them. Naturally, when we deal with a group of highly correlated predictors, when one is considered as a known factor and its effect on the response variable is removed, the other candidate predictors are highly likely to be found as insignificant. This is exactly where the subject knowledge is needed.

The user of Residual Modelling is free to combine the residual modelling approach with other variable selection approaches. For example, once the known effects are removed from the response variable, a stepwise variable selection can be carried out on the residual of this model. As the order of the candidate variables may be important especially when these predictors have a certain level of correlation among them, the user can carry out sensitivity analyses by shuffling the candidate variable list. Under the multicollinearity assumption, this would not matter; however, in real-life studies, we see at least mild level correlations among potential predictors more often than not.

Residual Modelling approach is also prone to similar weaknesses as in the other variable selection approaches such as overfitting (especially with Backward selection approach), existing multicolieariy among predictors, inflated Type-1 Error due to multiple testing. Therefore, as a future research, we plan to study who the regularisation techniques such as Lasso, Ridge, and Elastic Net are can be used as penalization and shrinkage approaches to improve the model performance of Residual Modelling [9-13].

Residual Modelling is proposed for continuous response variables specifically where we can obtain residuals that are continuous variables in nature as well. Therefore, in its current state, Residual Modelling is not applicable to other types of responses such as binary, count, time-to-event, etc.

5. Conclusions

We have shown that Residual Modelling has desirable power and type-1 error rate properties in model building when there are known predictors (KPs) we need to account for in testing the significance of a new set of candidate predictors.

Author Contributions: Conceptualization, ANK and MK.; methodology, MK.; software, MFD and MK.; validation, MFD and MK.; formal analysis, MK.; investigation, MK.; resources, MK.; data curation, MK.; writing—original draft preparation, ANK, MFD, and MK.; writing—review and editing, ANK, MFD, and MK.; visualization, MK.; supervision, MK.; project administration, MK.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We have no additional acknowledgements beyond the author contributions to be declared.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] W. Sauerbrei, A. Perperoglou, M. Schmid, M. Abrahamowicz, H. Becher, H. Binder, D. Dunkler, F. E. Harrell, P. Royston, and G. Heinze, "State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues," Diagnostic and Prognostic Research, vol. 4, pp. 1–8, Dec. 2020. doi: 10.1186/s41512-020-00074-3.

[2] M. Z. Chowdhury and T. C. Turin, "Variable selection strategies and its importance in clinical prediction modelling," Family Medicine and Community Health, vol. 8, no. 1, 2020. doi: 10.1136/fmch-2019-000262.

[3] G. Claeskens, "Statistical model choice," Annual Review of Statistics and Its Application, vol. 3, pp. 233–256, Jun. 2016. doi: 10.1146/annurev-statistics-041715-033659.

[4] G. Heinze, C. Wallisch, and D. Dunkler, "Variable selection–a review and recommendations for the practicing statistician," Biometrical Journal, vol. 60, no. 3, pp. 431–449, May 2018. doi: 10.1002/bimj.201700067.

[5] Y. Wang, Q. Chen, N. Zhang, and Y. Wang, "Conditional residual modeling for probabilistic load forecasting," IEEE Transactions on Power Systems, vol. 33, no. 6, pp. 7327–7330, Aug. 2018. doi: 10.1109/TPWRS.2018.2819668.

[6] P. S. de Mattos Neto, G. D. Cavalcanti, D. S. O. Santos Júnior, and E. G. Silva, "Hybrid systems using residual modeling for sea surface temperature forecasting," Scientific Reports, vol. 12, no. 1, p. 487, Jan. 2022. doi: 10.1038/s41598-021-04342-8.

[7] M. Goodwin, "Residual modeling in music analysis-synthesis," in 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, May 1996, vol. 2, pp. 1005–1008. IEEE. doi: 10.1109/ICASSP.1996.543310.

[8] X. Weng, Y. Li, L. Chi, and Y. Mu, "High-capacity convolutional video steganography with temporal residual modeling," in Proceedings of the 2019 on International Conference on Multimedia Retrieval, Jun. 2019, pp. 87–95. doi: 10.1145/3323873.3325047.

[9] R. Tibshirani, "Regression shrinkage and selection via the Lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[10] A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," Technometrics, vol. 12, no. 1, pp. 55–67, Feb. 1970. doi: 10.1080/00401706.1970.10488634.

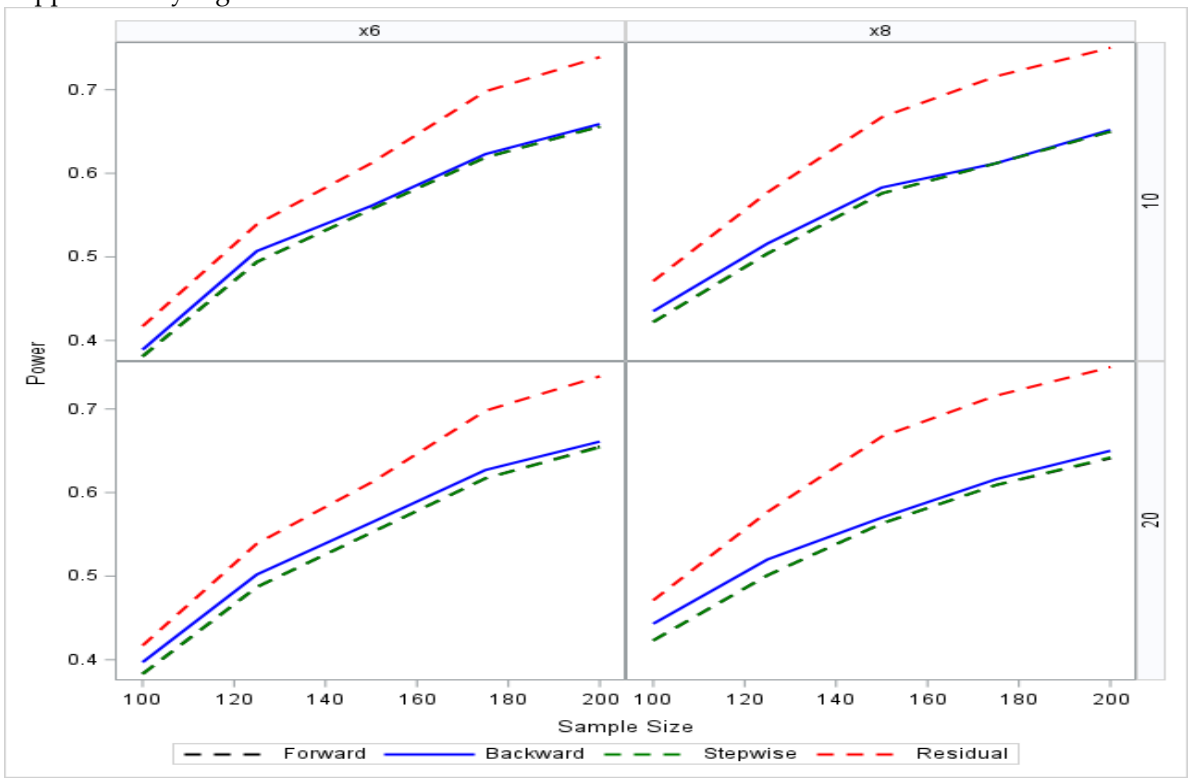
[11] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 2, pp. 301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

[12] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," Journal of Statistical Software, vol. 33, no. 1, pp. 1–22, 2010. doi: 10.18637/jss.v033.i01.

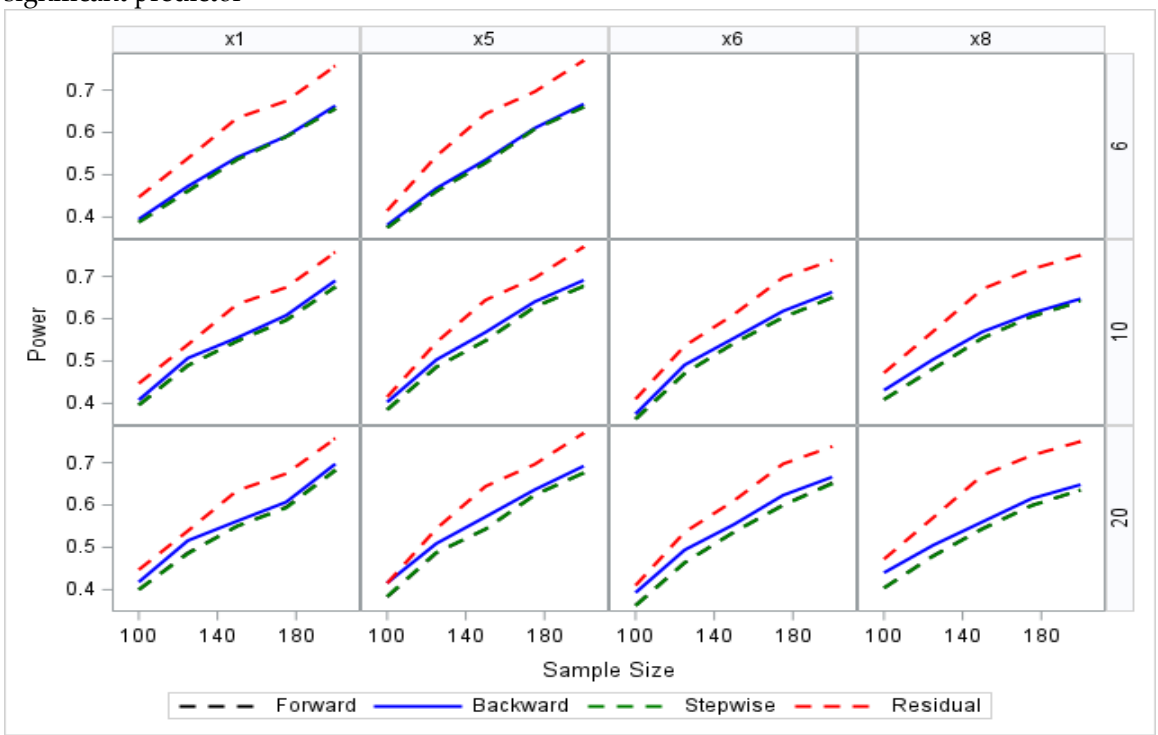
[13] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R. New York: Springer, 2013, ch. 6. doi: 10.1007/978-1-4614-7138-7.

Appendix A

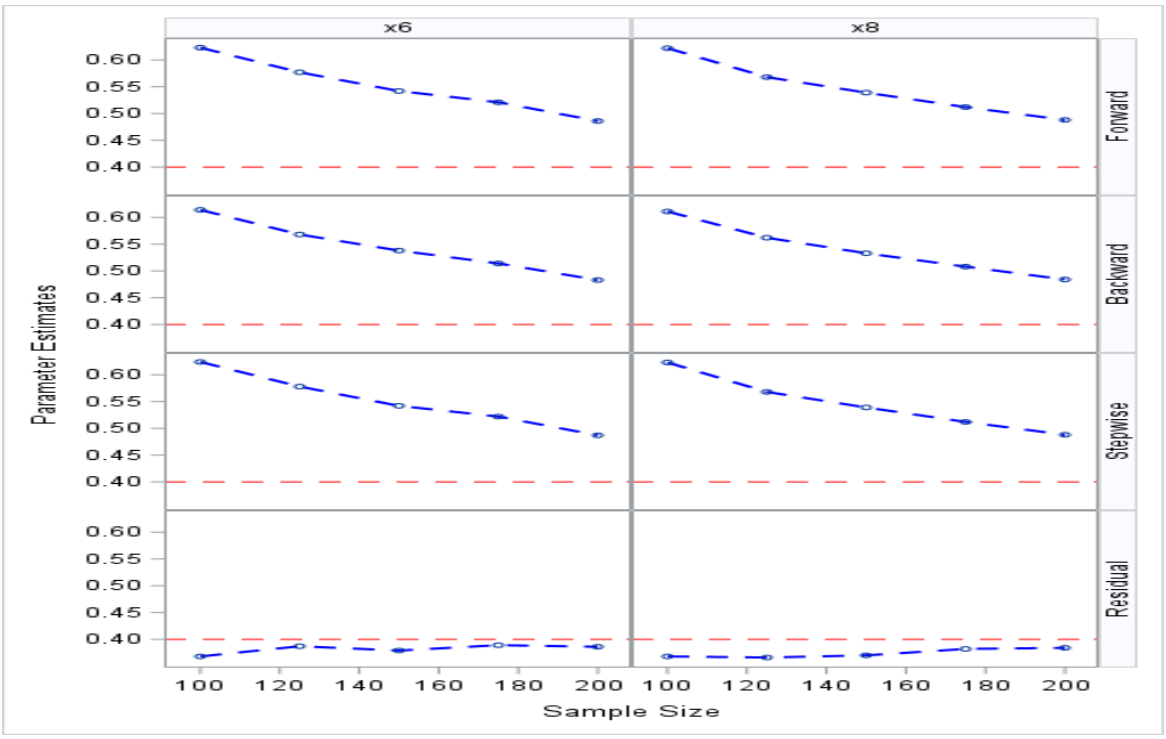
Supplementary Figures



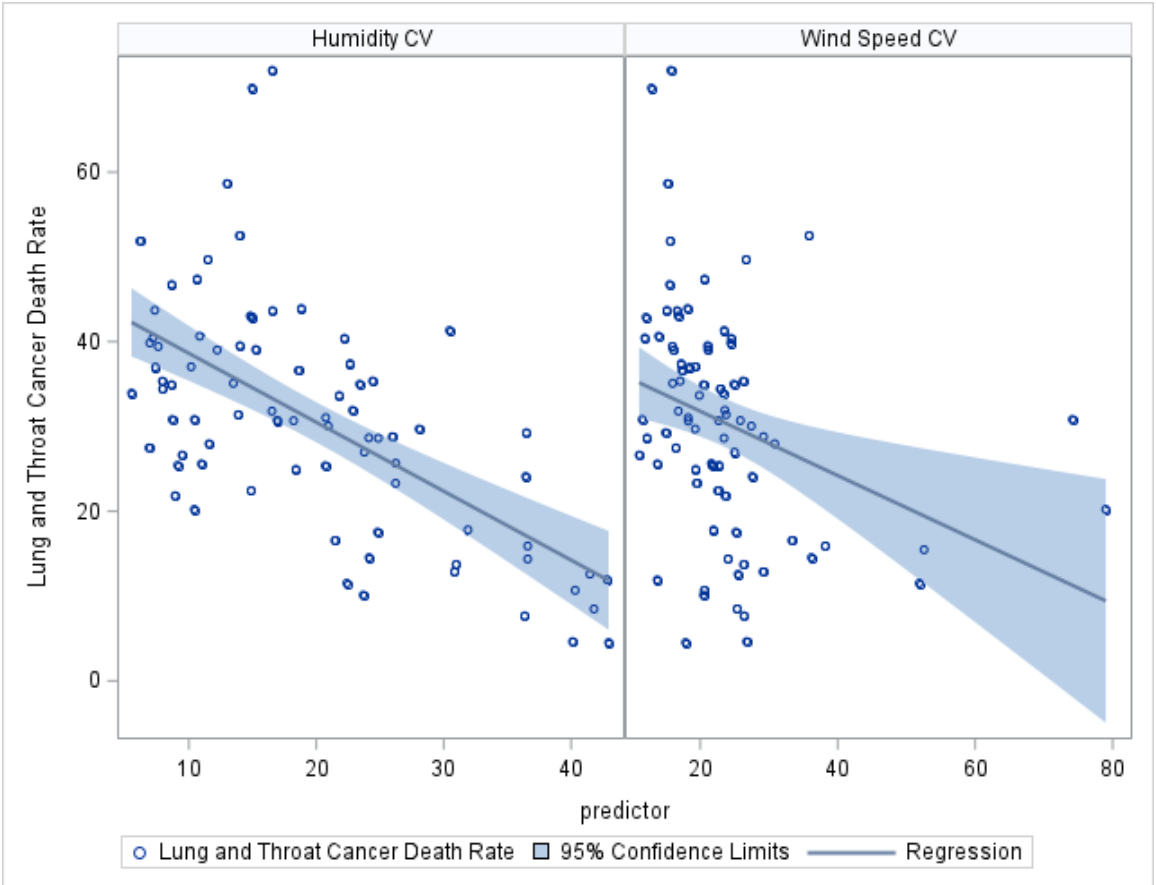
Supplementary Figure A1: Empirical Power when X1, X5, and X10 are considered to be the known significant predictor



Supplementary Figure A2: Empirical Power when X10 is considered to be the known significant predictor



Supplementary Figure A3: Parameter Estimates for the binary significant predictors by Variable Selection methods



Supplementary Figure A4: Association of Humidity and Wind Speed Coefficient of Variation (CV) with Lung and Throat Cancer Deaths per 100,000 population