STOR 390 Midterm Paper

Sofia Zhang

2024-10-25

The paper "The use of Logic regression in epidemiologic studies to investigate multiple binary exposures: an example of occupation history and amyotrophic lateral sclerosis" discussed the statistical method called logistic regression to investigate the association between occupational history and the risk of amyotrophic lateral sclerosis (ALS). Logistic regression has been an effective way to evaluate the risks of binary exposures to binary outcomes in epidemiological studies. However, when individuals are exposed to large dimensional binary exposures continuously and simultaneously , machine learning methods to deal with high-dimensional interactions have been developed. In this study context about Denmark population occupation histories, the mass of different occupations(68 primary occupation classifications and 746 sub-classifications among study participants) are interacting which are less effective to use the traditional independent logistic models. Therefore, the authors integrate simulated annealing to optimize the selection of exposure combinations, with presenting the interacting binary exposures as logic trees. I will examine how the stochastic process of simulated annealing to select significant variables introduced bias by choosing suboptimal solutions and overfitting. After that, I will discuss how biased conclusions generated by this algorithm may lead to the underrepresentation or misrepresentation of certain populations.

Methods Review

By preliminary knowledge of sex-specific differences of occupations and risk of ALS, the authors subset the data in four datasets: male with primary occupations, female with primary occupations, male with sub-classifications of occupations, female with sub-classifications of occupations. All the following model fitting and selection are constructed independently in the four sets.

Logic regression constructs Boolean combinations of exposures to determine significant predictors. In the context of this paper, the outcome variable(dependent variable) is a binary variable that 0 denotes not an ALS patient and 1 denotes an ALS patient. The exposure variables are 68 primary occupation classifications and 746 sub-classifications among study participants, where 0 denotes not one particular occupation and 1 denotes one particular occupation. It also includes newly constructed variables that indicate intersection and unions for some occupations by boolean operators. For example, there could be combinations like (((Textile and Clothing or (Paper or printing industry)) or (laundries or Others)) or ((Professional services or (CConstruction)) or (CWelfare))). Those Boolean expressions generated are selected by decision trees algorithms. However, when there are many predictors to consider, the number of possible combinations becomes so large that it's impractical to evaluate every potential tree. To tackle this challenge, the authors integrate simulated annealing to find an optimal model. Simulated annealing is an optimization technique inspired by heating and slowly cooling metal to remove imperfections. In simple terms, it helps find a good approximation of the best solution by iteratively making small changes to the decision trees—like altering one of the predictors or changing the operator between them. The method allows occasional acceptance of worse solutions to escape local optima, enabling a more thorough exploration of all possible trees and helping to identify a highly effective model without evaluating every single option. After multiple randomized iterations of this algorithm, the authors chose AUC(area under the curve) as the scoring function to evaluate model accuracy.

The authors first defined the binary exposure matrix, where one specific occupation is one column in the matrix that is ready to construct the logic trees. Then, after adjustments, the authors defined the parameters

required for the simulated annealing process: the acceptance rate in case of higher scores and the number of iterations. The fitted model has such a structure: The two important factors in the logistical model in this paper are model size and the total number of logic trees. Model size is the number of predictors, which in this paper's context are classifications of occupations. For example, A logic model with a size of 10 and 2 trees can be expressed as E[Y]=beta0 + beta1 * L1 + beta2 * L2, where L1 and L2 are 2 trees that represent combinations of up to 10 predictors in total (e.g., 7+3, 4+6, etc.). After iterations, the number of model sizes and tree numbers were moving, and the score function had lower and lower values . At the early stages all the moves are acceptable and the number of rejected moves gradually increases as the annealing chain progresses. They looked for a lower score(AUC in this paper), and if the scores are similar the authors chose the model with the lowest score and the least complexity. The final model in this paper has 3 trees and 20 predictors for primary classifications.

Because the model selection is highly based on the iterations algorithm, the result is sensitive to the noise as every run for the same parameters generates different trees. Therefore, the authors used 10-fold training/test set cross-validation to determine the size and the number of trees to achieve the best predictability. Researchers shuffle the dataset randomly and split it into 10 groups. For each unique group, they take the first group as a holdout or test dataset and use the remaining groups as a training dataset. A model is fitted on the training set and evaluated on the test set, retaining the evaluation score and discarding the model after each iteration. By summarizing the performance of the model using the sample of evaluation scores from all 10 iterations, researchers can assess the skill and generalizability of the model.. After this process, the authors found that larger model sizes generally perform better on the training set but introduce substantial noise in the test set, resulting in reduced predictive capability. Specifically, models with 3 trees have consistently high scores on the test dataset, indicating lower generalization. To address this, reducing model complexity is recommended, as a simpler model with 11 predictors and 2 trees achieves a better balance by maintaining a low training score and an acceptable test score. Finally, they figured out some consistently selected occupations, then manually added these occupations into classic logistic regression as independent predictors, thus getting odds ratio estimates to finally determine the influence of occupations on the risk of ALS.

Normative Concerns

The authors highlight several significant problems and limitations that warrant consideration when using simulated annealing in logic regression. They found that estimations are extremely sensitive to noise in the data, leading to inconsistent results in some analyses. Since the integration of simulated annealing and logistic regression is an iterative procedure that cannot feasibly search all possible combinations of predictors due to computational constraints, it inherently includes a random component introduced by simulated annealing. This stochastic nature can lead to model bias, as the algorithm might settle on suboptimal solutions or be influenced by initial conditions and hyperparameters, causing systematic errors in the model's predictions.

Previous applications of logic regression in genetics studies, as well as original publications, did not emphasize this issue. In genetic analyses, a few strong predictors or combinations are often identified, making the random component less problematic. However, the authors note that in many epidemiological analyses, there are a large number of predictors or combinations with weak associations to the outcome (e.g., odds ratios around 1.05 to 1.10). For example, in their analysis of 68 different occupation categories, dozens of Boolean combinations had similar weak associations with the outcome. The use of simulated annealing means that only a few combinations are selected from a larger pool of equally plausible options, potentially introducing bias by overlooking other relevant predictor combinations.

From a practical standpoint, the authors suggest that poor replicability of results may indicate the absence of strong predictors and increase the probability that selected exposures are due to chance. This issue becomes even more challenging when hundreds of covariates are of interest, as in their evaluation of over 700 specific occupations. They even can't find any significant conclusions from the 700 sub-classifications. They emphasize that due to the stochastic nature of simulated annealing, results from logic regression are not always reproducible, yet can still provide informative insights. The lack of computational tools like seed setting to replicate exact results further complicates the issue of model bias.

To compensate for the randomized nature and increase the accuracy, the authors use the cross-validation technique by training/test sets. However, the researchers acknowledged that they have to balance the accuracy and predictability of the whole model and the noise in the test datasets. Since this balance is not proved to be perfect in this paper, it is highly possible that the model fits the training data very well but performs poorly on new data. This overfitting can be a form of bias, as the model captures noise instead of the underlying trend. These biases of simulated annealing in the context of occupations and amyotrophic lateral sclerosis (ALS) raise important societal and ethical considerations. Model bias introduced through the stochastic processes of simulated annealing can lead to inconsistent identification of occupational risk factors for ALS. This inconsistency may result in misleading conclusions about which occupations are associated with higher risks, potentially impacting workers' livelihoods and well-being.

If certain occupations are incorrectly identified as high-risk due to biases in the model, and this study works as a reference to the policy-making, workers in those fields might face unwarranted stigma, job insecurity, or stress. This could influence employment opportunities and insurance premiums, disproportionately affecting individuals in those occupations. Conversely, genuine occupational hazards might be overlooked if the model fails to identify them due to bias, leaving workers uninformed about real risks and employers unaccountable for necessary safety measures.

Moreover, biases in the model can exacerbate existing societal inequalities. If the data used in the analysis is biased—perhaps underrepresenting certain demographic groups or occupations—the resulting model may fail to identify risks pertinent to marginalized populations, perpetuating health disparities. For instance, occupations predominantly held by minority groups might not be adequately studied, leading to a lack of protective policies for those workers. In the context of ALS, a disease with significant physical, emotional, and financial burdens, accurately identifying occupational risk factors is crucial for prevention and early intervention strategies. Therefore, Researchers have a moral responsibility to first ensure data quality and representativeness, and collect and use data that accurately reflect the diversity of the workforce, including all relevant occupations and demographic groups. Based on reliable data, researchers should keep developing complementary methods to mitigate model bias, such as other validation techniques or alternative optimization algorithms. After finding results, they should communicate the limitations and uncertainties associated with the model results, and then collaborate with occupational groups, healthcare providers, and policymakers to interpret findings in a way that considers the potential social impact.

As a conclusion, the pressing normative concern is the inability of simulated annealing to always select best occupation history predictors to predict the risk of ALS, which will lead to the moral dilemma to undercover risky populations or falsely identify populations in high-risk but actually not. It is essential to balance the technical advantages of methods like simulated annealing with a commitment to ethical responsibility, particularly when the outcomes have far-reaching implications for individuals' health and societal well-being.

Reference

Bellavia A, Rotem RS, Dickerson AS, Hansen J, Gredal O, Weisskopf MG. The use of Logic regression in epidemiologic studies to investigate multiple binary exposures: an example of occupation history and amyotrophic lateral sclerosis. Epidemiol Methods. 2020;9(1):20190032. doi:10.1515/em-2019-0032