Ethical Considerations in Using Logic Regression to Evaluate Multiple Binary Exposures: Addressing Overfitting

Sofia Zhang

Dec 13, 2024

Introduction

High dimensional data refers to datasets with a large number of variables (features) relative to the number of observations. In epidemiology, occupational classifications are considered high dimensional data because they involve numerous binary variables representing various job categories or exposures. While traditional high dimensional datasets often contain thousands of variables, occupational data can still pose similar challenges due to the large number of predictors relative to the sample size, the sparsity of certain occupations, and the complexity of interactions between exposures and health outcomes. These characteristics create issues such as overfitting and multicollinearity, making occupational datasets effectively high dimensional despite having fewer variables. The paper “The use of Logic regression in epidemiologic studies to investigate multiple binary exposures: an example of occupation history and amyotrophic lateral sclerosis” from Bellavia et al.applied machine learning techniques, specifically logic regression and decision trees, to analyze high dimensional data. They demonstrated this approach by evaluating the association between occupational history and the risk of developing amyotrophic lateral sclerosis (ALS) as a practical example.

This paper evaluates the validity of this method in identifying occupations associated with increased risks of certain diseases by applying it to a different dataset, and argues researchers should be careful in overfitting before using this method to real public health research, by a deontological framework.

Methods Analysis

To verify the ability of this method to identify risky occupation history, I applied the method to the 2013 and 2014 National Health And Nutrition Survey(NHANES) dataset. Since most datasets containing ALS information are of credential and subject to confidentiality, I used the publicly available NHANES dataset and evaluated associations between occupations and stroke history. This dataset had 23 occupation classifications and 3436 samples. The smaller sample size and fewer binary occupation variables may limit the statistical power and validity of this evaluation. However, they also provide an opportunity to assess the generalizability of this method to datasets with limited sample sizes, which are common in occupational epidemiology studies.

Prior to the analysis, the dataset was prepared to closely replicate the structure and characteristics of the analysis dataset described in the paper. From the raw dataset, observations with missing or unknown information on occupations and stroke histories were cleaned. Each occupation constructs one binary variable, which can be viewed as boolean expressions. For example, if one participant self reported their previous occupations including business financial and management, then the business boolean variable and management boolean variable equals to 1 while other variables equal to 0. The stroke histories are formatted as 0 1 binary variables as well. As Bellavia et al. states, the dataset should be divided into sex specific subsets. The following analysis is stratified of binary sex.

As Bellavia et al. introduced, Logic regression is a method used to estimate Boolean expressions that minimize a defined score function, such as the binomial deviance for logistic regression or least squares for linear regression. This approach integrates parameter estimation and the search for Boolean expressions simultaneously, often utilizing decision trees to explore combinations of predictors. Given the large number of potential predictor combinations in high dimensional data, logic regression employs simulated annealing to iteratively modify and evaluate trees. Moves that improve the score are accepted, while those that do not may still be accepted early in the process based on a temperature parameter, which decreases over time to limit acceptance of less optimal moves. This method ensures efficient exploration of the search space and has proven more effective than alternative algorithms.

A simulated annealing algorithm suited for the simulation dataset is defined in this novel analysis. After adjustments, the best setting for the temperature parameter limits was 2 to –4, and the number of iterations was 10000. Since the women(n=1546) and men(n=1890) were balanced, I used this searching system for both datasets.

For women, I simultaneously estimated several models by varying the tree number from 1 to 3 and the number of predictors from 1 to 12 and seeked the lowest score (typically representing the better model), implemented by the LogicReg package in R. Take inference from figure1 as what the paper mentioned, I chose the model with 11 predictors and 3 trees as an optimal one. As the paper presented, 3 different runs of the same Logic model with 11 predictors in 3 trees showed different occupations combinations, where I'm not able to consistently identify combinations of higher importance.

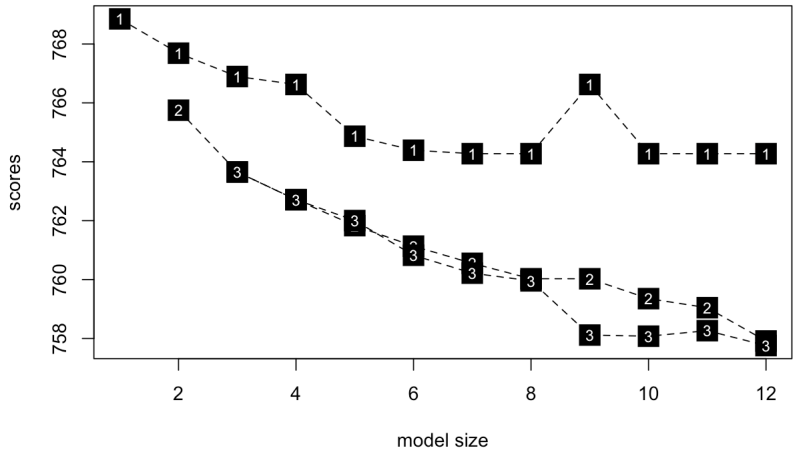


Figure1: Model scores for initial searching, 1 12 predictors and 1 3 trees

As what the paper suggests, a 10 fold cross validation was implemented for the model to evaluate the best model size. This effectively pruning the final model to 10 predictors and 2 trees. While the model with 11 predictors across 3 trees performs well in the training set, it exhibits one of the highest scores in the test dataset. This poor predictive performance could account for the inconsistencies observed in the previous section. Therefore, I chose a less complex model that still has a low score in the training dataset.

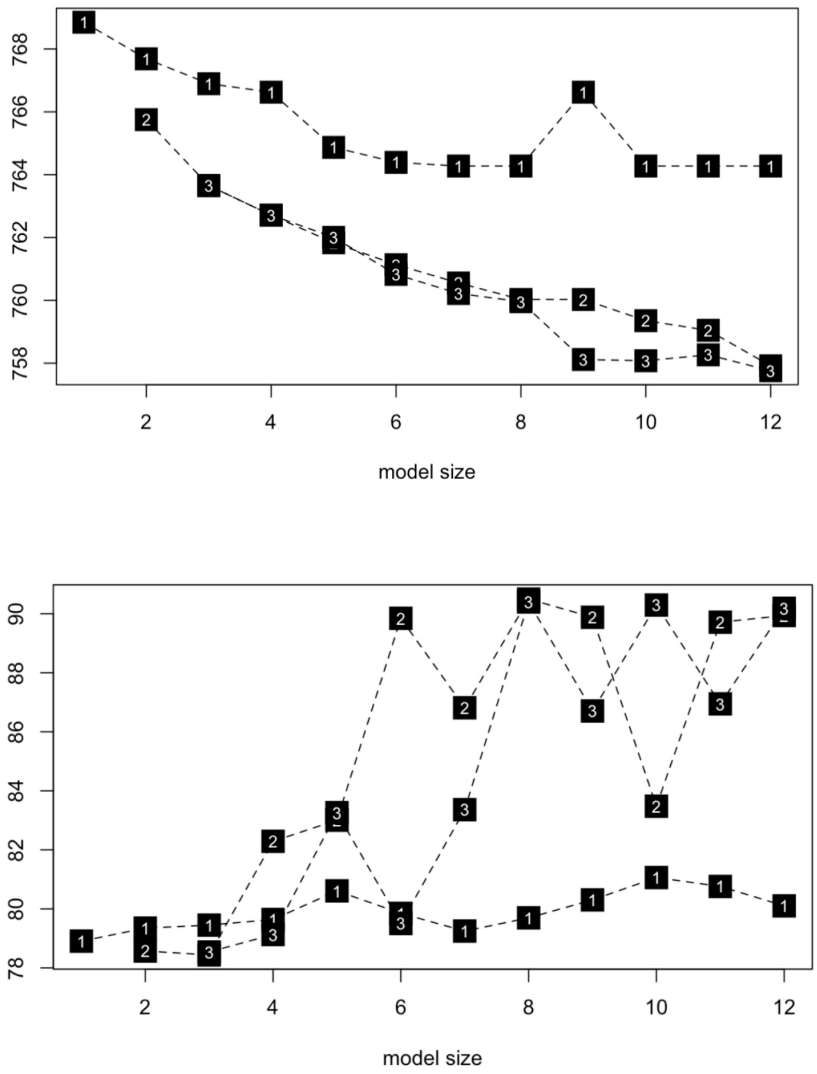
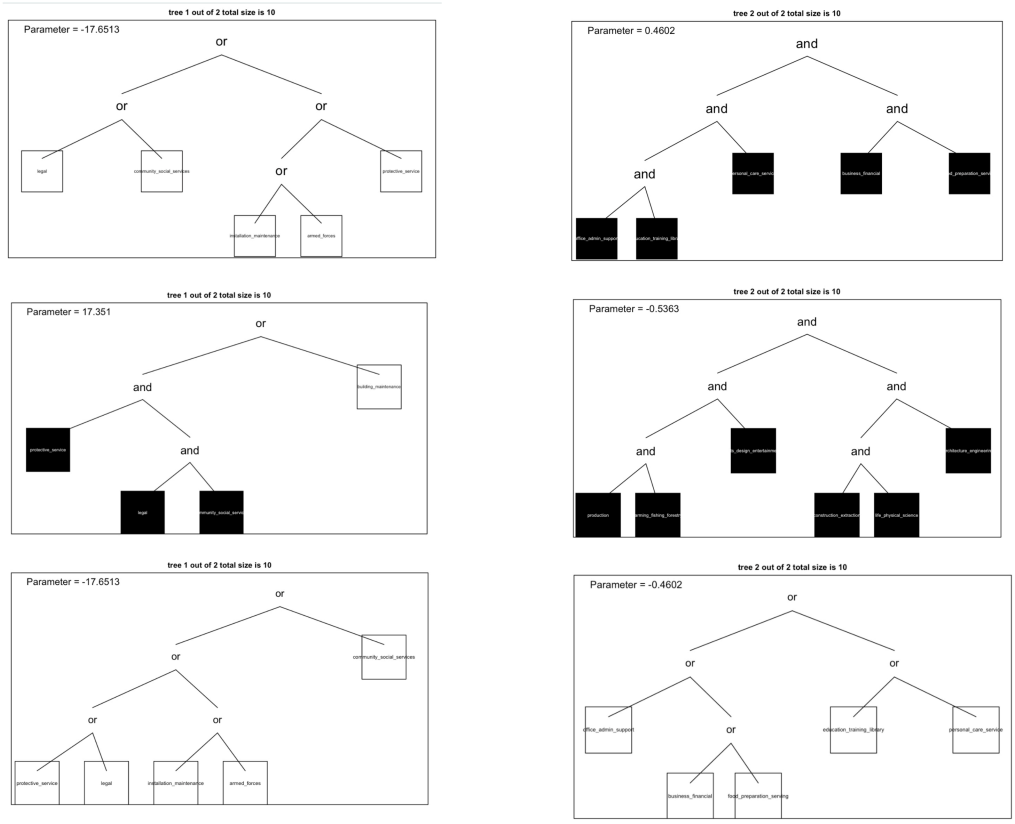


Figure2: Scores of Logic regression models of different sizes in the training (above) and test data (below) for women. The x axis represents the total size of the model (the maximum number of included predictors), while the number of trees is indicated inside the squares

Results from the pruned model provide much more consistent patterns, with several job categories being consistently selected (e.g. Protective Service, Legal, and Community Social Services). Table1 and figure 3 presents the selected occupations and their combinations. By incorporating the third model into classic logistic regression, only the L2 combination in model 3(shown in table1) were tested as significantly lower risk(p=0.03). However, those occupations were not consistently mentioned by repeated runs of our logic model. This inconsistency can be due to many reasons not limited to invalidity of the method. For example, there can be no detectable association between occupation history and stroke history for adult women or additional adjustments should be made.



Pruned Logic Trees with 10 predictors and 2 trees from 3 different runs

Score	β0	β1	β2	L1	L2
758.14	-3.01	-17.7	0.46	((Legal ∨ Community Social Services) ∨ ((Installation Maintenance ∨ Armed Forces) ∨ Protective Service)))	(((¬Office Admin Support ∧ ¬Education Training Library) ∧ ¬Personal Care Service) ∧ (¬Business Financial ∧ ¬Food Preparation Serving)))
760.028	-19.7	17.4	-0.536	((¬Protective Service ∧ (¬Legal ∧ ¬Community Social Services)) ∨ Building Maintenance))	(((¬Production ∧ ¬Farming Fishing Forestry) ∧ ¬Arts Design Entertainment) ∧ ((¬Construction Extraction ∧ ¬Life Physical Science) ∧ ¬Architecture Engineering)))
758.14	-2.55	-17.7	-0.46	((Protective Service ∨ Legal) ∨ (Installation Maintenance ∨ Armed Forces) ∨ Community Social Services))	((Office Admin Support ∨ (Business Financial ∨ Food Preparation Serving)) ∨ (Education Training Library ∨ Personal Care Service)))

Table1: Results of 3 different runs from trees with 10 predictors and 2 trees

For men, using the same simulated annealing algorithm and same initial estimation by varying the tree number from 1 to 3 and the number of predictors from 1 to 12, there are the same unidentifiable results. 10 fold cross validation pruned the tree to 11 predictors and 2 trees, which has relatively low scores in both training and testing datasets shown in figure 4.

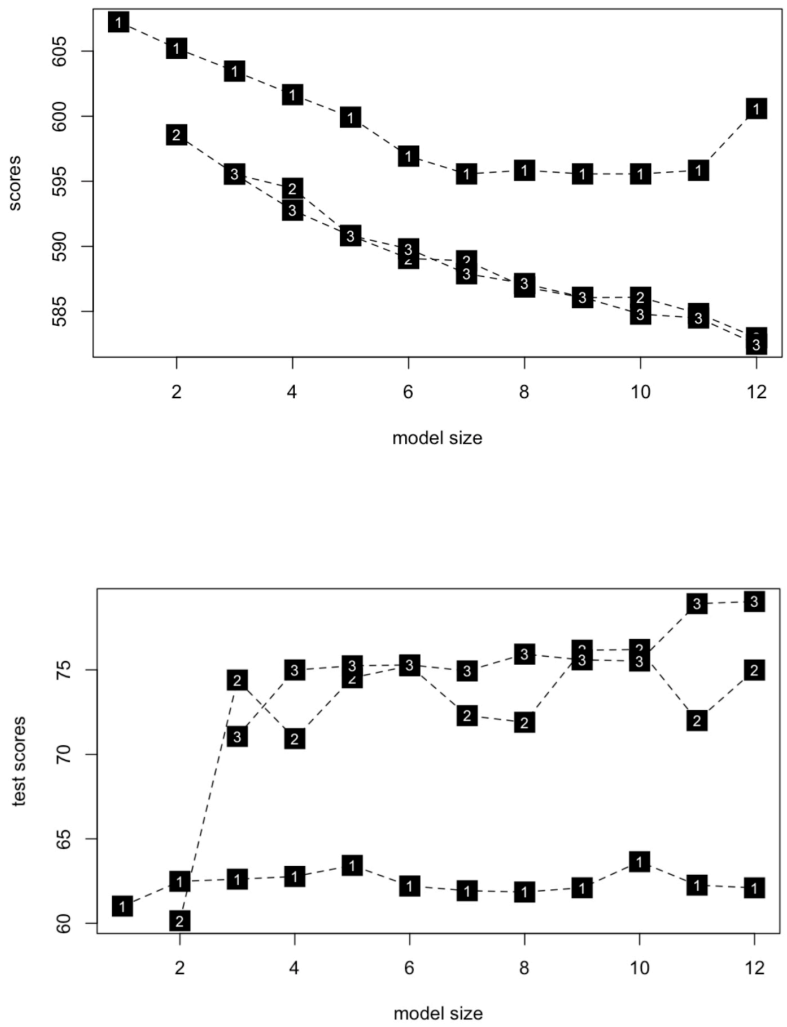


Figure3: Scores of Logic regression models of different sizes in the training (above) and test data (below) for men. The x axis represents the total size of the model (the maximum number of included predictors), while the number of trees is indicated inside the squares

Classic logistic regression models showed that several occupations were consistently identified(installation maintenance, architecture engineering, production, computer mathematica), the L2 combination in model 3 (Table2) were tested as significantly higher risk(p=0.001). Installation maintenance and production in this combination were identified as significantly increasing the risk of stroke by multivariate logistic regression (Table 3). This method successfully detects significant occupation history and their combination of higher risk.

Score	β0	β1	β2	L1	L2
585.546	-3.08	-18.1	0.773	((Education Training Library ∨ Arts Design Entertainment) ∨ (Food Preparation Serving ∨ Life Physical Science)) ∧ ¬Management)	((Computer Mathematical ∨ Installation Maintenance) ∨ (Architecture Engineering ∨ Legal)) ∨ (Armed Forces ∨ Production))
584.896	-21.2	18.2	0.679	(((¬Life Physical Science ∧ ¬Food Preparation Serving) ∧ (¬Personal Care Service ∧ ¬Community Social Services)) ∧ (¬Education Training Library ∧ ¬Arts Design Entertainment)))	((Installation Maintenance ∨ Architecture Engineering) ∨ (Production ∨ Computer Mathematical)) ∧ ¬Food Preparation Serving)
584.09	-21.2	18.1	0.755	(((¬Food Preparation Serving ∧ ¬Personal Care Service) ∧ ¬Education Training Library) ∧ (¬Arts Design Entertainment ∧ ¬Life Physical Science)))	((Production ∨ Installation Maintenance) ∨ (Armed Forces ∨ Architecture Engineering)) ∨ (Legal ∨ Computer Mathematical))

Table2: Results of 3 different runs from trees with 10 predictors and 2 trees

Table3: Results from logistic regression models based on Logic models results

Occupation	log.OR.	OR	P.value
Installation Maintenance	3.04	1.11	0.001
Production	1.87	0.63	0.128
Architecture Engineering	1.97	0.68	0.038
Computer Mathematical	2.59	0.95	0.131

Analysis of Normative Concerns

Bellavia et al. acknowledged that they have to balance the accuracy and predictability of the whole model and the noise in the test datasets. The main purpose of the 10 fold cross validation was to reduce noise in the model, while it simultaneously deals with overfitting issues. Overfitting is a significant concern in logic tree modeling, as it can lead to results that perform well on training data but fail to generalize to unseen data. While the authors addressed overfitting by employing 10 fold cross validation to prune the trees and select a model with relatively low scores in both the training and testing datasets, this approach remains a relative measure. It appears that the pruning was aimed more at reducing noise to uncover consistent patterns rather than rigorously ensuring that the pruned model is no longer overfit. In my simulations, pruning resulted in minimal changes—only reducing one tree and one predictor—which raises doubts about the effectiveness of this technique in mitigating overfitting. Although my analysis in men detected results, it's still doubtful that the non significant results in the analysis of women were possibly due to uneliminated overfitting. The main issue is that the authors did not propose a method to thoroughly evaluate whether the pruned model avoids overfitting, leaving a gap that could mislead future researchers applying this method. Without careful examination, the supposedly consistent patterns identified by the model might actually be artifacts of overfitting, undermining the reliability of the findings.

Speaking on the research side, this issue mirrors a philosophical example often discussed in epistemology: the problem of induction. Just as drawing broad conclusions from limited or biased observations can lead to faulty generalizations, reliance on an inadequately pruned model risks identifying patterns that appear consistent but are not truly generalizable. In both cases, the failure to critically test the validity of the observed patterns or results leads to a risk of mistaking noise or coincidence for meaningful insights. For researchers, this highlights the importance of implementing robust methods, such as additional validation techniques or sensitivity analyses, to

ensure the reliability and generalizability of findings, particularly in high dimensional settings where overfitting is a pervasive challenge. When viewed through the lens of the Ends Not Means Maxim from Deontology, by only concerning the significant results but ignoring overfitting, the research process loses its commitment to treating participants as ends. The findings may provide convenient and publishable results, but at the expense of accuracy and fairness. This violates the fundamental ethical obligation to use rigorous and transparent methods that prioritize the well being of participants. From a deontological perspective, ensuring that participants' data is analyzed in a way that minimizes biases and prevents overfitting is critical to respecting their dignity. Occupational epidemiology research must prioritize the well being of the individuals being studied rather than merely instrumentalizing their data to produce results. Participants are not just sources of information for identifying statistical patterns; they are the very population whose health and safety the research should aim to protect. In this context, the act of identifying risky occupations must go beyond achieving statistically significant or policy relevant outcomes. Instead, it should directly address the needs and concerns of the workforce, ensuring that the findings are robust, equitable, and actionable in ways that genuinely enhance workers' quality of life.

For instance, if certain occupations are inaccurately identified as high risk due to biases in the model, this could lead to unwarranted stigma, job insecurity, or elevated insurance premiums for individuals in those fields. Such outcomes would not only misuse participants' data but also fail to respect their dignity, treating them as a means to an academic or policy oriented end rather than as individuals deserving of fairness and protection. Conversely, if genuine occupational risks are overlooked due to methodological shortcomings, the research fails to fulfill its duty to inform workers and employers about real hazards, leaving populations exposed to preventable harm.

By adhering to the Ends Not Means Maxim, researchers can ensure that their methods are rigorously designed to avoid overfitting, minimize bias, and prioritize transparency. This approach aligns the research process with the ethical obligation to treat participants as ends in themselves, making their well being the primary objective of both the study and its applications. Through such an ethical framework, occupational epidemiology can balance scientific rigor with moral responsibility, ensuring that its findings serve to protect and empower the populations it studies.

Conclusion

The paper "The use of Logic regression in epidemiologic studies to investigate multiple binary exposures: an example of occupation history and amyotrophic lateral sclerosis" from Bellavia et al. introduced logic trees to evaluate associations between multiple binary exposures and binary outcomes. They used associations between occupation history and ALS as an example, in order to identify single occupations and occupation combinations in one person's personal history that increase the risk of ALS. They verified results from this innovative machine learning method into a classical epidemiological approach(logistic regression) to generate meaningful epidemiological results(odds ratio). My analysis replicated their proposed methods into examining associations between occupational histories and stroke history from 2013 to 2014 NHANES dataset. Result in my analysis for adult males verifies that the method is effective in identifying occupations in risk. As there is a lack of modern methods to examine multiple binary exposures, logic trees can be viewed as a potentially promising tool.

However, it still needs further work to resolve the overfitting issues in this model, before implementing this method into real epidemiological research to generate any meaningful conclusion. This paper argues the importance of this carefulness by the Ends Not Means Maxim from Deontology. By focusing solely on reducing noise while neglecting overfitting, researchers risk treating participants as mere tools to achieve statistically significant results rather than respecting their dignity and well being. Overfitting can lead to misleading conclusions that either falsely label occupations as high risk, causing unnecessary harm, or fail to identify genuine hazards, leaving

workers unprotected. Adhering to rigorous methods that minimize bias and validate findings ensures that participants are treated as ends in themselves, aligning research with ethical principles and its ultimate goal of improving public health and workplace safety.

Reference

Bellavia, A., Rotem, R. S., Dickerson, A. S., Hansen, J., Gredal, O., & Weisskopf, M. G. (2020). The use of Logic regression in epidemiologic studies to investigate multiple binary exposures: an example of occupation history and amyotrophic lateral sclerosis. Epidemiologic methods, 9(1), 20190032.