Lifestyle and Wellness: Analyzing the Impact of Personal Habits on Health

Jiahao Gu, Fanke Qin, Pengfeng Yang, Zilong Zhao, Hilary Zou

December 2023

1 Contributions

Jiahao Gu	Coding of Model Transformation and Selection and relative parts in the result section
Hilary Zou	Method Section, Coding of Model Validation and relative parts in the result section
Fanke Qin	Coding of Model Inference, checking multicollinearity and relative parts in the result section
Pengfeng Yang	Introduction and Discussion Section, Coding of Model Building and relative parts in the result section
Zilong Zhao	Coding of Assumption Checking

2 Introduction

Personal habits can significantly influence health conditions. Habits such as regular physical activity and a balanced diet contribute to overall wellness and the prevention of chronic diseases. However, negative habits like smoking and excessive alcohol consumption are known risk factors for various health issues. Understanding the direct impact of personal habits on health conditions is essential, as this knowledge could lead to more effective strategies for promoting healthier lifestyles and preventing various diseases. We intend to conduct a thorough analysis of the following research question: How do personal habits, including the number of days per week the person engages in physical activities, whether he regularly performs physical activities (categorical variable), whether he consumed at least 12 alcoholic beverages in any one year (categorical variable), whether he is experiencing trouble sleeping (categorical variable), and whether he smoked at least 100 cigarettes in his lifetime (categorical variable) impact health conditions as measured by the self-reported number of days per month the participant's physical health was not good? Linear regression allows us to study the impact of multiple independent variables on a single dependent variable. Each predictor's coefficient quantifies the significance of the associated predictor, thereby providing insights into their contributions to the dependent variable.

The past research articles mainly focused on the impact of a specific habit on physical health. For example, there is a significant association between physical activity and health conditions. Any level of regular physical activity, compared to inactivity, reduces the probability of having diabetes, high blood pressure, arthritis, and reporting being in fair or poor health [1]; smoking can lead to many serious health issues such as cancer, coronary artery disease, and autoimmune disorders [2]; low to moderate consumption might be associated with reduced risks or benefits, but as consumption increases, the risks also increase [3]. Our research aims to provide a comprehensive analysis of the combined effects of multiple personal habits. We seek to provide a complete understanding of lifestyle factors and their relation to physical health conditions. This approach allows us to explore the interplay and relative impact of these habits in a broader context, offering an in-depth perspective on how lifestyle choices cumulatively affect health and well-being.

3 Methods

We split the dataset into training (three-fourths) and testing (one-fourth) subsets, then built a multiple linear regression model using the training set.

Then, we made sure that our model met the key assumptions, including linearity, homoskedasticity, and normality. We used two graphs to do this: a scatter plot of ”Residuals versus Fitted Values” and a Normal QQ Plot. For linearity, we looked at the scatter plot to see if the residuals showed any clear patterns, like curves to determine if our model was capturing the relationship properly. For homoskedasticity, using the scatter plot, we checked if the spread of residuals changed (like fanning out or in) as the predicted values increased or decreased. For Normality, with the Normal QQ Plot, we checked if the data followed a straight diagonal line. In the case of any assumptions being violated, we used the Boxcox method to find an appropriate transformation for the response or predictors and applied the transformation to our model.

After the transformation was applied and all assumptions were satisfied, we selected our model using three methods: Firstly, we used hypothesis tests and partial F test to check if there were any significant linear relationships existing between a subset of predictors and the response. We used the reduced model as a candidate for the final model. Secondly, we used all possible subsets selection method and used the model built by the resulting predictors as a candidate. Thirdly, we used step-wise automated selection methods and used the resulting model as a candidate. We then evaluated the models based on the assumptions, the coefficients of determination, AIC, BIC, and VIFs.

Subsequently, we identified problematic observations, including leverage observations, outlying points, and influential observations by comparing respective measures with certain cutoffs. By identifying these problematic observations, we could understand if individual observations affected our estimated relationship. Unless there was a contextual reason, we noted their presence and impact as a limitation of the model but did not remove them from the data set.

Proceeding, we made inferences on the final model. First, we used the ANOVA test for overall significance and checked if the p-value was small to determine if there was a statistically significant linear relationship existing for at least one predictor. Then, we applied the hypothesis test to each predictor and checked if the p-value was small to conclude if there linear relationship existed between each predictor and the response. Lastly, we calculated confidence intervals for all predictors, which indicated how each predictor influenced the response.

Finally, we validated the model by comparing it with one built with the testing dataset, checking for consistency in estimated coefficients, significance of predictors, adjusted coefficient of determination, and multicollinearity. Any substantial differences highlighted limitations in the model's performance.

4 Result

We used the NHANES data set, a survey data collected by the US National Center for Health Statistics which had conducted a series of health and nutrition surveys since the early 1960s, to build our model and answer our research question. Our response variable, DaysPhysHlthBad, reflects the self-reported number of days in the past month with poor physical health. The predictors include PhysActiveDays (weekly days of physical activity), PhysActive (regular physical activity), AlcoholDay (average daily alcohol consumption), Alcohol12PlusYr (annual consumption of 12 or more alcoholic beverages), SleepHrsNight (typical nightly sleep duration), SleepTrouble (presence of sleep difficulties), Smoke100n (history of smoking over 100 cigarettes), HardDrugs (use of substances like cocaine or heroin), and Marijuana (Marijuana use). These variables were selected because they were habits that had intuitive relationships with health conditions and we liked to investigate their relationships through data analysis. Additionally, we observed the scatter plots of each continuous predictor and the response, they showed that no data point was significantly distant from the others, suggesting the absence of obvious leverage points or outliers for model diagnosis purposes.

Variable Name	Variable Type	Min	Mean	Max	P-Value
DaysPhysHlthBad	Response	1.000	9.874	30.000	/
PhysActiveDays	Predictor	0.00	1.46	7.00	0.493389
AlcoholDay	Predictor	1.000	3.394	21.000	0.575245
SleepHrsNight	Predictor	2.000	6.733	12.000	0.876599

Table 1: Numerical Summaries of the Continuous Variables in the Data Set, including the Minimum, Mean, Maximum, and P-Value of the Hypothesis Test

Variable Name	Variable Type	YES	NO	P-Value
Smoke100n	Predictor	201	203	0.027155
HardDrugs	Predictor	99	305	0.260159
PhysActive	Predictor	165	239	0.007815
Marijuana	Predictor	245	159	0.666017
SleepTrouble	Predictor	127	277	0.000101
Alcohol12PlusYr	Predictor	364	40	0.559915

Table 2: Numerical Summaries of the Categorical Variables in the Data Set, including the Number of Observations in the Two Categories, and P-Value of the Hypothesis Test

After checking the assumptions, we discovered that the residual vs fitted plot displayed a random band of residuals with no obvious patterns, but there were stark deviations from the diagonal line in the QQ plot, suggesting a violation of the normality assumption. We applied the Boxcox method, and the resulting plot suggested a reasonable transformation to the response would be a logarithmic transform since the power that maximizes the likelihood is estimated to be about 0. After applying the logarithmic transform to the response, we found the residual vs fitted plot showed an approximately null plot, and the data in the QQ plot follows the diagonal line, indicating that all the assumptions are satisfied.

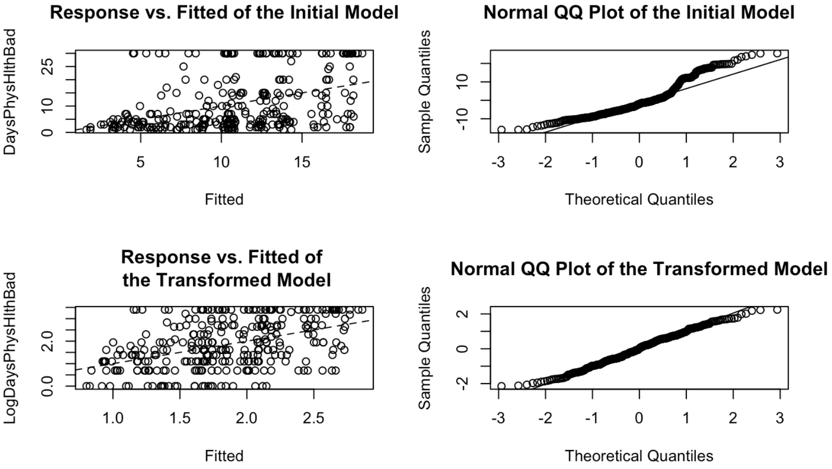

Figure 1: Response vs. Fitted and QQ Plot of Both the Initial Model and the Transformed Model

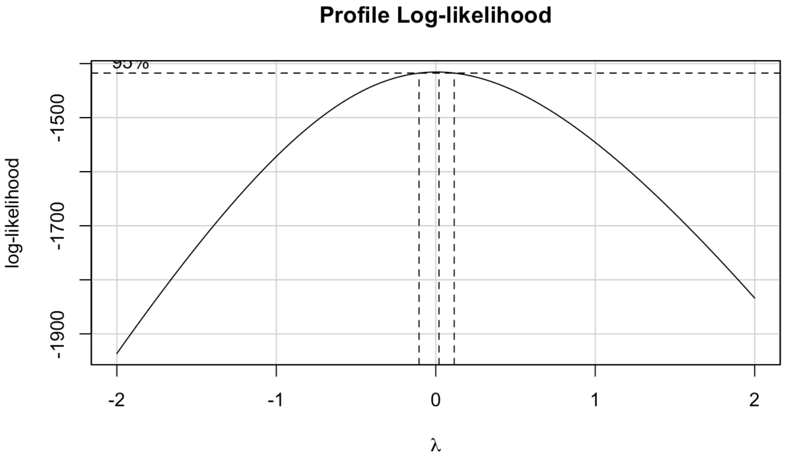

Figure 2: Result of the Boxcox Method

For the model selection process, by applying the hypothesis tests on all predictors, we derived four reduced models, as shown in the table below. Based on the result of the partial F tests, there did not exist significant linear relationships between the response and the removed predictors in all four models. Both the all possible subsets selection method and the step-wise automated selection methods agree on the first reduced model shown in the table below. By comparing the models, we found that all of them satisfied the assumptions. Although the first model had higher VIFs, indicating some multicollinearity present, it has the highest adjusted coefficient of determination, lowest AIC, and BIC, indicating the first model explained the most variation in the response, so we selected the first model, which includes the predictors PhysActiveDays, Smoke100n, PhysActive, SleepTrouble, and Alcohol12PlusYr, as our final model. Based on the residual vs. fitted plot, all assumptions were preserved. According to the VIFs, no severe multicollinearity is present.

Predictors	Adjusted R^2	AIC	BIC	Average VIF	Highest VIF	Violated Assumptions	P-Value
PhysActiveDays, Smoke100n, PhysActive, SleepTrouble, Alcohol12PlusYr	0.1773753	-5.601957	16.68044	1.9197	3.193651	None	0.5448
AlcoholDay, SleepHrsNight, Smoke100n, PhysActive, SleepTrouble	0.1621478	-0.04443291	22.23796	1.0778	1.149672	None	0.07579
SleepHrsNight, HardDrugs, Smoke100n, PhysActive, SleepTrouble	0.1650399	-1.092118	21.19028	1.0881	1.167152	None	0.1138
Marijuana, PhysActiveDays, Smoke100n, PhysActive, SleepTrouble	0.1647939	-1.002888	21.27951	1.9080	3.129368	None	0.1099

Table 3: The Comparison of Adjusted R^2, AIC, BIC, Average VIF, Highest VIF, and Violated Assumptions of the Four Reduced Models and Their Respective P-value in the Partial F Test

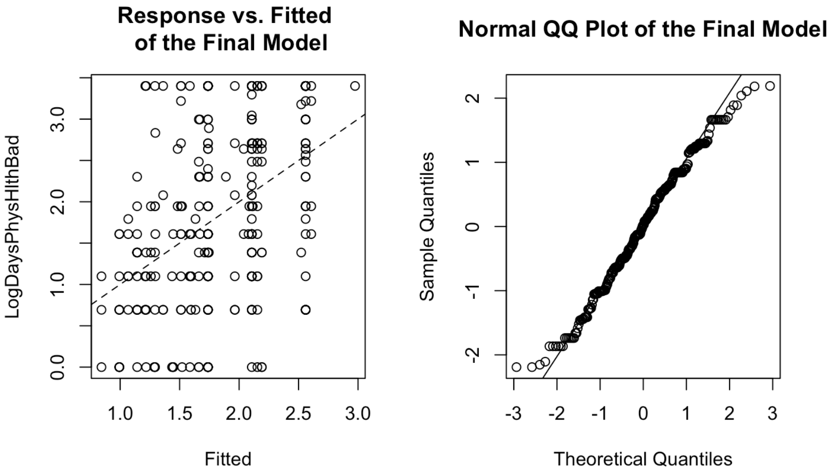

Figure 3: Response vs. Fitted and QQ plot of the Final Model

PhysActiveDays	Smoke100n	PhysActive	SleepTrouble	Alcohol12PlusYr
3.165778	1.116055	3.193651	1.040779	1.082211

Table 4: The VIFs of the Final Model

Conclusively, we applied the ANOVA test for overall significance and got a p-value of 1.765e-10, indicating a statistically significant linear relationship exists for at least one predictor. Then, we conducted hypothesis tests on all predictors and calculated confidence intervals for them, the result is displayed in the following table.

Predictor Name	Estimate	Confidence Interval	P-Value
(Intercept)	2.15368	(1.7904339, 2.51693410)	< 2e-16
DaysPhysHlthBad	-0.07498	(-0.1675808, 0.01762458)	0.112128
Smoke100n	0.36838	(0.1338552, 0.60290392)	0.002183
PhysActive	-0.37052	(-0.7752732, 0.03422634)	0.072628
SleepTrouble	0.45277	(0.2108570, 0.69468462)	0.000273
Alcohol12PlusYr	-0.41538	(-0.7961525, -0.03461392)	0.032612

Table 5: The 95% Confidence Intervals of the Predictors and Their Respective P-values

5 Discussion

Our final model is $DaysPhysHlthBad = 2.15368 - 0.07498 \times DaysPhysHlthBad + 0.36838 \times Smoke100n - 0.37052 \times PhysActive + 0.45277 \times SleepTrouble - 0.41538 \times Alcohol12PlusYr$, where Smoke100n, PhysActive, SleepTrouble, and Alcohol12PlusYr is 1 if it's YES, and 0 if it's NO. The model reveals that with all other predictors held constant, engaging in an additional day of physical activity corresponds to an average decrease of 0.07498 in the logarithm of the number of days spent in poor health. Additionally, individuals who have smoked at least 100 cigarettes exhibit an increase of 0.36838 in the logarithm of the average number of days with poor health compared to non-smokers. Regular physical activity is associated with a reduction of 0.37052 in the logarithm of the average number of days with poor health. Those experiencing sleeping troubles show an increase of 0.45277 in the logarithm of the average number of days with poor health relative to those without such issues. Lastly, individuals who have consumed at least

12 alcoholic beverages annually demonstrate a decrease of 0.41538 in the logarithm of the average number of days with poor health in comparison to those who have not. Based on the p-values, there is a significant linear relationship between Smoke100n, SleepTrouble, Alcohol12PlusYr, and DaysPhysHlthBad, but there may not be a significant linear relationship between PhysActiveDays and PhysActive, we can conclude that smoking and having trouble sleeping harm health conditions, consuming alcohol have a positive impact on health conditions, and the impact of regular physical activities on health conditions requires further research. The observed conclusion that smoking and experiencing sleep troubles negatively affect health conditions is expected. However, the findings related to alcohol consumption and regular physical activity are counterintuitive. Typically, alcohol consumption is associated with negative health effects and regular physical activity is widely recognized for its positive impact on health. Our conclusion is consistent with existing literature that identifies smoking as a contributor to health issues [2] and suggests that low to moderate alcohol consumption may be associated with better health outcomes [3]. However, due to the absence of a variable quantifying the extent of alcohol consumption in our model, we are unable to confirm the finding that a higher amount of alcohol consumption correlates with an increased risk of poor health.

During the model-building process, we noticed that a large portion of the participants had 0 days of bad health, making our response severely right-skewed and our model faced a serious violation of the normality assumption that could not be fixed by the Boxcox method. We therefore focused our investigation only on participants who had experienced at least one day of poor health. A small dataset can lead to an unreliable and overfit model in machine learning, as it may not capture the underlying patterns or generalize well to new data due to a limited sample size. Upon checking for problematic observations, we found 58 leverage observations, no outliers, and 99 influential points. Problematic observations like outliers can significantly skew the results, as the model tries to fit these anomalies, leading to inaccurate or biased estimates of the relationships between variables. These observations are not removed because their removal is not justifiable based on domain knowledge, they may be genuine data points. In addition, while the model built using the testing data set satisfies all assumptions and has similar, proportions and types of problematic observations and a similar amount of multicollinearity as the model built using the training data set, the p-value for ANOVA test for overall significance is 0.1615, indicating no statistically significant linear relationship exists for any predictors. The reason for this may be because we have a relatively small data set and collecting more data may not be feasible in this case. The lack of model validation can result in a model that may not generalize well to unseen data, leading to potential inaccuracies, overfitting, and reduced model reliability.

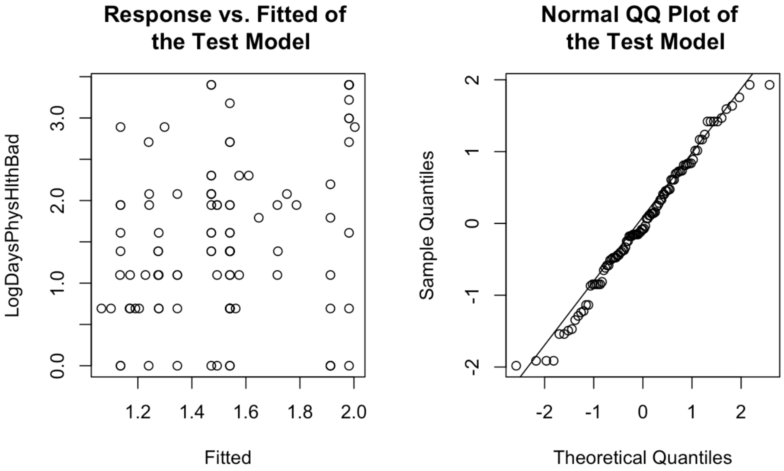

Figure 4: Response vs. Fitted and QQ plot of Both the Model Built with the Testing Data Set

Predictor Name	Estimate	P-Value
(Intercept)	1.49330	< 1.59e-05
DaysPhysHlthBad	-0.03538	0.6130
Smoke100n	0.06883	0.7205
PhysActive	-0.15917	0.6127
SleepTrouble	0.44118	0.0401
Alcohol12PlusYr	-0.02147	0.9490

Table 6: Summary of the Model Built with the Testing Data Set

PhysActiveDays	Smoke100n	PhysActive	SleepTrouble	Alcohol12PlusYr
2.716124	1.040256	2.752496	1.048997	1.037393

Table 7: The VIFs of the Model Built with the Testing Data Set

Appendix

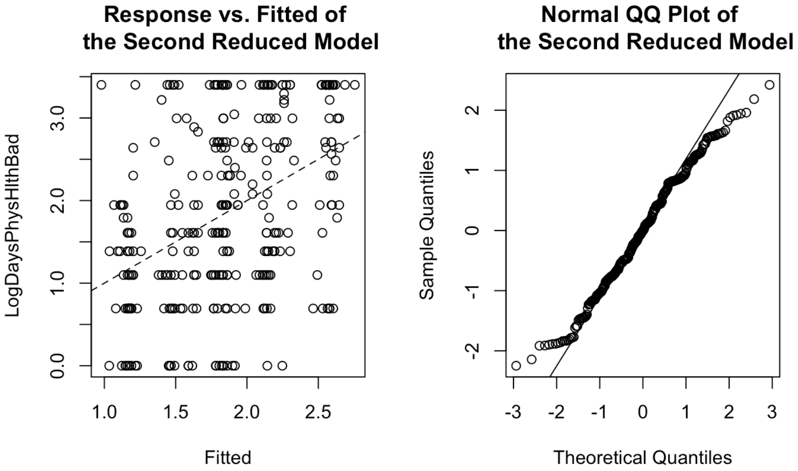

Figure 5: Response vs. Fitted and QQ plot of the Second Reduced Model

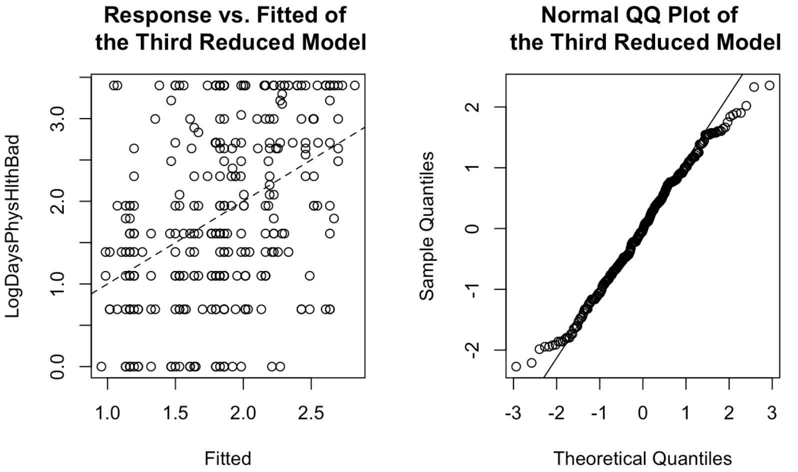

Figure 6: Response vs. Fitted and QQ plot of the Third Reduced Model

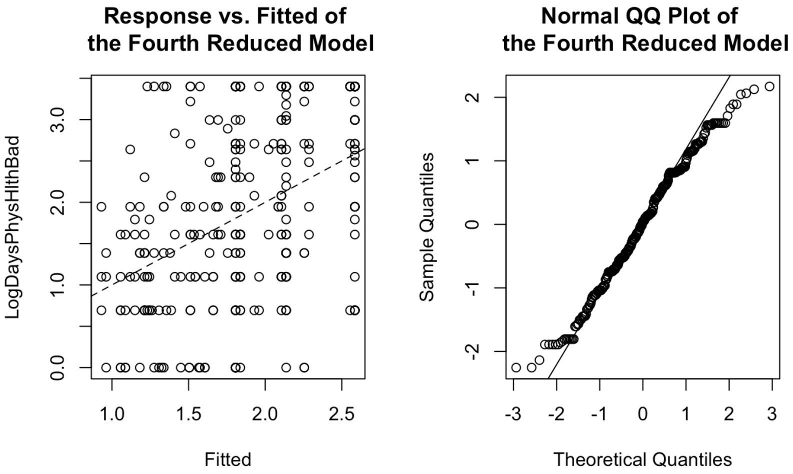

Figure 7: Response vs. Fitted and QQ plot of the Fourth Reduced Model

References

[1] Humphreys, B. R., McLeod, L., & Ruseski, J. E. (2014). Physical Activity and Health Outcomes: Evidence from Canada. *Health Economics*, 23(1), 33–54. `https://doi.org/10.1002/hec.2900`

[2] Babizhayev, M. A., & Yegorov, Y. E. (2011). Smoking and health: association between telomere length and factors impacting on human disease, quality of life and life span in a large population-based cohort under the effect of smoking duration. *Fundamental & Clinical Pharmacology*, 25(4), 425–442. `https://doi.org/10.1111/j.1472-8206.2010.00866.x`

[3] O'Keefe, J. H., Bybee, K. A., & Lavie, C. J. (2007). Alcohol and Cardiovascular Health: The Razor-Sharp Double-Edged Sword. *Journal of the American College of Cardiology*, 50(11), 1009–1014. `https://doi.org/10.1016/j.jacc.2007.04.089`