Kelly Lwin

STA2260 Final Project

Kevin Bailey

8 December 2024

Regression Analysis for Health and Lifestyle Data

Objective

The data set I am going to use is Health and Lifestyle Data for Regression that I found on kaggle.com. I chose this data set because I find it interesting to explore the relationship between lifestyle factors and health scores. My goal is to analyze how different predictors relate to health scores using regression analysis. I will identify predictors, examine their relationships with the response variable, and create models that predict an individual's health score based on their lifestyle.

Analysis

Based on the data set information from Kaggle, I have chosen Health Score as my response variable since it represents overall health status. The Health Score is a continuous variable ranging from 0 to 100; it reflects an individual's overall health. The remaining columns in the dataset will be potential predictor variables.

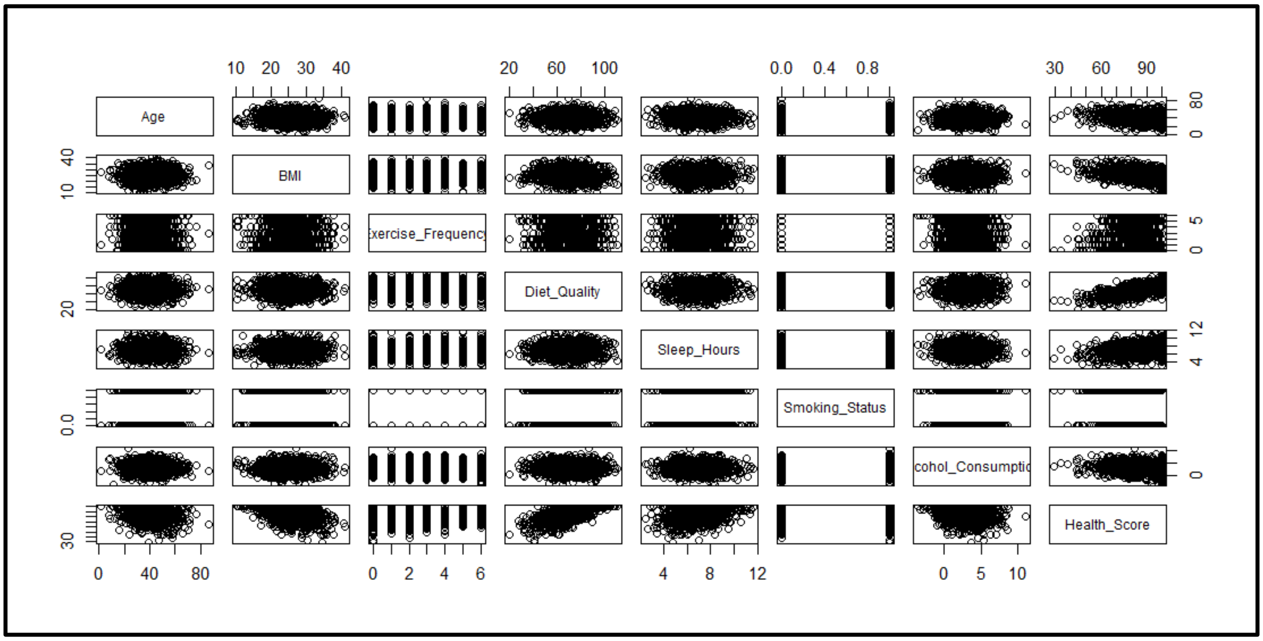
- Age: Age of the individual in years
- BMI: Body Mass Index of the individual
- Exercise Frequency: Number of days per week the individual exercises (categorical, values 0-7)
- Diet Quality: An index reflecting diet quality, with higher values indicating healthier

dietary habits (continuous, values 0-100)

- Sleep Hours: Average hours of sleep per night
- Smoking Status: Binary variable where 0 = Non-smoker, 1 = Smoker
- Alcohol Consumption: Average alcohol units consumed per week

Model 1 (m1)

I will create a one-predictor linear model to explore the relationship between a single predictor and my response variable (Health Score). By fitting a one-predictor model, I can clearly determine which individual predictors have the strongest relationship with the response variable when considered in isolation. As seen below, Smoking Status definitely is not a good predictor because there is not enough variability to explain significant changes in the response variable.



Similarly, Exercise Frequency is also categorical. It represents the number of days per week an individual exercises and only ranges from 0 to 7. Since I believe both variables have discrete categories rather than continuous numeric values, they are not ideal for predicting health score in this analysis.

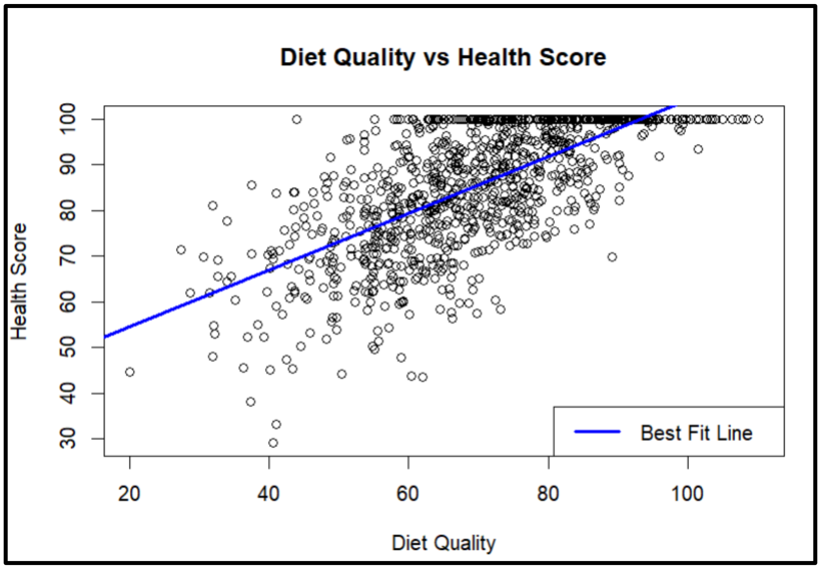
Additionally, Age might not be a strong predictor compared to others. People of the same age can have very different health outcomes due to lifestyle, genetics, and environmental factors. Factors like exercise, diet, and sleep have a more direct impact on health than age as they are modifiable. Therefore, I will focus on the other variables: BMI, Diet Quality, Sleep Hours, and Alcohol Consumption.

Computing the R^2 and AIC values

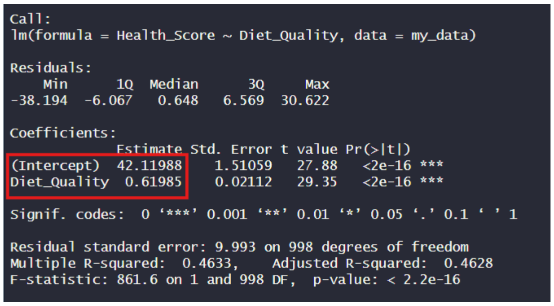
To determine the best predictor for the health score, I will compute and compare the R-squared (R^2) and Akaike Information Criterion (AIC) values for each potential predictor variable. R^2 represents the proportion of variance in the response explained by the predictor. A higher R^2 value means the predictor explains a larger portion of the variability in the response. Adjusted R^2 penalizes for adding variables that do not agree with the model. AIC is the measure of model fit, and lower AIC values indicate better-fitting models.

Predictors	**BMI**	**Diet Quality**	**Sleep Hours**	**Alcohol Consumption**
R^2	0.1723	0.4633	0.07238	0.01912
Adjusted R^2	0.1715	0.4628	0.07145	0.01813
AIC	7878.826	7445.608	7992.855	8048.685

Based on the results, Diet Quality appears to be the best predictor of health score among the four variables tested, as it has the highest R-squared and the lowest AIC. Following that, BMI shows a meaningful relationship with the Health Score. Sleep Hours and Alcohol Consumption are statistically significant but have low explanatory power. Therefore, I settled on using Diet Quality as a predictor for how high the Health Score would be.



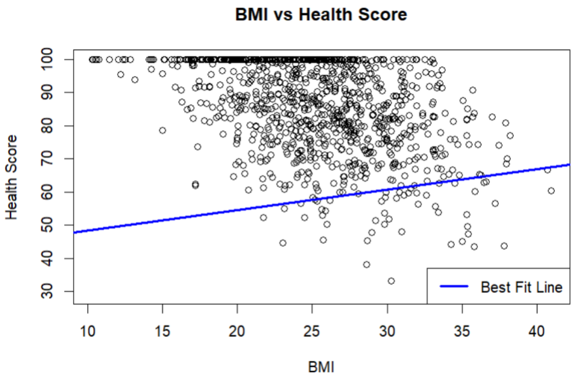
Linear regression assumes a linear relationship between the dependent variable y and independent variable x of the form y = mx + b. The model is expressed as: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. To find the line of best fit that minimizes the error between the predicted and observed values, we choose the parameters β_0 and β_1 that best fit the data and make predictions using the fitted line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

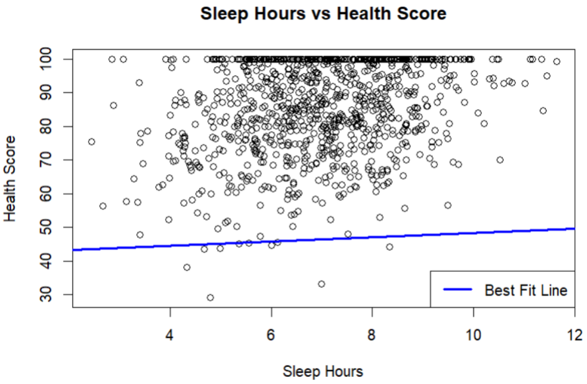


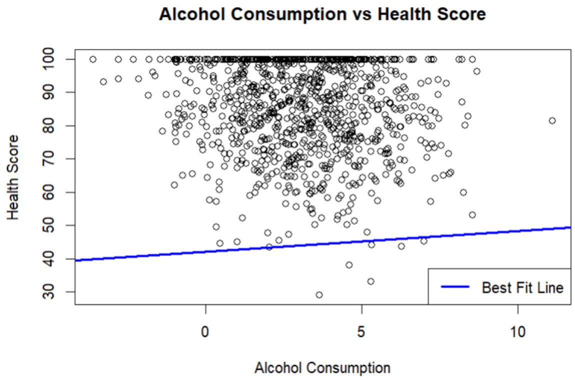
```
Call:
lm(formula = Health_Score ~ Diet_Quality, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.194  -6.067   0.648   6.569  30.622 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  42.11988    1.51059   27.88   <2e-16 ***
Diet_Quality  0.61985    0.02112   29.35   <2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.993 on 998 degrees of freedom
Multiple R-squared:  0.4633,	Adjusted R-squared:  0.4628 
F-statistic: 861.6 on 1 and 998 DF,  p-value: < 2.2e-16
```

Based on the model and the data, R-square = 0.4628 and AIC = 7445.608. The equation for this model is calculated in R: y = 0.61985x + 42.11988.

Approximately 46.3% of the variation in Health Score can be explained by Diet Quality alone. For every unit increase in Diet Quality, the Health Score is expected to increase by approximately 0.62, and when Diet Quality is 0, the Health Score is around 42.12.







These are examples of other predictors in comparison to the response of the Health Score. As shown, these three predictors have a fairly low slope. While they all are in linear regression, they are not as apparent as the m1 model of the Diet Quality vs. Health Score.

Model 2 (m2)

Linear models minimize squared error by selecting coefficients (β) that reduce the distance between predicted and observed values. Adding more terms to a model always improves in-sample performance by reducing the residual sum of squares (RSS). For example, if we fit:

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$ and then fit: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2^2$ the second model will always have at least the same squared error as the first.

Earlier, I found that BMI also has a decent relationship with Health Score following the Diet Quality. Therefore, I would create two new models using these two predictors and find the R^2 and AIC values.

Predictors	**Diet Quality vs. BMI** **(m2.dq.bmi)**	**Diet Quality vs. BMI** **(squared) (m2.dq.bmi.sq)**

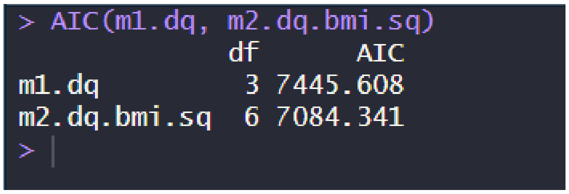
R^2	0.6141	0.6283
Adjusted R^2	0.6133	0.6268
AIC	7117.880	7084.341

Since the second model—I will call this Model 2 (m2)— has higher R^2 and lower AIC. Hence, I will compare it with Model 1 (m1) again.

Model m1 and Model m2 Comparison

Predictors	**m1: Diet Quality vs. Health Score**	**m2: Diet Quality vs. BMI (squared)**
R^2	0.4633	0.6283
Adjusted R^2	0.4628	0.6268
AIC	7445.608	7084.341

When comparing these models, Model m2 explains a higher proportion of the variability in the data, as indicated by its higher R² value (0.6283 compared to 0.4633 in m1). The higher R² suggests that Diet Quality is a better predictor of BMI than of Health Score; it can explain more of the variation in the response variable.



```
> AIC(m1.dq, m2.dq.bmi.sq)
             df      AIC
m1.dq         3 7445.608
m2.dq.bmi.sq  6 7084.341
> 
```

Lastly, Model m2 has a lower AIC (7084.341 compared to 7445.608 in m1). This indicates that it fits the data more efficiently and that it can provide a better balance between model complexity and accuracy. Therefore, it is more likely to generalize well when making predictions on new or unseen data.