The Weight of Time : Exploring the Age - BMI relationship through Descriptive Regression Analysis

DATASCI 203: Statistics for Data Science
Branndon Marion

Table of Contents

Introduction

In the realm of health research, age and Body Mass Index (BMI) are pivotal markers, each playing unique roles in offering insights into an individual's well-being. This descriptive analysis explores the relationship between age and BMI, while considering a range of demographic and lifestyle variables as operationalized indicators. BMI serves as a widely-used metric for assessing body composition, reflecting the balance between weight and height and providing insights into potential health risks, such as obesity-related conditions. By understanding how age interacts with factors such as dietary habits and physical activity levels, this study aims to elucidate patterns and trends in BMI across different population groups, capturing some of the diversity and complexity inherent in human health. Understanding these relationships can contribute to public health discourse, aid in/serve as a starting point for informing evidence-based strategies for health promotion and disease prevention across the lifespan. This leads us to the following research question, which we aim to address in this study:

What is the relationship between age and BMI, while taking into account other demographic and lifestyle factors?

Data & Methodology

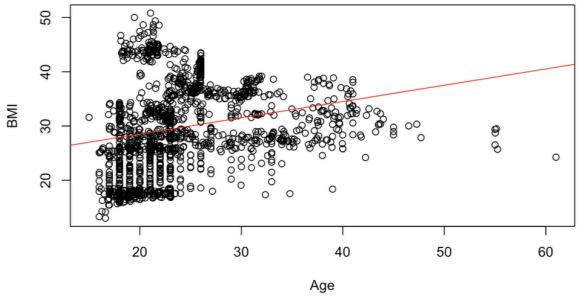
The observational dataset used in this analysis originates from the article *Obesity Level Estimation Software based on Decision Trees*. This dataset contains 2111 records and comprises 17 attributes pertaining to an individual's demographic characteristics, eating and lifestyle habits, and family weight history. 23% of the records were collected from surveys administered to nationals from Columbia, Mexico, and Peru through a web platform, and the remaining 77% of records were generated synthetically using the Weka tool and the SMOTE filter[1]. BMI was not originally presented as part of the dataset, and was added using the weight and height variables in the following equation $BMI = \frac{weight\ (kg)}{height^2\ (m^2)}$

While exploring the data, a scatter plot was generated to visualize the shape of the relationship between age and bmi as seen in Figure 2. The sampling is dense when age is less than 30 years, with large variance in the data which made it difficult to observe the shape of the. Once age is greater than 30 years, the density of the sampling changes drastically. We created an indicator variable for when age is greater than 30 to see if we could learn more about the data. Key features relevant to this analysis can be found in Figure 1.

Figure 1: Variable Descriptions

Variable Name	Description	Type	Units
age		continuous	years
age_30	Indicator for age > 30	binary	yes/no
hist_obesity	Has a family member suffered or suffers from being overweight?	binary	yes/no
high_calorie	Do you eat high calorie food frequently?	binary	yes/no
active	How often do you have physical activity?	continuous	days per week
ch2o	How much water do you drink daily?	continuous	glasses per day
bmi		continuous	kg/m^2

Figure 2: Scatter Plot of Age & BMI



We started off with a set of simple restricted models, which aims to elucidate the relationship between age and BMI. Our first model represents age in its original form, whilst our second model captures age as the indicator variable previously mentioned.

$$(1) \quad \widehat{bmi} = \beta_0 + \beta_1 age \qquad (2) \quad \widehat{bmi} = \beta_0 + \beta_1 age\_30$$

Given the complex nature of health and the multitude of potential factors that may contribute to the relationship, we expanded our model by selecting additional variables that resonated intuitively. We chose to include family history as this variable substantially impacts health and health trajectories. However, since family history is inherently out

[1] De la Hoz Manotas, Alexis & De la Hoz Correa, Eduardo & Mendoza, Fabio & Morales, Roberto & Sanchez, Beatriz. (2019). Obesity Level Estimation Software based on Decision Trees. Journal of Computer Science. 15. 10. 10.3844/jcssp.2019.67.77.

of our control, we sought to also investigate factors within our control that are widely applicable and easily interpretable. This resulted in the inclusion of physical activity levels and dietary habits, such as consumption of high caloric foods and water consumption levels. By incorporating both uncontrollable variables such as family history and controllable variables such as lifestyle habits, our analysis strives to offer a holistic understanding of how age interacts with BMI within the broader context of individual lifestyles and demographic characteristics.

$$(3)\ \widehat{bmi} = \beta_0 + \beta_1 age + \beta_2 history\_overweight + \beta_3 activity + \beta_4 high\_caloric + \beta_5 ch2o$$

$$(4)\ \widehat{bmi} = \beta_0 + \beta_1 age\_30 + \beta_2 history\_overweight + \beta_3 activity + \beta_4 high\_caloric + \beta_5 ch2o$$

Results

The regression output for our models can be seen in Figure 3. Model (1) and (2) represent our simple model and capture age in its original form, and age_30 as an indicator for when age is greater than 30 years, respectively. In both models, the age concept was statistically significant in explaining the variance of the bmi concept. Model (3) and (4) represent the set of expanded models which capture other intuitive factors in addition to the age concept. A t-test was performed to conduct hypothesis testing for statistical significance. Within the set of expanded models, each of the additional variables was statistically significant except for the age_30 variable in Model (4).

Figure 3: Linear Model Regression Outputs

Regression Models of Age, Family History, and Lifestyle Habits on BMI

	Dependent variable:			
	bmi			
	(1)	(2)	(3)	(4)
Age	0.311***		0.164***	
	(0.032)		(0.029)	
Age Over 30		1.879***		0.525
		(0.569)		(0.486)
Family History of Overweight			8.542***	9.013***
			(0.474)	(0.474)
Physical Activity			−1.222***	−1.386***
			(0.212)	(0.212)
Water Consumption			1.346***	1.295***
			(0.296)	(0.301)
Frequent Consumption of High Caloric Foods			3.691***	3.737***
			(0.541)	(0.547)
Constant	22.083***	29.330***	14.068***	17.799***
	(0.812)	(0.230)	(1.041)	(0.809)
Observations	1,477	1,477	1,477	1,477
R2	0.059	0.007	0.315	0.300
Adjusted R2	0.058	0.007	0.312	0.298
Residual Std. Error	7.872 (df = 1475)	8.085 (df = 1475)	6.728 (df = 1471)	6.799 (df = 1471)
F Statistic	92.433*** (df = 1; 1475)	10.898*** (df = 1; 1475)	135.022*** (df = 5; 1471)	126.071*** (df = 5; 1471)
Note:				*p<0.1; **p<0.05; ***p<0.01

An ANOVA test was performed between the simple and expanded models that captured the age concept in its original form, Model (1) and (3). We did not repeat the ANOVA test for Model (2) and (4), as the age_30 variable was not statistically significant in the expanded Model (4). Results from the ANOVA test can be found below in Figure 4. Cohen's D was calculated to evaluate practical significance for each model, respectively. The results can be seen in Figure 5.

Figure 4: ANOVA Test Outputs

```
Analysis of Variance Table

Model 1: bmi ~ age + hist_overweight + active + ch2o + high_caloric
Model 2: bmi ~ age
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   1471 66577
2   1475 91405 -4    -24827 137.14 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: Cohen's D Values

Model	Cohen's D
Model 1	0.0627
Model 2	0.0074
Model 3	0.4589
Model 4	0.4285

Discussion

In Model (1) and (2), we see that age is statistically significant, leading us to reject the null hypothesis that there is no relationship between age in its original form and bmi, as well as age_30 and bmi. In Model (1), the coefficient of age can be interpreted as for every year increase in age, we see a .311 unit increase in BMI. In Model (2), because age_30 is an indicator, we interpret the coefficient to mean that for people over 30 years old, BMI is higher by 1.879 units on average than for people under 30 years old. However, the low R^2 value in both restricted models suggests that they are not a great fit for the observed data. In Model (3), all variables were statistically significant, therefore we can reject the null hypothesis that there is no relationship between the variables in the model and bmi. Additionally, observing the sign of the coefficients, we can see that increased physical activity correlates with a decrease in BMI while frequent consumption of high caloric foods correlates with an increase in bmi. Family history stands out in this model, as the magnitude of the coefficient is much larger than that of the other variables. In Model (4), we see that age_30 is no longer statistically significant although all the other variables are. Because of this, we can conclude that the age indicator variable is less robust than age itself, which is statistically significant even when other variables are included. This suggests age to have a continuous relationship with bmi, and there is no specific point in time where a drastic change can occur.

We further evaluated this inclusion of additional variables in Model 3 using an ANOVA test to compare it with Model (1). The test showed that by expanding our model, we greatly improved our regression sum of squares. The statistically significant high F-test value of 137.14, and large reduction in RSS led us to the interpretation that the additional variables are useful at describing the relationship between age and bmi. While the total RSS is still high, suggesting that there is still a lot of unexplained variance in the models, the ANOVA test further demonstrates the value of the variables chosen.

Since all variables present in Models (1), (2), and (3), had statistically significant relationships with BMI, we examined the practical significance of these models using Cohen's D. Practical significance increased as we captured more variables in the relationship, as seen in the .396 change from Model (3) from Model (1), echoing the multifaceted relationship between age and health. However, while we could have added additional variables to potentially further increase the practical significance of the model and potentially further reduce the RSS of the model, we saw that there was a diminishing marginal reduction in RSS with additional variables while adding complexity to the model.

The effect size of age varies significantly based on the model. According to the Center of Disease Control (CDC), a healthy BMI range is 20-25 for men and 18.5-23.5 for women, with anything over being considered overweight, and BMIs another 5 points higher (30 for men, 28.5 for women) being considered obese. In Model 1, the coefficient of .311 means that a population could move from a healthy to obese range over the course of 16 years. With the non linear relationship with age modeled in Model 2, the jump of 1.8 BMI points at age 30 also takes a significant jump within the health classifications of BMI ranges. However, with these ranges in mind, family history demonstrates an especially large effect size – covering over 8 BMI points in Model 3, and thus representing the difference between a healthy population and an obese population.

Conclusion

After training multiple models to determine the relationship between age and BMI, including several other lifestyle and demographic features, we saw that even the best performing model evaluated still had a lot of unexplained variance, highlighting how complex it is to describe BMI. However, we did find that increasing age had a statistically significant correlation with increasing BMI, indicating that public health efforts to address weight concerns may need to be tailored to different age groups. Even so, BMI is only one measure that can correlate to general health and obesity, and should not be considered an all-encompassing metric of population health. Although there is more work to be done, we hope that this descriptive analysis provides some insight into the multi-faceted relationship between variables to describe bmi.