

Healthcare and Medical Analytics Individual assignment

Bernardo Carvalho
24/06/2024

Table of Contents

| | |
|---|-----------------|
| <i>Introduction.....</i> | <i>3</i> |
| <i>Descriptive statistics.....</i> | <i>3</i> |
| <i>Regressions</i> | <i>4</i> |
| <i>Discussion and limitations</i> | <i>6</i> |
| <i>References</i> | <i>8</i> |
| <i>Appendix</i> | <i>9</i> |
| Appendix 1 – Simple regression, no panel | 9 |
| Appendix 2 - Panel 1: Children that were not overweight and became overweight. | 10 |
| Appendix 3 - Panel 2: Children who did not transition..... | 11 |
| Appendix 4 – Logit baseline model | 11 |

Introduction

The objective of this study is to investigate the habits and education in early childhood and the development of overweight at a later stage in life. We use as a starting point the work from Jensen et. al. where the relationship of cardiovascular risk factors in the parents of the children in wave 1 is compared to the cardiovascular risk factors of the adults in wave 5.

Our goal is to check if the education the children are receiving in school about dieting is showing any effect later in life. We also check if other household factors, such as typical breakfast choices and times that the children have dinner with the parents impact the metric in adulthood.

The article “Life-course trajectories of body mass index from adolescence to old age: Racial and educational disparities” argues that:

“[...] adolescence and young adulthood are critical life stages when excess weight can rapidly accumulate and racial/ethnic or educational disparities emerge, most significantly among recent cohorts of young people [...]” (Yang et. al, 2021)

The study, however, concentrates on the social aspects of the children’s upbringing. We will aim to control for these factors, while also investigating other educational factors to see if there are clear actions that could be taken from a public health standpoint that would help to mitigate the risk of becoming overweight at a later stage in life. The study “Socioeconomic inequalities in childhood and adolescent body-mass index, weight, and height from 1953 to 2015:” also found that inequalities in childhood widen gaps in BMI in later stages in life.

Descriptive statistics

Our dataset consists of the merged data of Wave 1 and Wave 5 questionnaires, with Wave 1 also having information about the family context. The main dependent variable we wish to explain is the BMI of individuals at wave 5. This is constructed from the height and weight values. Then, we have 4 groups of dependent variables:

1. Wave 5: habits
 - a. Exercise → Self-reported exercise frequency in bands
2. Wave 1: background
 - a. Gender → Self-reported gender at birth
 - b. Ethnicity
 - c. Urbanity
 - d. Household income
3. Wave 1: baseline health indicators
 - a. BMI at wave 1 → Calculated measure from self-reported weight and height
 - b. Self-reported general health
4. Wave 1: Dieting educational factors.
 - a. Learned about diet in school.
 - b. Learned about importance of exercise.
 - c. Learned about obesity in school.
 - d. How often they have dinner with parents.
 - e. Usual breakfast. Grouped by [Cereal, Fruit, Eggs, Meat, Snack and Bread]

Questions 4.a, 4.b and 4.c are from topics that the children see in school: which foods you should eat, the importance of exercise and the risk of obesity.

Our focus are the dieting educational factors, although we need to try to control for the social background and baseline health in Wave 1. We consider having dinner with parents and the typical breakfast as “educational” under the assumption these are things that can enhance the relationship with food and the understanding of the importance of a balanced meal. Our assumption is that teenagers that do not have dinner with their parents will more often revert to easier meal options, such as frozen food.

Regressions

After preparing the data and checking our assumptions on distributions, we start fitting a regression model to understand the relationship between the independent variable and the dependent ones. We start with a simple model:

$$\begin{aligned}
y_{w5\ bmi} = & \beta_{w1\ bmi} w1_{bmi} + \beta_{\log\ income} \log(w1_{income}) + \beta_{w1\ gen\ health} w1_{gen\ health} \\
& + \beta_{w1\ dinner\ parents\ freq} w1_{dinner\ parents\ freq} + \theta_{female} + \theta_{ethnicity:black} + \theta_{ethnicity:other} \\
& + \theta_{urbanity:not\ completely\ urban} + \theta_{learned\ diet\ importance} + \theta_{learned\ exercise\ importance} \\
& + \theta_{learned\ obesity\ risks} + \theta_{usual\ breakfast:cereal} + \theta_{usual\ breakfast:fruit} + \theta_{usual\ breakfast:eggs} \\
& + \theta_{usual\ breakfast:meat} + \theta_{usual\ breakfast:snack} + \theta_{usual\ breakfast:bread}
\end{aligned}$$

In our first regression, we observe expected results like what Yang et. al. had, in which the main factors for high BMI in wave 5 are household income and ethnicity. We also observe that wave 1 BMI is a strong predictor of wave 5 BMI. The full result of the regression is available in [Appendix 1](#). This baseline model has an adjusted R² score of 36%. Most of the explainability of the model comes from the obvious candidates.

However, we observe that regularly consuming meat during breakfast also presented a significant value (p<.01). To investigate this trend further, we split the data into two panels: Children who had a healthy BMI (Prentice A. M, 1998) and became obese by wave 5 (BMI > 30) (All About Adult BMI, 2022) and the rest. Our strategy here is to isolate the effect of maintaining the BMI trend throughout adulthood and focus on finding the explanations for BMI in those children who transitioned from one group to the other. The full regression result is available in [Appendix 2](#). In this model, the female variable BIO_SEX:Female had the highest impact on the BMI score with a high significance. Although we must note that the overall explainability of the model (R²) dropped significantly to 8%. Nevertheless, this is consistent with the finding that women are more prone to obesity (All About Adult BMI, 2022).

By analyzing the other panel, we continue to observe the similar effects as with the general data, except that we don't attribute any of the effects to the breakfast habits (see [Appendix 3](#))

Finally, we fit a last set of models that explain the transition from healthy BMI to obese status. We start by creating a baseline model, where we only keep the variables that presented the highest impact so far. The results of the baseline model are available in [Appendix 4](#). For this model, we observe a Log likelihood of -2164 and AIC of 4343. We then fit another model adding the "Education" variables, so that we can compare the results.

| Dependent variable: | |
|-----------------------------------|-----------------------------|
| w5_become_overweight | |
| BIO_SEXFemale | 0.160* (0.082) |
| w1_ethnicityBlack | 0.062 (0.120) |
| w1_ethnicityOther | -0.150 (0.229) |
| w1_urbanityNot completely urban | -0.046 (0.081) |
| w1_income_1994_log_win | -0.569*** (0.123) |
| w1_general_health_int | 0.074 (0.047) |
| w1_learned_dietYes | 0.083 (0.139) |
| w1_learned_importance_exerciseYes | -0.077 (0.176) |
| w1_learned_problems_obesityYes | 0.128 (0.086) |
| w1_dinner_with_parents | 0.002 (0.020) |
| w1_usual_breakfast_cerealYes | -0.071 (0.082) |
| w1_usual_breakfast_fruitYes | -0.217** (0.089) |
| w1_usual_breakfast_eggsYes | 0.068 (0.126) |
| w1_usual_breakfast_meatYes | 0.277* (0.146) |
| w1_usual_breakfast_snackYes | 0.045 (0.157) |
| w1_usual_breakfast_breadYes | -0.018 (0.089) |
| Constant | 1.030* (0.607) |
| Observations | 3,574 |
| Log Likelihood | -1,915.189 |
| Akaike Inf. Crit. | 3,864.379 |
| Note: | *p<0.1; **p<0.05; ***p<0.01 |

Table 1 – Final Logit model results

Adding the education variables has increased the performance of the fit (although only slightly).

Discussion and limitations

In this study, we aimed to investigate the importance of nutritional education in teenage years and how that correlates with obesity metrics later in life. Our goal was to understand if there are clear indication of actions that can be publicly targeted that have the potential of decreasing the prominence of obesity among adults.

Although some variables were shown to have a significant relationship to the dependent variable from a statistical standpoint, there are several limitations. For example, we ignored the correlations between income and education quality. We know

that in the U.S, school quality is highly tied to the district it serves. Furthermore, although we saw some level of explainability between breakfast habits and the likelihood of becoming obese, these habits can be influenced by social-economic factors. Nevertheless, it still speaks to reason that breakfast habits in teenage years could be important indicators of lifelong eating habits. It is worth noting that investing in the awareness of breakfast habit importance has very little downside, and even with the slim evidence of improvement, when applied to such a large population its impact can be very large.

Other limitations discussed throughout this study are the fact that we had to perform data imputation on both the exercise and income variables. Although in the case of exercise the variable was not so relevant, we used a very simplistic approach without any further investigation on it. We took a similar approach with the income data, where we made the imputation based on the mean. It could be the case that the missing income data is biased. For example, if high income families tend not to disclose their income, then our approach would not be correct.

Another limitation of our study is that we did not perform any re-weighting of variables. Add health provides methods to reweigh the data based on the expected population distribution at the time, but we didn't do this. Examples of these are the gender data, where in our final dataset women were overrepresented. That is probably explained by the fact that more women stuck to the program (and probably we observed less obits) than men. Furthermore, the ethnic distribution became unbalanced in comparison to the original dataset after the EDA. We also note that the classification used is simplistic (only between white, black and others). These considerations are further explored in the code.

References

- Jensen, T. M. et al., 2021. Like Parent, Like Child: Intergenerational Patterns of Cardiovascular Risk Factors at Midlife. *Journal of Adolescent Health*. 68 (3), 596-603. [10.1016/j.jadohealth.2020.06.039](https://doi.org/10.1016/j.jadohealth.2020.06.039).
- Yang, Y. C. et al., 2021 Life-course trajectories of body mass index from adolescence to old age: Racial and educational disparities. *Proceedings of the National Academy of Sciences of the United States of America*, 118(17), e2020167118.
<https://doi.org/10.1073/pnas.2020167118>
- Prentice A. M. (1998). Body mass index standards for children.
Are useful for clinicians but not yet for epidemiologists. *BMJ (Clinical research ed.)*, 317(7170), 1401–1402. <https://doi.org/10.1136/bmj.317.7170.1401>
- All about adult BMI*. (2022, June 3). Centers for Disease Control and Prevention.
https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html
- Bann, D. et al., 2018. Socioeconomic inequalities in childhood and adolescent body-mass index, weight, and height from 1953 to 2015: an analysis of four longitudinal, observational, British birth cohort studies. *The Lancet Public Health*, 3(4), pp. 194-204.

Appendix

Appendix 1 – Simple regression, no panel

Simple regression - No changes

| Dependent variable: | |
|-----------------------------------|---------------------------|
| w5_bmi | |
| w1_income_1994_log_win | -1.295*** (0.316) |
| w1_ethnicityBlack | 0.781** (0.306) |
| w1_ethnicityOther | -0.352 (0.548) |
| w1_urbanityNot completely urban | 0.047 (0.199) |
| BIO_SEXFemale | 0.253 (0.208) |
| w1_bmi | 0.932*** (0.023) |
| w1_general_healthVery good | 0.513** (0.239) |
| w1_general_healthGood | 0.634** (0.279) |
| w1_general_healthFair | 0.398 (0.480) |
| w1_general_healthPoor | -1.265 (1.957) |
| w1_learned_dietYes | 0.418 (0.339) |
| w1_dinner_with_parents | 0.106** (0.050) |
| w1_learned_problems_obesityYes | 0.252 (0.211) |
| w1_learned_importance_exerciseYes | -0.613 (0.431) |
| w1_usual_breakfast_cerealYes | -0.283 (0.203) |
| w1_usual_breakfast_fruitYes | -0.106 (0.216) |
| w1_usual_breakfast_eggsYes | 0.631** (0.317) |
| w1_usual_breakfast_meatYes | 0.749** (0.373) |
| w1_usual_breakfast_snackYes | -0.256 (0.392) |
| w1_usual_breakfast_breadYes | -0.172 (0.219) |
| w5_exercise1 or 2 times | 0.748 (0.587) |
| w5_exercise3 to 5 times | 0.346 (0.550) |
| w5_exercise6 or 7 times | 0.277 (0.602) |
| w5_exerciseMore than 7 times | 0.640 (0.590) |
| Constant | 12.993*** (1.729) |
| Observations | 3,533 |
| R2 | 0.369 |
| Adjusted R2 | 0.364 |
| Residual Std. Error | 5.819 (df = 3508) |
| F Statistic | 85.394*** (df = 24; 3508) |

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix 2 - Panel 1: Children that were not overweight and became overweight.

Panel 1: Children who were not overweight and became overweight

| Dependent variable: | |
|-----------------------------------|-------------------------|
| ----- | |
| | w5_bmi |
| ----- | |
| w5_exercise1 or 2 times | -0.134 (0.857) |
| w5_exercise3 to 5 times | 0.430 (0.814) |
| w5_exercise6 or 7 times | 0.196 (0.892) |
| w5_exerciseMore than 7 times | -0.096 (0.875) |
| BIO_SEXFemale | 1.724*** (0.322) |
| w1_ethnicityBlack | -0.043 (0.453) |
| w1_ethnicityOther | -0.489 (0.877) |
| w1_urbanityNot completely urban | -0.224 (0.308) |
| w1_income_1994_log_win | 0.616 (0.473) |
| w1_bmi | 0.503*** (0.076) |
| w1_general_health_int | -0.136 (0.187) |
| w1_learned_dietYes | -0.215 (0.533) |
| w1_dinner_with_parents | 0.089 (0.077) |
| w1_learned_problems_obesityYes | 0.036 (0.329) |
| w1_learned_importance_exerciseYes | -0.291 (0.678) |
| w1_usual_breakfast_cerealYes | 0.152 (0.314) |
| w1_usual_breakfast_fruitYes | 0.524 (0.340) |
| w1_usual_breakfast_eggsYes | 0.564 (0.481) |
| w1_usual_breakfast_meatYes | 0.783 (0.536) |
| w1_usual_breakfast_snackYes | 0.090 (0.589) |
| w1_usual_breakfast_breadYes | -0.372 (0.341) |
| Constant | 20.030*** (2.921) |
| ----- | |
| Observations | 831 |
| R2 | 0.103 |
| Adjusted R2 | 0.080 |
| Residual Std. Error | 4.319 (df = 809) |
| F Statistic | 4.430*** (df = 21; 809) |

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix 3 - Panel 2: Children who did not transition

Panel 2: Children who did not transition

| Dependent variable: | |
|-----------------------------------|-----------------------------|
| w5_bmi | |
| w5_exercise1 or 2 times | 0.747 (0.583) |
| w5_exercise3 to 5 times | 0.603 (0.543) |
| w5_exercise6 or 7 times | 0.425 (0.595) |
| w5_exerciseMore than 7 times | 0.975* (0.584) |
| BIO_SEXFemale | -0.335* (0.203) |
| w1_ethnicityBlack | 0.778** (0.304) |
| w1_ethnicityOther | -0.187 (0.529) |
| w1_urbanityNot completely urban | 0.232 (0.195) |
| w1_income_1994_log_win | -0.598* (0.314) |
| w1_bmi | 1.020*** (0.020) |
| w1_general_health_int | -0.246** (0.115) |
| w1_learned_dietYes | 0.587* (0.329) |
| w1_dinner_with_parents | 0.098** (0.049) |
| w1_learned_problems_obesityYes | 0.025 (0.206) |
| w1_learned_importance_exerciseYes | -0.557 (0.419) |
| w1_usual_breakfast_cerealYes | -0.238 (0.198) |
| w1_usual_breakfast_fruitYes | 0.164 (0.210) |
| w1_usual_breakfast_eggsYes | 0.397 (0.311) |
| w1_usual_breakfast_meatYes | 0.225 (0.375) |
| w1_usual_breakfast_snackYes | -0.483 (0.386) |
| w1_usual_breakfast_breadYes | -0.072 (0.214) |
| Constant | 7.321*** (1.740) |
| Observations | 2,702 |
| R2 | 0.535 |
| Adjusted R2 | 0.531 |
| Residual Std. Error | 4.968 (df = 2680) |
| F Statistic | 146.675*** (df = 21; 2680) |
| Note: | *p<0.1; **p<0.05; ***p<0.01 |

Appendix 4 – Logit baseline model

Logit regression - Become overweight

| Dependent variable: | |
|------------------------|-------------------|
| w5_become_overweight | |
| BIO_SEXFemale | 0.148* (0.077) |
| w1_ethnicityBlack | 0.198* (0.106) |
| w1_ethnicityOther | -0.168 (0.216) |
| w1_income_1994_log_win | -0.648*** (0.114) |
| w1_general_health_int | 0.005 (0.044) |

| | |
|-------------------|-----------------------------|
| w1_bmi | -0.055*** (0.009) |
| Constant | 2.909*** (0.603) |
| ----- | |
| Observations | 4,014 |
| Log Likelihood | -2,164.876 |
| Akaike Inf. Crit. | 4,343.752 |
| ===== | |
| Note: | *p<0.1; **p<0.05; ***p<0.01 |