Does obesity affect academic performance ?

Roxane Dumont, Loïc Brunner and Alexandre Schroeter

June 10, 2018

Abstract

In this paper, we analyze the impact of obesity on academic performance during adolescence. We measure academic performance in two ways: the final year GPA, which summarizes non-standardized work of a student during the entire last school year, and the standardized aptitude test (SAT), taken on one exam day to earn college entry. Using *the National Longitudinal Survey of Youth 1997*, we show that overweight and obese teenagers score lower on the GPA during their final year of high school. We also discuss the effect of obesity on the SAT standardized test. We reach the conclusion that the SAT scores of overweight and obese teenagers do not differ significantly from healthier teenagers' scores. These findings suggest that a temporality and selection effect may play a role on the academic performance of teenagers.

1 Introduction

Today one third of the children in the United States suffer from obesity. This number keeps increasing year after year and as a result, obesity has become one of *WHO*'s main concerns. The way people eat and live lies at the core of this worldwide problem. Many organizations are trying to understand and fight the bad habits creating obesity.

It is also said that : *"Obesity is the reason that the current generation of youth is predicted to live a shorter life than their parents."* [1]. This could create a long-term problem and additional costs in our societies (i.e health care systems). "*The number of seriously overweight people is increasing almost everywhere in the world, including in the least-developed nations. Switzerland has not been spared and is among the 73 countries where the number of obese people has more than doubled in the past 35 years.*" [2]

Our study concerns the United States but, as a matter of fact, it could be conducted elsewhere. Obesity is a worldwide phenomenon with short and long-term implications. How is "being overweight" defined? Obesity is related to medical problems, inactivity, diabetes and other side effects. What about the psychological, academic and professional consequences of obesity ?

We decide to focus on the impact of obesity on academic performance among a population of teenagers. We study subjects from 14 to 22 years old because we believe that the consequences

[1] http://www.obesitycampaign.org/

[2] https://www.swissinfo.ch/eng/growing-problem_swiss-failing-to-control-obesity-levels/43316314

of obesity at this age on school performances would be stronger than in a population of younger children or adults. Adolescence is a particular period of life. Not only is it a "hormones party", but it also engages many psychological processes and as we all know, high school is a crucial moment in one's academic path. For these reasons, the question/ problematic of our paper is as follows:

The effect of obesity on school performance is unclear. On one hand, children who stay at home and do less sports have more time to work, but on the other hand, obesity could be an indicator of a poor psychological status and thus lead to worse academic performance. Our work tries to identify the mechanism of obesity and analyze the impact of obesity on school results.

2 Literature

As the first step of our research paper, we examine the existing literature on the subject. Obesity is an important research subject and its impact on academic results has also been tested but only on younger children (for example: children in kindergarden). Researchers have not tested teenagers so far. We have selected four relevant articles. All four of them discuss the relationship between obesity and school performance and have reached the same result; a negative correlation between the 2 variables. If one is overweight or suffers from obesity, his academic results will decrease. The four papers have used different statistical tools and different control variables; they all have a different approach and therefore are highly interesting and valuable to us.

Does weight status affect academic performance? Evidence from Australian children [3]. This article starts the research by presenting a problem of endogeneity among the dependent and independent variable. The first OLS hypothesis is not respected: $\mathbb{E}(\epsilon|x) = 0$. Indeed, the effect of the obesity variable is not explained enough and this could introduce a bias in the regression. To prevent the bias from happening, a useful tool is an instrumental variable. The experimenters used the Body Mass Index of the parent as instrument. It is valid because the parent BMI is exogenous to the regression and highly correlated with the subject's BMI. Another interesting point of the article is the fact that panel data was used which gives a temporality dimension highly valuable in a research subject such as obesity. Unfortunately, due to incomplete database, we could not use panel data. In their conclusion and thanks to panel data, Hui Shi and Chuhui Li state that the effect of obesity on academic performance changes with age and school degree. In upper classes, the effect of obesity is higher.

Obesity and Student Performance at School [4] is the second article we consider very interesting because it tests obesity and its impact on school scores among children by using a relevant control variable; the absenteeism rate. Absenteeism at school plays a strong role in children's academic success. If the student misses school often, he will have problems to follow courses and will get bad marks. Using a control variable as relevant as the absenteeism helps to capture the exact effect of obesity and improve the significance of the regression. In our study, we also use the number of absences at school (Hopefully, this variable can be found in the *NLSY* database). Although this paper's model to test obesity and its impact is strong, researchers conclude by writing: « *Data on the association of child overweight or obesity with levels of attendance are too*

[3] https://www.tandfonline.com/doi/abs/10.1080/00036846.2017.1418077

[4] https://www.ncbi.nlm.nih.gov/pubmed/16179079

sparse to draw conclusions."

The third article which has proven helpful for our study is, **Overweight children, weight-based teasing and academic performance** [5] This paper considers a psychological variable into its regression: weight-based teasing. It is certain that psychological influence must be considered to analyze the relationship between obesity and school. The stigmatization of overweight children at school may play a strong role in poor academic performance. If a person suffers psychologically at school, he will not be able to study properly. Incorporating the psychological variable in our research has proven more difficult than we expected. Unfortunately our database, *NLSY*, did not contain this type of data. The closest variable we found for the psychology mindset of a person (for our regression) is: "Is in bad health". (We will talk about it in the data description). In their conclusion, the researchers give a worthy advice: « *Psychosocial variables, such as weight-based teasing, should be considered in future research examining the impact of childhood obesity on school performance and in future intervention studies* »

Childhood Overweight and Academic Performance: National Study of Kindergartners and First-Graders [6] is the last article that we analyze, and the research was done among young children, in a kindergarden. The paper proposes an interesting point of view. The researchers use relevant control variables such as *parent-child interaction, birth weight, physical activity, and television watching*. The effect of obesity is significant, but it is important to keep in mind that other socio-demographic factors have a stronger impact on academic results than obesity. The conclusion of the paper leads to "hard to see" psychological aspects: *" However, overweight is more easily observable by other students compared with socioeconomic characteristics, and its significant (unadjusted) association with worse academic performance can contribute to the stigma of overweight as early as the first years of elementary school."*

3 Data

3.1 Source of the data

3.1.1 Variables

All our data come from the *National Longitudinal Survey of Youth* [7] and the respondents from the *NLSY97* cohort. The American youth of our sample were born between 1980 and 1984. We also use the data of the *National Center for Health statistics (CDC)* [8] to compare the BMI (which we compute for every individual we have in our sample) percentile of our sample and the BMI percentile measured by the *CDC* in the United States. The data are only used for this purpose, consequently we do not put them in the final regression.

3.2 Tested variables

To properly test the relationship between obesity and school performance, we need to use school results as the dependent variable. Therefore we use the GPA of the individual at the end of high

[5] https://onlinelibrary.wiley.com/doi/full/10.3109/17477160902846203

[6] https://onlinelibrary.wiley.com/doi/full/10.1038/oby.2004.9

[7] https://www.nlsinfo.org/content/cohorts/nlsy97

[8] https://www.cdc.gov/growthcharts/html_charts/bmiagerev.htmmales

school in our regression. We also take the SAT test as a second dependent variable.

We take two dependent variables because we want to test the selection effect that could occur between high school and the college, and if this selection is due to the hypothetical overweight of the individual. As all respondents do not pass their exams in the same year, we have to make a distinction between two types of variables: intertemporal variables (which do not change across time) and temporal variables.

The intertemporal variables we first consider are the following: the gender of the respondent, his ethnicity, the date of birth (month + year), the height and weight of his biological parents and the education level of his father and mother.

The temporal variables are more complicated to compute. These variables change over time, therefore we have to extract them at the time of the GPA or the SAT. The temporal variables we use are the following: the general health of the respondent (he has to give a grade of how healthy he feels), his height and weight, his urbanity level (if he lives in city or not), the income of his family and the number of lessons he has dropped the past semester (absences).

To compute the temporal variables in our regression, we need to extract other variables that allow us to select at which time each variable must be taken. We consequently extract the SAT exam date and the date of the end of high school.

3.2.1 Indicators

We use the BMI in our regression as the indicator of obesity. We arbitrarily decide to consider overweight and obesity together for simplification. We use this indicator because we believe it is the easiest approach to determine if a person is overweight or not. We know that this way of estimating obesity is not perfect, but we decide to use it because it is the most commonly used estimator of obesity. The formula of the BMI is:

$$\text{BMI} = \frac{\text{Weight in kg}}{(\text{Height in m})^2}$$

To compute the index, we have to make a unit transformation in the calculation because the database gives us the height in feet (1 foot = 0.3048 meter) and the weight in pounds (1 pound = 0.453592 kg).

3.3 Computation and description of data

The GPA at the end of high school is an overall grade which is calculated for each student in his final year of school. The *NLSY* team computed this grade by weighing each grade obtained in every study field. The result is scaled between 0 and 400.

The SAT grade for college is given in two parts by the database. The first part is the SAT math score and the second part is the SAT language score. To be as precise as possible while reflecting the school level, we decide to sum up these two grades into one. This makes sense because we want to know the general level of an individual, not only his math or language level. The total SAT score is included between 400 and 1600.

3.3.1 The intertemporal variables

The variable gender equals 1 when the respondent is a man and 2 when she is a woman in our database. The date of birth is the date when the respondent was born (computed in months). The ethnicity variable has the value of 1 if the respondent is black, 2 if he is Hispanic, 3 if he is mixed race (but not Hispanic) and 4 if he is neither black nor Hispanic. The level of education (mother and father) are two variables which represent the number of education years of the mother or father.

The last intertemporal variable we need in our regression is the BMI of the biological parent (we take it for the parent 1 and 2). To compute it, we use the height and weight of the biological parents and use the classical BMI formula.

3.3.2 The temporal variables

The temporal variables need to be adjusted for two specific periods. The first adjustment period is the end of high school and the second adjustment period is the year of the SAT exam. For a specific individual who passed his SAT test in 2002, we need to select his height and weight in the year 2002, but as the year of GPA is 2001, we also need the height and weight of that year. The height and weight of the respondents are crucial because they help us to compute the BMI to approximate their level of overweight.

General health is an interesting temporal variable as well. To compute this variable, people were asked how they felt about their general health and had to give a grade between 1 and 5. This is interesting because we can use it as an approximation of the psychological status. As it has been done before in the literature, we want to add that kind of subjective variable to make our model more accurate. To perform our tests, we compute a dummy: is in bad health. This variable is 1 if the general health is evaluated under 4 and 0 if it is evaluated at 4 or 5.

Then we add an urbanity variable in our model because it could impact the grade the individual would achieve. Indeed, we think that the size of the classroom in smaller towns could be smaller which would lead to higher grades because of a larger availability of the teacher. This variable is worth 1 if the respondent lives in a rural area and 0 if he does not.

We also add the last income of the family in the pool of variables. The wealth of a family could also impact the grade of a student because less wealthy parents may provide less resources for the child's education.

The last temporal variable we add is the number of absences during the last semester of high school. On the first hand, we could interpret this variable as a proxy of a lack of student's motivation. On the other hand, we could argue that the less a child goes to school, the less he understands the lessons. This variable could be another kind of psychological variable, but we only interpret it as a standard variable so as a "not learning" effect of the absences.

3.4 Approximations

Our study suffers from a big bias which is due to the variable "height" we use in our BMI calculation. Indeed, it just represents an approximation of the real height of the respondent. The methodology used by the *NLSY* to obtain this variable could be problematic: Interviewers ask the individuals directly for their height and round it to the closest integer. As the variable is in feet, this is a huge approximation. For example, in our sample, a lot of respondents measure 5 feet, some measure 6 and nearly nobody measures less than 5 feet. Moreover, as the height was directly asked from the individuals, they could introduce bias in the estimation of their own height. In these conditions, it was difficult to differentiate them by their height and the representation of the overweight given by the BMI could be imprecise.

To have an idea of the imprecision of our BMI indicator, we compare the distribution of our BMI's sample with the real distribution of the BMI in the United States. For this purpose, we use the data of the *CDC*. Our BMI indicator is overestimated by 2 points in mean, which is huge and can lead to misinterpretations. As we want to compare individuals between themselves we decide to create our own overweight percentile frontier which depends directly on our sample and ignores the real distribution of the BMI.

As, in the health sector, the frontier of the overweight is the 85th percentile of the population sample for a child below 20 years old, we decide to keep this idea but with the percentile of our own sample. To be as precise as possible, we take the age of an individual in months because, during childhood, the BMI may vary a lot with the growth and in-between the years. To have enough observations of BMIs, the 85th percentile of a given monthly age is computed by picking the 85th percentile of a set of BMIs created by adding the BMIs of the target monthly age, the BMIs of monthly age minus one month and BMIs of the monthly age plus one month. We use a 3-month interval to have enough observations of people in the same age range and to have a more precise 85th percentile. The 3-month interval appears to give the best estimation for a given month's 85th percentile.

With this method, we hope to reduce the bias of the height approximation and keep the internal validity of our model. We know its weaknesses and the first of them is that it is arbitrary. We choose the 85th percentile limit as it is commonly used, but this criterion is already controversial. Secondly, the problem of the bias of the height does not disappear, we just compare data with that bias. Thirdly, with this kind of approximation, we know that we can not precisely quantify the effect of the BMI on the grade, but just get the direction of this effect. Despite this criticism, it appears to be the most efficient method to evaluate the impact of obesity on grades.

3.5 Preliminary tests

We first perform some preliminary tests to choose the most effective variables in our regressions. Firstly, the family income is not significant, since it probably has to be adjusted to the inflation rate. Indeed, we have a comparison problem between time periods because it is an temporal variable, which leads to insignificantly and potentially correlated error effect on the regression. Secondly, the urbanity level is insignificant, because the impact of the lifestyle or the living place is too unclear to be described. Finally, we decide to give up the BMI of the second biological parent because few respondents gave that information for both parents. As we needed just the

BMI of one of biological parent to make our instrumental variable regression, we decide not to take that second variable for the same effect.

4 Descriptive analysis

As we explain in our data section, our main variables of interest are the GPA in the final year of high school and the SAT. Before computing any regression, we first decide to plot the distributions of the dependent variables and the independent variable BMI. We also represent graphically the following relationships : GPA vs. BMI and SAT vs. BMI.

4.1 Distribution of BMI

We choose to represent the empirical distribution and the theoretical distribution using a kernel-based estimation of a p.d.f of a standard normal distribution. Our first remark is that the empirical distribution does not look normal. This can be explained by the fact that the height of the respondents was rounded to the nearest integer (in the *NLSY* database) and also because the respondents were asked the following question: *"Can you tell me approximately what is your height ?"*

These two points introduce a lot of variability in the estimated height and thus in the estimated BMI. It is unfortunately expected that the empirical distribution is not normal.

By performing a 5-number summary, we observe that half of the respondents have a BMI below 27.34 and the average BMI equals 28.32 with a standard deviation of 6.12. However, at this stage, we cannot conclude anything about the BMI because our population of interest is teenagers whose height is not stable. We therefore have to work with percentiles to identify individuals who are overweight.

4.2 Distribution of GPA

Using the same method as the one for the BMI, we can deduce that the GPA is approximately but not perfectly normal. The mean and median are respectively 297.93 and 304 with a standard deviation of 56.01.

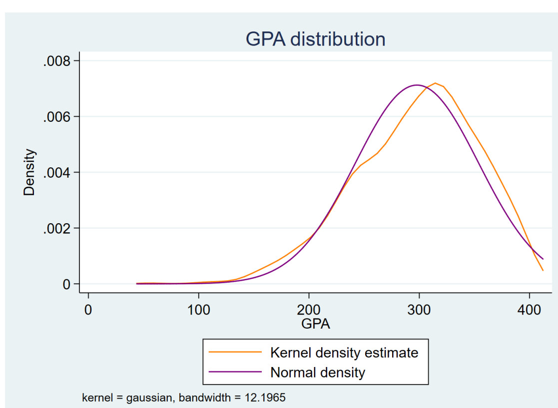

Figure 1: GPA distribution

4.3 Distribution of SAT

The Scholastic Aptitude Test or SAT is a standardized score composed of a verbal and quantitative section. Students sit this exam for a day and receive a score on a scale from 400 to 1600. Using the same method as the one for BMI and GPA, we can deduce that the SAT is approximately but not perfectly normal. The mean and median are respectively 1022.63 and 1010 with a standard deviation of 202.01.

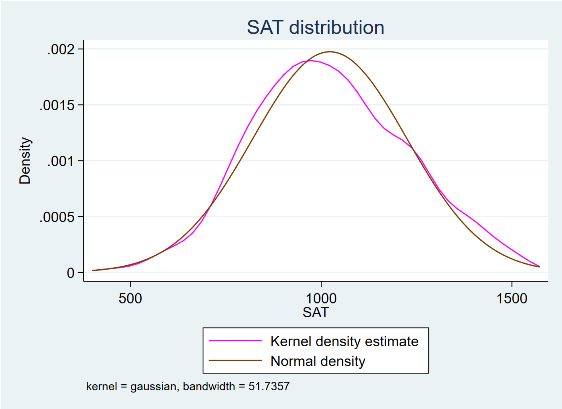

Figure 2: SAT distribution

4.4 GPA vs BMI

When we plot the GPA vs the BMI, we notice that most observations are below the arbitrary line of 30. As we mentioned before, the BMI variability is a problem. To better estimate graphically this relationship, we decided to create a plot for each rounded height (121.92, 152.4 and 182.88 cm)

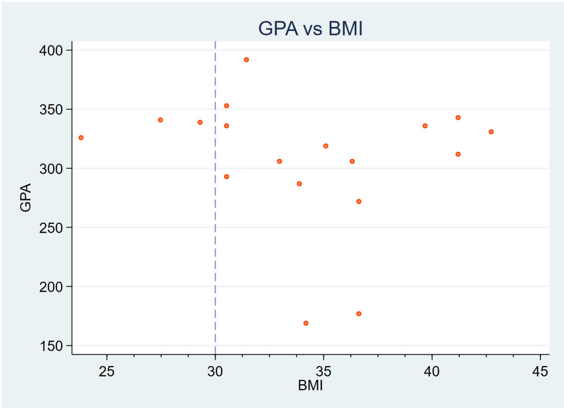

Figure 3: Individuals whose rounded height equals 121.92 cm

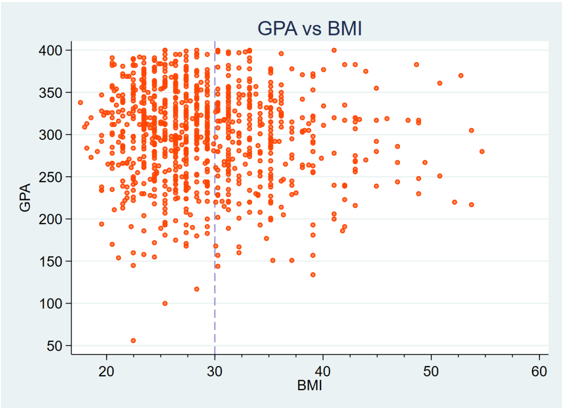

Figure 4: Individuals whose rounded height equals 152.4 cm

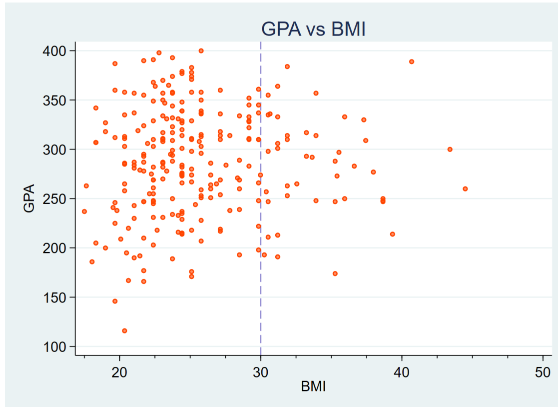

Figure 5: Individuals whose rounded height equals 182.88 cm

Looking at the 3 plots, we immediately see that most of our observations are included in the 150 to 180 cm interval.

4.5 SAT vs BMI

When we plot the SAT vs the BMI, we reach the same conclusions as for the GPA vs. BMI graphs. It is difficult to infer anything from a visual perspective. Therefore, the regressions we are going to perform will prove very useful to estimate the impact of the BMI on the GPA and SAT.

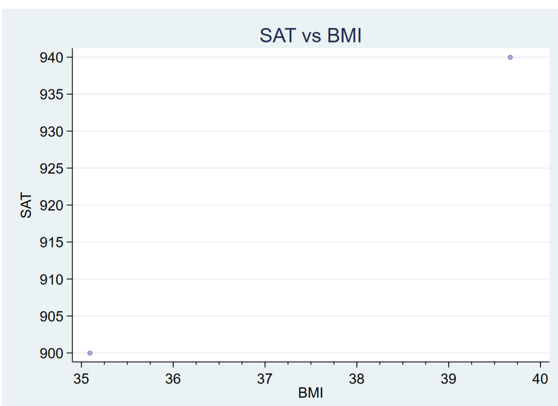

Figure 6: Individuals whose rounded height equals 121.92 cm

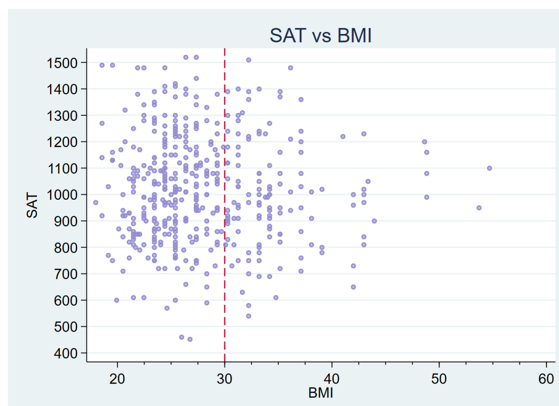

Figure 7: Individuals whose rounded height equals 152.4 cm

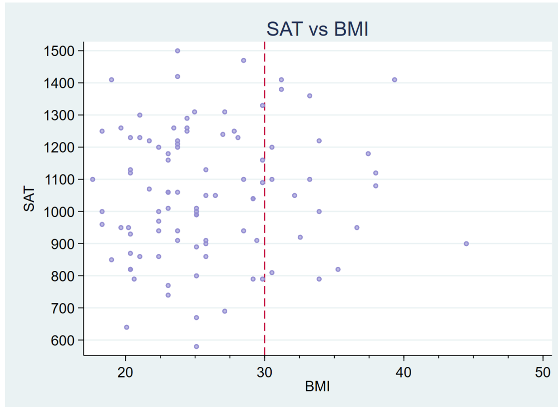

Figure 8: Individuals whose rounded height equals 182.88 cm

4.6 Correlation matrix GPA vs BMI

At first sight, we see that the correlation between our variables of interest is slightly negative. We could almost conclude that the variables are uncorrelated. However this is not surprising given the fact that we expect ex-ante the BMI not to be the main influencing variable on the GPA.

Other variables such as the number of years of education of the parents may have more impact. This is proven with a positive correlation of 0.2629 and 0.2705 with the GPA for both. The number of years of education of the mother is relatively heavily correlated with the father's.

Quite surprisingly being an African-American seems to be negatively correlated with the BMI. This is in opposition to the theory that, according to the *CDC*, a higher percentage of African-Americans suffer from obesity.

In the meantime, if one considers himself in poor health, this is slightly positively correlated with the GPA. Of course, this seems also surprising, given that people who have health problems are less likely to be able to attend school and to have good grades.

Finally, the number of absences during the GPA year is negatively correlated with the GPA and this is completely understandable. With lower attendance, the student is less likely to have good grades or even to pass the year.

4.7 Correlation matrix SAT vs BMI

If we observe the correlation between SAT and BMI, our variables of interest, we notice an an even slightly stronger negative correlation. Once again, we could almost conclude that the variables are uncorrelated. Still, this is not surprising given the fact that we expect ex-ante the BMI not to be the main influencing variable on the SAT.

Other variables such as the number of years of education of the parents seem to have more impact on the SAT than on the GPA with positive respective correlation of 0.3201 and 0.2913.

The number of years of education of the mother is relatively heavily correlated with the father's.

As for the GPA, being an African-American seems to be negatively correlated with the BMI. In the meantime, being an African-American is slightly positively correlated with the SAT.

Contrary to the GPA, considering oneself in poor health is negatively correlated with the SAT which seems logical. Finally, the number of absences during the GPA year (which is usually either the year before or during which one takes the SAT) correlates negatively with the SAT score.

5 Results

5.1 GPA

5.1.1 Standard regression model

The first results we want to discuss are regression (4) results. It has the highest level of significance and the highest adjusted R^2 (as 21,1 % of the variance of the model is explained in this regression). We test 8 control variables and the BMI as the independent variable on our dependent variable, the GPA, to explain at best the relationship between obesity and school results.

The relationship between the BMI and GPA is positive at a level of 1% of significance. To interpret this result properly, we must simultaneously consider the dummy variable "is overweight" and its negative correlation with the GPA. We interpret these results according to the following theory: as long as a person's BMI increases in the "interval of good health", her GPA will also increase and she will get better school results. However when her BMI crosses the threshold of obesity, her school results will significantly drop. At this stage in our project, we realize how ambiguous the variable BMI looks and this is a challenging interpretation for us.

We also decide to analyze the relationship between GPA and the parental education because many researches have proven that there is a significant positive relationship between these variables. The parental education appears to play a strong role on the teenager GPA and, according to our regression, the mother's education has more effect than the father's on the child's school results. Adding more general and binary variables, we test the person's gender and whether she is African American. Both dummies are significant at 1% and positively correlate with the BMI. After reading many articles about obesity in the USA, we realised that there was a higher percentage of obesity among the African-American population. Therefore we decide to create an interaction variable "is African American and overweight" to see if there is a particular impact on school performance. Our regression presents a negative relationship between the two variables but this associated coefficient is not significant. We also add a variable which we consider as highly relevant: the number of days of absence from school during the year of the GPA. The GPA and the absence variable have a negative correlation, which is coherent. If the teenager is absent from school and misses many courses, he could see a decrease in his marks. Finally, we test the variable "considers himself in poor health ". It is the closest proxy variable we found in our database to psychological health.

Surprisingly, it appears that this binary variable negatively correlates with the GPA which suggests that our regression is biased. We would tend to believe that if a person considers herself in poor health, this means that she feels bad in her life. She could suffer either physically or psychologically, she would get lower grades but our regressions do not confirm this intuition.

5.1.2 Extensions: polynomial model

Our standard model gives us an intuition of a potential quadratic relationship between BMI and GPA. In table 7.1.2, we try to model this in the regression by dropping the overweight indicator (and the interaction term between ethnicity and overweight) and including a new variable: BMI squared. We also take additional specifications trying to be more precise in our determination of the effect: the height of individuals. In specification (2) we only include people whose height is 5 feet (152.4 cm) and in specification (3) we consider people whose height is 6 feet (182.8 cm).

In regression (1), the sign of the BMI is positive, and the BMI squared's sign is negative which confirms our intuition. On the one hand, the BMI could be a "good health" indicator, helping students to perform at school. On the other hand, if the BMI is too high, the student could be considered in bad health and his school results decline. The signs of the other variables are the same as in our standard model which confirms the validity of this extension.

In specification (2) we separately analyse the situation for people whose height is rounded to the closest integer, which equals 152.4 cm in our case. We can see that the signs of the variables are the same as the polynomial model, but the coefficient associated with the BMI and BMI squared are not significant. This is probably due to the fact that the number of respondents is diminished by the specification.

In specification (3) the analysis is the same for people whose height is rounded to 182.8 cm. The signs are the same as for specifications (1) and (2) but our two main variables are only significant at a level of 5%.

In conclusion, the polynomial model is a more precise model than our standard model, but the usefulness of the specification (2) and (3) is unclear. With more data, it could be possible to test these specifications better, but with present results we recommend to use a polynomial model to explain the impact of obesity on the school results.

5.1.3 Extension: Instrumental variable model

After considering a polynomial regression, we decide to try avoiding endogeneity bias by using an instrumental variable regression (IV). As we mention it in this study's introduction, the paper **Does weight status affect academic performance? Evidence from Australian children** tries to estimate the relationship between BMI and the GPA with a IV regression using the BMI of the parent as a relevant instrument. We decide to use this helpful idea for our sample.

As for the polynomial model, we decide to split our observations according to the height (rounded to the nearest integer). We then obtain 2 main groups of data; people whose height equals 152.4 cm and those whose height equals 182.88 cm.

We decide to run 2 regressions for each group; an OLS polynomial regression and an IV regression using the subject's biological parent BMI. Furthermore, if we wanted to have both biological parent's BMIs this would reduce our sample to a very small size which is not desirable. To ensure that this instrument is relevant and exogenous, we run 2 tests. First, we regress the teenager's BMI on the biological parent's BMI to ensure that our instrument is relevant. We s that the coefficient associated with the biological parent's BMI is statistically significant at a 1% level.

Secondly, after we run our IV regression, we test the endogeneity. We obtain a very high p-value of around 0.54. Therefore we cannot reject the null hypothesis according to which the biological parent's BMI is exogenous.

We also proceed this way for the other group whose rounded height equals 182.88. For the OLS polynomial specification we reach the following conclusion. The BMI's coefficient is positive and significant at a 10% level. The BMI squared's coefficient is slightly negative and also significant at a 10% level. The coefficients of the control variables are very similar to the ones we find in the polynomial regression table.

If we now analyze the IV 2SLS regression, we find a positive but not statistically significant coefficient for the BMI variable. All the other control variable coefficients are closely the same in the IV 2SLS regression. To summarize, although our instrument is relevant and exogenous, the IV regression does not help us estimate the relationship between the BMI and the GPA because the coefficient is not statistically significant.

If we proceed the same way for individuals whose rounded height equals 182.88 cm, our OLS regression does not give us a significant coefficient for the BMI. When we use the IV regression, we get a surprisingly negative coefficient for the BMI despite having a relevant and exogenous instrument.

Our 182.88 cm sample does not include any women and the number of observations is relatively low with only 152 people. Therefore, we cannot reach any useful conclusion using the IV regression.

5.2 SAT

5.2.1 Standard regression model

As we mentioned it in our regression we also decide to test the relationship between the BMI and the SAT to confirm the impact of obesity on school performance. Surprisingly, the regression results are completely different from the results of the regression with the GPA. First, the regression of the SAT score on BMI is not significant as well as the coefficient of the variable "is overweight". There is a temporality difference between the SAT and GPA which may explain why the results are different.

The SAT is an exam taken on one day to apply for university programs whereas the GPA is a weighted average on a year. Indeed, according to our model, it could seem that a person's BMI has an effect on school performances in the long term but does not affect school capacities in the short term. In other terms, if a person is overweight, this will not play a significant role on an individual exam result. It will have an effect on the school continuity and evolution. We also suppose a selection effect could play a role. For example, if a person suffers from obesity and has bad school results, she might not take the SAT to go to the university.

As we can see in the table 7.2.1, the SAT is significantly affected by others factors such as the parent's education,gender, absences during the year of the GPA and how one considers his health.

5.2.2 Extension: polynomial model

In table 7.2.2, with the regression (1), we can not improve the significance of the model for SAT. As for the GPA model we try this because the function that drives the effect on the SAT could be polynomial. But, as the effet is unclear in the standard SAT model, we can not say what is the sign of the effect. This confirms the intuition we discussed. We cannot reject the fact that the selection effect or the temporality effect could play a role between highschool and the college.

Unsurprisingly, the specifications (2) and (3) do not help to improve the significance of the model. These specifications do not help either in the GPA polynomial model or in the SAT polynomial model. We decide to keep these specifications because they could be interesting before the test results and also to show the efficiency of our other methods.

5.2.3 Extension: instrumental variable regression

As for the GPA regression, we decide to estimate the relationship between SAT scores and the BMI using a IV regression. We proceed the exact same way as for the GPA IV regression.

We split our observations into 2 groups, the 152.4 cm group and the 182.88 cm group. We first want to ensure that the instrument, biological parent's BMI is relevant and exogenous. We regress the teenager's BMI on the biological parent's BMI. We do this for both groups (152.4 cm and 182.88 cm). The coefficients for the 2 regressions of the teenagers's BMI on the biological parent's BMI are statistically significant, respectively at a 5 and 10% level.

Performing the exogeneity test after the IV regression, we cannot reject the null hypothesis under which the biological parent's BMI is exogenous. We therefore have a relevant and exogenous instrument.

For the 152.4 cm group, both the OLS and the IV regression give us a negative but non-significant BMI coefficient. This is the same as for the polynomial model we discuss above. Apart from the number of years of the father's education, all other coefficients are not statistically significant. The adjusted R-square is also very small. Unfortunately, we cannot conclude anything, using the IV regression for the 152.4 cm group. When we look at the 182.88 cm group, once again, the BMI coefficients of both the OLS and IV regression are not significant and the small size of the sample makes our model unreliable.

However, after running these regressions, our theory would be that first of all a selection effect may play a role. Only the people who passed their GPA year can take the SAT. Therefore, if obese people fail during their GPA year, they will not try to pass the SAT. Secondly, a temporality effect may affect the relationship between the SAT and one's BMI. While the GPA is a score obtained throughout a year, the SAT is a score on one day. It is therefore not surprising that this score is very slightly affected by one's BMI.

5.3 Limits of our model

We would like to draw attention on 3 main criticisms our models suffer. First, as we mentionned, the respondents' height are rounded to the nearest feet which is an important approximation. Secondly, we did not compute the 15th percentile to take into account the underweight people's effect on the BMI. Thirdly, we use an instrument, the biological parent's BMI. As for every instrument, it is not perfect but represents a useful tool to isolate the effect of the BMI.

6 Conclusion

Does obesity affect teenager's academic performances in the USA? According to the comparison between our GPA and SAT models, obesity could affect academic performance on the long term but not significantly on the short term because of a temporality effect. Obesity could also prevent students from attending college because of a selection effect. Still other variables such as demographic and socio-economic backgrounds of a person play a strong role. Obesity is an ambiguous variable and finding the right interpretation has been our biggest challenge. We think that if a teenager's obesity affects his school results, this could possibly impact his entire academic path and professional future. Obesity does not only have long-term health implications, it can also impact life choices and societies.

7 Appendix

7.1 GPA vs BMI

7.1.1 First model

	(1) GPA	(2) GPA	(3) GPA	(4) GPA	(5) GPA in log
BMI	-0.354	0.111	0.930**	0.994***	
	(0.262)	(0.257)	(0.365)	(0.368)	
Is overweight			-17.20***	-16.96**	-0.0581**
			(6.044)	(6.907)	(0.0241)
Is a woman		22.55***	23.72***	24.41***	0.0868***
		(3.041)	(2.883)	(2.911)	(0.0110)
Number of years of father education		2.933***	2.511***	2.522***	0.00863***
		(0.593)	(0.571)	(0.571)	(0.00228)
Number of years of mother education		3.409***	2.686***	2.603***	0.00917***
		(0.653)	(0.630)	(0.629)	(0.00228)
Is African-American			9.474***	8.359***	0.0330***
			(2.915)	(3.199)	(0.0119)
Number of absences during GPA year			-1.162***	-1.164***	-0.00442***
			(0.149)	(0.149)	(0.000593)
Is African-American and overweight				-1.515	-0.0146
				(8.521)	(0.0330)
Considers himself in poor health				6.623*	0.0237*
				(3.466)	(0.0131)
BMI in log					0.118***
					(0.0381)
Constant	307.9***	200.1***	203.8***	198.4***	5.039***
	(7.686)	(12.02)	(13.46)	(13.68)	(0.135)
Observations	1206	1206	1206	1206	1206
Adjusted R^2	0.001	0.125	0.210	0.211	0.200
F	1.815	47.01	41.81	33.43	28.04

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

7.1.2 Extension: polynomial model

	(1) GPA	(2) GPA	(3) GPA
BMI	3.163**	2.811	9.407**
	(1.593)	(1.885)	(4.329)
BMI^2	-0.0472*	-0.0419	-0.158**
	(0.0249)	(0.0284)	(0.0742)
Is a woman	24.66***	24.79***	30.84*
	(2.918)	(3.861)	(16.18)
Number of years of father education	2.582***	2.648***	2.693**
	(0.575)	(0.666)	(1.291)
Number of years of mother education	2.586***	2.420***	3.318**
	(0.632)	(0.720)	(1.433)
Is African-American	8.192***	9.377***	-0.492
	(3.024)	(3.483)	(6.559)
Number of absences during GPA year	-1.166***	-1.078***	-1.352***
	(0.150)	(0.172)	(0.301)
Considers himself in poor health	5.990*	4.596	13.15
	(3.475)	(3.942)	(8.120)
Constant	173.5***	179.4***	79.56
	(26.84)	(33.44)	(61.20)
Observations	1206	932	256
Adjusted R^2	0.209	0.183	0.253
F	36.87	25.33	8.766

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

(1) = all data (2) = 152.4 (3) = 182.8

7.1.3 Extension: IV for individuals whose height = 152

	BMI
IMCBiol	0.146***
	(0.0457)
Is a woman	-4.539***
	(0.508)
Number of years of father education	-0.160
	(0.116)
Number of years of mother education	-0.0326
	(0.106)
Is African-American	-0.586
	(0.483)
Number of absences during GPA year	-0.0139
	(0.0217)
Considers himself in poor health	-1.306**
	(0.621)
Constant	30.68***
	(2.117)
Observations	508
Adjusted R^2	0.156
F	14.51

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Test of relevance of instrument "BMI of biological parent"

	OLS	IV 2SLS
BMI	4.431*	0.884
	(2.480)	(2.389)
BMI^2	-0.0620*	
	(0.0373)	
Is a woman	23.20***	23.30**
	(5.215)	(11.57)
Number of years of father education	3.473***	3.475***
	(0.914)	(0.971)
Number of years of mother education	2.434**	2.511**
	(0.989)	(1.004)
Is African-American	13.63***	14.03***
	(4.492)	(4.629)
Number of absences during GPA year	-1.200***	-1.202***
	(0.223)	(0.227)
Considers himself in poor health	5.226	6.379
	(5.386)	(6.387)
Constant	141.7***	188.1**
	(43.97)	(86.48)
Observations	508	508
Adjusted R^2	0.217	0.212
F	19.41	

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

7.1.4 Extension: IV for individuals whose height = 182

	BMI
IMCBiol	0.192***
	(0.0652)
Is a woman	0
	(.)
Number of years of father education	0.294*
	(0.174)
Number of years of mother education	-0.172
	(0.164)
Is African-American	-1.618*
	(0.929)
Number of absences during GPA year	-0.0277
	(0.0305)
Considers himself in poor health	-1.508
	(1.256)
Constant	20.46***
	(3.269)
Observations	152
Adjusted R^2	0.094
F	2.969

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Test of relevance of instrument "BMI of biological parent"

	OLS	IV 2SLS
BMI	8.117	-3.816
	(5.015)	(3.167)
BMI^2	-0.133	
	(0.0851)	
Is a woman	0	0
	(.)	(.)
Number of years of father education	1.918	2.996
	(1.551)	(1.936)
Number of years of mother education	3.622**	2.850
	(1.608)	(1.940)
Is African-American	8.940	2.320
	(8.334)	(11.09)
Number of absences during GPA year	-0.955**	-1.041**
	(0.405)	(0.431)
Considers himself in poor health	19.13*	13.97
	(9.981)	(11.77)
Constant	89.50	309.4***
	(75.81)	(90.19)
Observations	152	152
Adjusted R^2	0.181	0.003
F	5.037	

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

7.2 SAT vs BMI

7.2.1 First model

	(1) SAT	(2) SAT	(3) SAT	(4) SAT	(5) SAT in log
BMI	-0.835	-0.316	-0.145	-0.471	
	(1.346)	(1.277)	(2.082)	(2.083)	
Is overweight			-0.126	-38.32	-0.0375
			(35.72)	(36.24)	(0.0361)
Is a woman		-30.42*	-27.02	-29.04*	-0.0263
		(16.84)	(16.88)	(16.87)	(0.0171)
Number of years of father education		14.73***	14.46***	14.18***	0.0131***
		(3.522)	(3.535)	(3.534)	(0.00358)
Number of years of mother education		12.13***	12.24***	11.91***	0.0126***
		(3.997)	(3.996)	(3.960)	(0.00415)
Is African-American			14.43	0.159	0.00109
			(16.60)	(18.17)	(0.0185)
Number of absences during SAT year			-1.954**	-1.935**	-0.00181*
			(0.929)	(0.934)	(0.000935)
Is African-American and overweight				87.23*	0.0818*
				(45.15)	(0.0449)
Considers himself in poor health				-29.31*	-0.0311*
				(17.30)	(0.0171)
BMI in log					-0.00825
					(0.0604)
Constant	1045.7***	667.6***	670.1***	716.5***	6.622***
	(38.52)	(61.61)	(74.86)	(77.29)	(0.206)
Observations	536	536	536	536	536
Adjusted R^2	-0.001	0.117	0.118	0.126	0.116
F	0.385	17.52	11.08	10.46	10.33

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

7.2.2 Extension: polynomial model

	(1) SAT	(2) SAT	(3) SAT
BMI	-2.756	-5.681	9.736
	(8.106)	(9.612)	(23.81)
BMI^2	0.0381	0.0789	-0.118
	(0.121)	(0.139)	(0.405)
Is a woman	-30.91*	-22.28	-193.0***
	(16.90)	(22.69)	(48.95)
Number of years of father education	14.02***	11.73***	25.41***
	(3.560)	(4.037)	(6.995)
Number of years of mother education	12.05***	12.30***	11.97
	(3.994)	(4.445)	(9.026)
Is African-American	15.65	20.11	-6.912
	(16.91)	(18.84)	(42.25)
Number of absences during SAT year	-2.029**	-1.877	-2.034
	(0.926)	(1.211)	(1.375)
Considers himself in poor health	-28.76*	-25.49	-51.30
	(17.34)	(19.20)	(42.43)
Constant	743.0***	806.2***	413.6
	(136.9)	(168.0)	(351.0)
Observations	536	437	97
Adjusted R^2	0.121	0.095	0.202
F	9.796	6.209	.

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

(1) = all data (2) = 152.4 (3) = 182.8

7.2.3 Extension: IV for individuals whose height = 152

	BMI
IMCBio1	0.171**
	(0.0721)
Is a woman	-5.671***
	(0.756)
Number of years of father education	-0.0335
	(0.137)
Number of years of mother education	-0.248*
	(0.146)
Is African-American	0.430
	(0.675)
Number of absences during SAT year	-0.0109
	(0.0468)
Considers himself in poor health	-0.584
	(0.736)
Constant	30.33***
	(3.033)
Observations	250
Adjusted R^2	0.221
F	10.35

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Test of relevance of instrument "BMI of biological parent"

	OLS	IV 2SLS
BMI	-11.93	-12.18
	(12.50)	(13.22)
BMI^2	0.188	
	(0.175)	
Is a woman	-13.95	-77.98
	(29.85)	(82.50)
Number of years of father education	15.11***	14.23**
	(5.689)	(6.053)
Number of years of mother education	11.11	8.329
	(6.925)	(7.626)
Is African-American	10.40	18.37
	(24.67)	(25.51)
Number of absences during SAT year	0.383	0.0706
	(1.832)	(1.907)
Considers himself in poor health	-7.726	-16.56
	(25.22)	(27.22)
Constant	827.6***	1085.8**
	(222.3)	(487.1)
Observations	250	250
Adjusted R^2	0.090	-0.022
F	4.242	

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

7.2.4 Extension: IV for individuals whose height = 182

	BMI
IMCBiol	0.236*
	(0.126)
Is a woman	0
	(.)
Number of years of father education	0.0255
	(0.258)
Number of years of mother education	0.0474
	(0.306)
Is African-American	-4.669***
	(1.532)
Number of absences during SAT year	-0.00374
	(0.0390)
Considers himself in poor health	0.739
	(1.715)
Constant	19.85***
	(4.983)
Observations	59
Adjusted R^2	0.176
F	2.397

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Test of relevance of instrument "BMI of biological parent"

	OLS	IV 2SLS
BMI	13.99	-18.99
	(26.96)	(17.39)
BMI^2	-0.198	
	(0.454)	
Is a woman	0	0
	(.)	(.)
Number of years of father education	14.37*	13.66*
	(7.796)	(8.281)
Number of years of mother education	25.95**	28.23**
	(10.95)	(12.51)
Is African-American	21.48	-84.89
	(56.03)	(100.0)
Number of absences during SAT year	-2.685	-3.102**
	(1.638)	(1.243)
Considers himself in poor health	-70.12	-45.73
	(54.99)	(75.11)
Constant	349.4	1092.2**
	(399.0)	(474.6)
Observations	59	59
Adjusted R^2	0.242	-0.074
F	7.025	

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

7.3 GPA vs BMI correlation Matrix

	BMI	GPA	# of years of education of father	# of years of education of mother	Is African-American	Considers himself/herself in poor health	# of absences during GPA year
BMI	1						
GPA	-0.0386	1					
# of years of education of father	-0.0755	0.2629	1				
# of years of education of mother	-0.0591	0.2705	0.6184	1			
Is African-American	-0.1208	0.1061	0.1	0.0587	1		
Considers himself/herself in poor health	-0.1191	0.0663	0.0683	0.0874	0.2448	1	
# of absences during GPA year	-0.022	-0.3145	-0.1217	-0.1541	-0.0689	-0.0165	1

7.4 5-number summary for GPA vs BMI

					Quantiles			
Variable	n obs	Mean	S.D.	Min	0.25	0.5	0.75	Max
BMI	1206	28.32	6.12	17.5	23.73	27.34	31.25	54.68
GPA	1206	297.93	56.01	56	261	304	339	400

7.5 SAT vs BMI correlation Matrix

	BMI	SAT	# of years of education of father	# of years of education of mother	Is African-American	Considers himself/herself in poor health	# of absences during GPA year
BMI	1						
SAT	-0.0243	1					
# of years of education of father	-0.0994	0.3201	1				
# of years of education of mother	-0.0911	0.2913	0.6001	1			
Is African-American	-0.0757	0.0514	0.0472	-0.0107	1		
Considers himself/herself in poor health	-0.053	-0.0874	-0.1037	-0.097	0.0791	1	
# of absences during GPA year	-0.0004	-0.0873	-0.0356	-0.0259	-0.0127	-0.0513	1

7.6 5-number summary for SAT vs BMI

					Quantiles			
Variable	n obs	Mean	S.D.	Min	0.25	0.5	0.75	Max
BMI	536	27.59	5.88	17.63	23.44	26.37	30.86	54.68
SAT	536	1022.63	202.01	452	870	1010	1160	1520