Obesity in California:

An Econometric Exploration

By Anaelia Ovalle

The obesity epidemic in the United States represents a critical public health issue that has the potential to incur major healthcare costs because of the substantial risks associated with excess body fat. Whereas many recognize the significant risk of cardiovascular disease and diabetes mellitus associated with excess body fat, a myriad of other health problems can accompany overweight and obesity, potentially leading to early morbidity and mortality. Public recognition of obesity as an important health crisis, and not simply a matter of cosmetics or lifestyle choice, is clearly needed. A greater awareness of the health risks associated with excess weight will facilitate more frequent obesity screenings and discussions about healthy weight management that have the potential to result in a greater commitment of healthcare resources to effective obesity prevention and management strategies[1].

I. Introduction

A. California

As of 2015, more than one-third of adults in the United States are obese. In California, 24.7% of its resident are obese. These alarming statistics raise concerns and questions for policy makers and citizens alike. Obesity-related conditions include heart disease, stroke, type 2 diabetes and certain types of cancer, leading to an annual medical cost of $1, 429 higher than those of normal weight[2]. Speculation to the question of its cause lead us to dive into not just what the condition is but how it is affected by everyday factors citizens may tend to overlook. Within this report, I seek to shed light on certain societal characteristics and conditionals that may contribute to the disease on a more local scope within the state of California.

For adults, the range of obesity is determined using weight and height to calculate a number called the body mass index. BMI is used because, for most people, it correlates with

[1] http://www.ncbi.nlm.nih.gov/pubmed/19410676
[2] http://www.cdc.gov/obesity/data/adult.html

their amount of body fat. For example, an adult who has BMI of 30 or above is considered obese[3]. Not to be confused with an exaggerated weight, a 21-year old male who is is 6feet tall and weighing 230 pounds is considered to have a BMI within the obesity range. Among children of the same age and sex, obesity is defined on CDC growth charts as a BMI at or above the 95th percentile.

B. The Data

For this analysis, I used cross-sectional data found in most recent sources within the United States Agriculture Department[4] and county health rankings sponsored by the University of Wisconsin Population Health Institute[5]. For my first model, I was interested in testing the effect and relationship of 11 variables on adult obesity rates in California. In my second model, I look to test these same variables on child obesity rates in California. Variables and their descriptions using the USDA's database documentation are listed below.

1. Food Environment Index in 2016
    1. Indicator of access to healthy foods - 0 is worst, 10 is best.
2. % population that are food insecure, 2016
    1. Prevalence of household-level food insecurity by state. Food-insecure households were unable, at times during the year, to provide adequate food for one or more household members because the household lacked money and other resources for food. For most food-insecure households, inadequacy was in quality and variety of foods; for about a third—those with very low food security—amounts were also inadequate.
3. % population that have low access to stores, 2010
    1. Percentage of people in a county living more than 1 mile from a supermarket, supercenter or large grocery store if in an urban area, or more than 10 miles from a supermarket or large grocery store if in a rural area.
4. Median household income, 2016
    1. Median income by household: income level that divides county households in half, one half with income above the median and the other half with income below the median; includes income of all household members 15 years old or older.
5. Count of recreational facilities, 2016

---

[3] http://www.cdc.gov/ncbddd/disabilityandhealth/obesity.html
[4] http://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads.aspx
[5] http://www.countyhealthrankings.org/app/california/2016/downloads

    1. Number of "fitness and recreation centers" in a county, where "fitness and recreation centers" (defined by North American Industry Classification System (NAICS) code 713940) are establishments primarily engaged in operating fitness and recreational sports facilities featuring exercise and other active physical fitness conditioning or recreational sports activities, such as swimming, skating, or racquet sports.
6. Metropolitan, 2010
    1. Classification of counties by metro or non-metro definition, where 1=metro county; 0=non-metro county; metro areas include all counties containing one or more urbanized areas: high-density urban areas containing 50,000 people or more; metro areas also include outlying counties that are economically tied to the central counties, as measured by the share of workers commuting on a daily basis to the central counties. Non-metro counties are outside the boundaries of metro areas and have no cities with 50,000 residents or more.
7. Population Density, 2016
    1. Current population of the county based on current U.S Census data per square mile.
8. Soda price 2010
    1. Regional average price of sodas relative to the national average price. Sodas include carbonated diet and caloric-sweetened beverages.
9. Fast food chains per 1000 people 2012
    1. The number of limited-service restaurants in the county per 1,000 county residents. Limited-service restaurants (defined by North American Industry Classification System (NAICS) code 722211) include establishments primarily engaged in providing food services (except snack and nonalcoholic beverage bars) where patrons generally order or select items and pay before eating.
10. High School Graduation Rate, 2016
    1. Number of students expected to graduate.
11. % Some College 2016
    1. Adults age 25-44 with some post-secondary education

C. Methods

By implementing the ordinary least squares method, I seek to minimize the difference between the observed and predicted residuals within my linear regression model. In order to highlight a potential relationship between adult obesity and each variable, I focus on finding statistical significance on each of the variable's coefficients, considering a p-value less than or equal to 0.10. If I do not find statistical significance, I look for joint significance among the

variables using the F-test. The goal at each point is to maximize the value of the R-Squared, known as the percent of explanatory power the model provides against the total residuals.

II. Descriptive Statistics

A. Summary Statistics

Table 1 contains a summary of each variable, detailing the count of observations, min and max value, mean, standard deviation, and skew. Later, we correct for skew by transforming each variable whose distance from 0 is ±0.5.

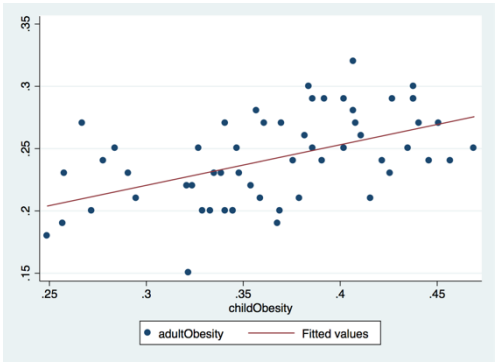
Table 1. Summary Statistics

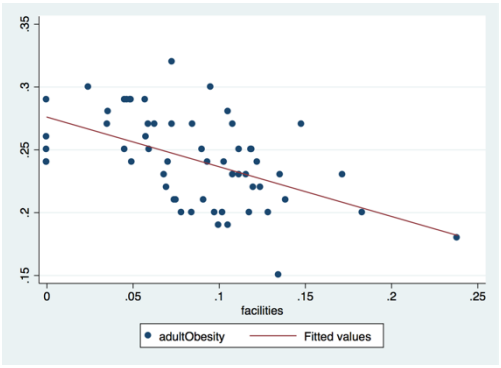
	count	min	max	mean	sd	variance	skewness
adultobesity	56	.15	.32	.2423214	.0363314	.00132	-.0969019
childobesity	56	.249	.469	.3665536	.057133	.0032642	-.2690827
fei	56	4.2	8.6	7.173214	.8795642	.7736331	-.7377464
foodinsec	56	10	21	15.41071	2.640777	6.973701	.0084952
lowaccess	56	3.31486	48.3044	18.13144	9.264357	85.82832	1.547373
income	56	34961	100806	56260.23	15819.77	2.50e+08	1.01745
facilities	56	0	.238217	.0886265	.0447174	.0019996	.4889198
metro	56	0	1	.6607143	.4777518	.2282468	-.6788829
sodaprice	56	1.08949	1.16442	1.123494	.0269244	.0007249	.0926667
fastfood	56	7	7204	501.8393	1074.811	1155218	4.689779
highschool	56	81	96	87.39286	3.901881	15.22468	.2198046
somecollege	56	39.6	85	59.24464	10.3175	106.4509	.1001404
popden	56	1.8	3575.3	402.6964	753.9257	568404	2.787323

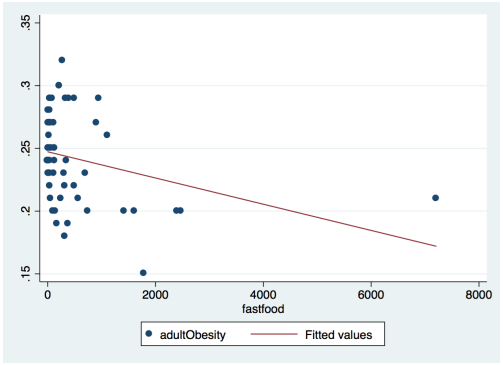
B. Scatter Plots

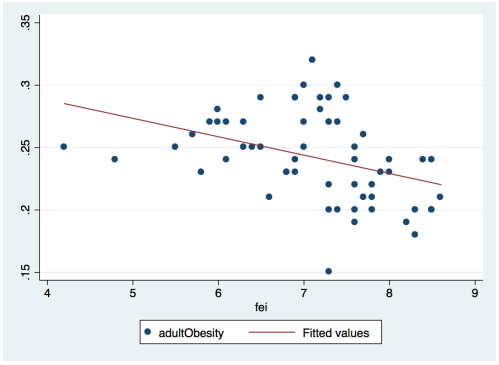
Below is Figure 1, containing the scatter plots for each variable plotted against adult obesity.

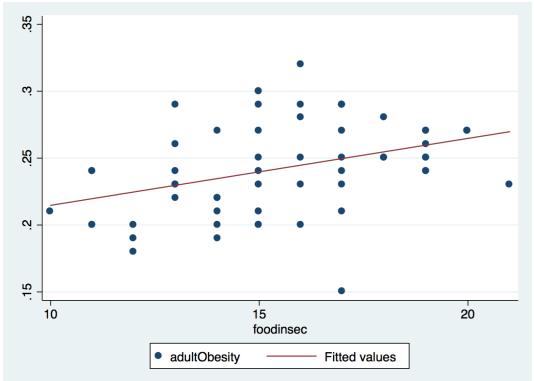
Figure 1. Scatter plot of each variable

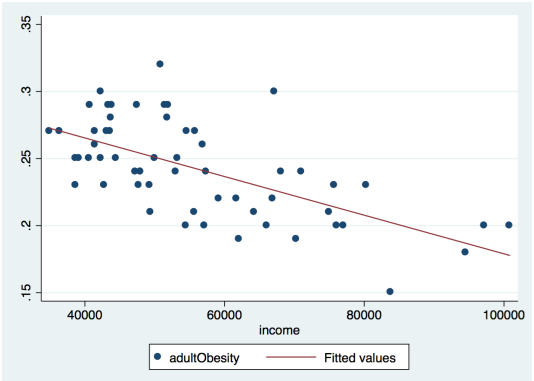


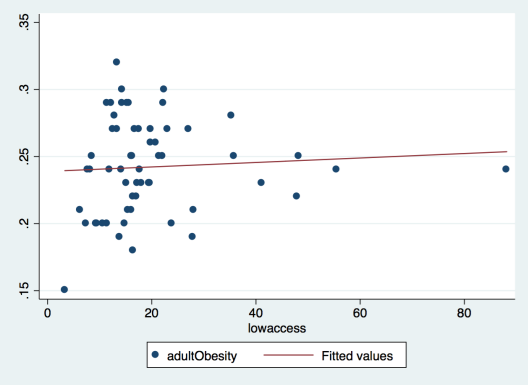


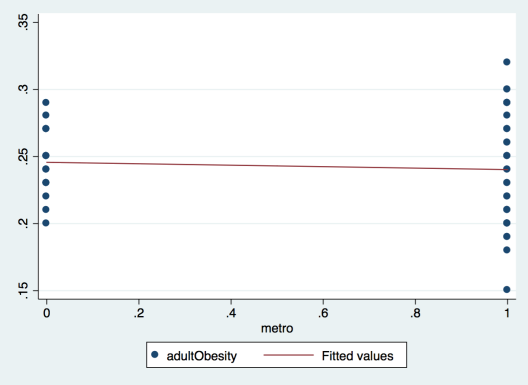


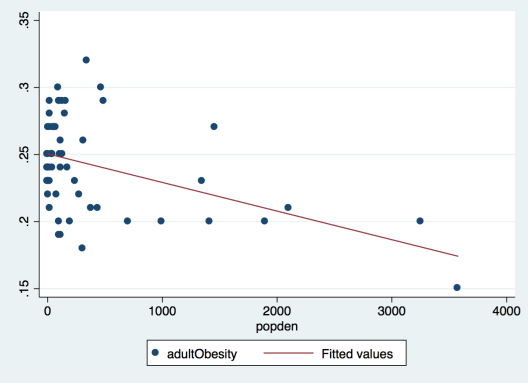


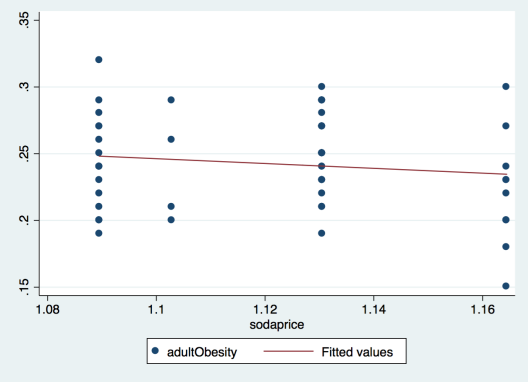


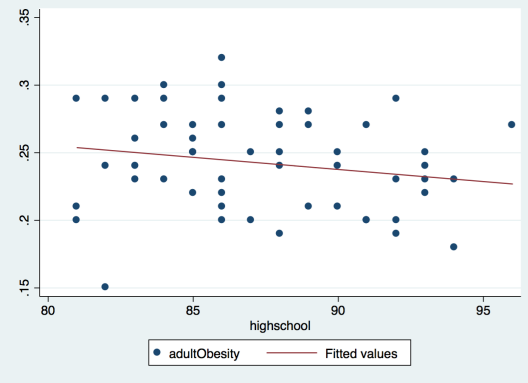


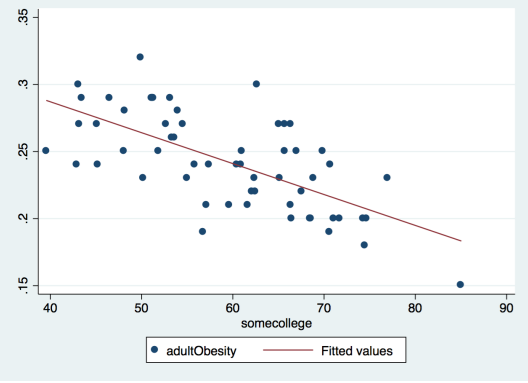












C. Transformations

Before regressing adult obesity as a function of the other 12 variables, I first used the summary statistics and scatterplots to tweak my model. First, I transformed variables that had a skew or lack of symmetry of the frequency distribution about the mean (±0.5). This resulted in 7 transformations. The variables whose distance from 0 was negative were squared to increase the

dispersion of larger values and reduce the relative dispersion of smaller values. Likewise, variables who showed positive skew were reduced to their log function in order to decrease the dispersion of larger values and likewise increase the relative dispersion of smaller values. Table 2 shows the results of such transformations. The scatter plots also indicate outliers, further adjusting the initial regression by not containing variables with the outlier values, as that would skew the regression.

Table 2. Variables corrected for skew

	skewness
childobesity	**-.2800491**
childobesi~2	**.0003062**
fei	**-.7861687**
fei2	**-.3780345**
lowaccess	**3.039017**
llowaccess	**.2023698**
income	**1.044663**
lincome	**.5457161**
metro	**-.6249324**
metro2	**-.6249324**
fastfood	**4.72734**
lfastfood	**.1125318**
popden	**2.816657**
lpopden	**-.1175134**

III. Results

A. The initial regression

Surprisingly enough, the initial regression of adult obesity on the other 12 variables (Table 3) results in very little significant variables. Only a college education and population density seem to hold a relevant relationship with adult obesity. The negative coefficient on college education depicts a subtle but significant negative relationship with adult obesity implying that someone receiving a college education is less likely to be obese. On the other hand, there is a positive and significant coefficient on population density. This tells us that the greater population density, the greater the presence of individuals who are obese. Logically, a higher population density naturally leads to a broader scope of individuals within the area, including those of varying degrees of health.

Table 3. Regression output for the initial regression.

VARIABLES	(1) Adult Obesity
childobesity2	-0.121
	(0.187)
fei2	0.00107
	(0.00128)
foodinsec	0.00436
	(0.00576)
llowaccess	0.0168
	(0.0124)
lincome	-0.0678
	(0.0417)
facilities	-0.152
	(0.114)
metro2	0.00490
	(0.0132)
sodaprice	0.0808
	(0.149)
lfastfood	-0.00420
	(0.00485)
highschool	0.000494
	(0.00134)
somecollege	-0.00186***
	(0.000661)
lpopden	0.00943*
	(0.00537)
Constant	0.793
	(0.521)
Observations	56
R-squared	0.623
Adj. R-squared	0.518
Insignificant variables (I)	10
Degrees of Freedom (dF)	43
F (I, dF)	0.056
Mean VIF	7.19

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

While these coefficients are interesting, what's more illustrative is the lack of significant coefficients. To confirm that all other 10 variables are truly insignificant, I ran an F-test in order see whether or not there was joint-significance within my model. The data in Table 3 allows me to reject the claim that all tested coefficients are statistically insignificant. While this model does not capture the full story of factors influencing adult obesity, the joint influence of these variables seem to contribute to the unraveling of elements associated with adult obesity.

An R-squared of 62.3% shows that model is able to account for 62.3% of the total residuals. While possible multicollinearity does not affect the model's fit of the data, it is equally important to observe the degree of its presence as it causes coefficient estimates to be unstable and difficult to interpret. Indicated by a high variance inflation factor and high correlations between variables, severe multicollinearity results in each coefficient partially capturing the effect of other variables. Luckily, Table 3 points to a mean VIF of 7.19, a relatively low level of mulitcollinearity.

A. Specifying the Model

Given the weak results from the original regression, I set out to re-specify my model. Since adult obesity seemed to be unaffected by most of my variables, I instead decided to regress the rate of child obesity as a function of the same 12 variables used in the original model. With this, the data could potentially belong to a slightly different narrative.

The first regression labeled "Child Obesity v.1" in Table 4 has the child obesity rate as the dependent variable and places all other variables, including adult obesity, as independent variables. Doing this simple switch, the results change drastically. The R-squared and adjusted R-squared surge about 20%, indicating that the choice to switch variables was indeed a good one! We now see 6 new variables that are statistically significant. The food environment index is significant at the 5% level with a positive coefficient. This is actually pretty surprising, since one would assume that an increase in a location's food quality would influence its residents to consume more nutritious foods. We also see income with a negative coefficient with a 1% p-value, confirming that the higher an individual's income is, the less it is that their child would be obese. A higher income allows an individual to spend more on high quality foods. Given their resources, they are also more likely to be exposed to a healthier lifestyle. The number of facilities per square mile also seem to be significant at the 5% level. Its negative relationship with the child obesity rate implies that recreational facilities allow kids to stay active and therefore less likely to put on excess weight. However, what seems strange is that soda price holds a significant and positive value in regards to the child obesity rate. Since this does not make sense intuitively, I suspect omitted variable bias. I will keep it in the model for now but consider its potential sneaky bias. The last factor to consider is the coefficient on high school. We see that it is negative and significant at a 1% p-value. This tells a similar story to the coefficient on college education. The more an individual is involved in obtaining an education, the more likely it is that they will learn how to maintain a healthy lifestyle (i.e. a healthy diet and lifestyle) and extend it to their family.

Table 4. Regression output for all models

VARIABLES	(1) Adult Obesity	(2) Child Obesity v.1	(3) Child Obesity v.2	(4) Child Obesity v.3
childobesity2	-0.121			
	(0.187)			
fei2	0.00107	0.00285**	0.00270*	0.00237***
	(0.00128)	(0.00138)	(0.00140)	(0.000646)
foodinsec	0.00436	0.00256	0.00119	
	(0.00576)	(0.00650)	(0.00676)	
llowaccess	0.0168	-0.000556	-0.00146	-0.00381
	(0.0124)	(0.0142)	(0.0143)	(0.0114)
lincome	-0.0678	-0.151***	-0.157***	-0.158***
	(0.0417)	(0.0425)	(0.0433)	(0.0377)
facilities	-0.152	-0.306**	-0.278**	-0.309**
	(0.114)	(0.124)	(0.130)	(0.123)
metro2	0.00490	0.0111	0.0133	0.0104
	(0.0132)	(0.0147)	(0.0150)	(0.0144)
sodaprice	0.0808	0.499***	0.471***	0.501***
	(0.149)	(0.151)	(0.157)	(0.150)
lfastfood	-0.00420	-0.00917*	-0.00904*	-0.00930*
	(0.00485)	(0.00529)	(0.00532)	(0.00523)
highschool	0.000494	-0.00384***	0.00193	-0.00373***
	(0.00134)	(0.00140)	(0.00757)	(0.00136)
somecollege	-0.00186***	-0.00193**	0.00646	-0.00179***
	(0.000661)	(0.000740)	(0.0108)	(0.000637)
lpopden	0.00943*	0.0126**	0.0114*	0.0129**
	(0.00537)	(0.00599)	(0.00621)	(0.00588)
adultobesity		-0.0837	-0.0655	-0.0765
		(0.171)	(0.173)	(0.168)
high_college			-9.55e-05	
			(0.000123)	
Constant	0.793	1.742***	1.359*	1.875***
	(0.521)	(0.550)	(0.741)	(0.431)
Observations	56	56	56	56
R-squared	0.623	0.807	0.810	0.807
Adj. R-squared	0.517	0.754	0.751	0.758
Insignificant variables (I)	10	4	7	3
Degrees of Freedom (dF)	43	43	42	44
F (I, dF)	0.056	0.927	0.001	0.865
Mean VIF	7.19	6.62	156.06	3.67
Variables omitted	NO	NO	NO	YES
Interaction Term	NO	NO	YES	NO

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

Deciding to look further into the relationships in the model, I was interested in asking how a high school education influences the marginal effect of a college education on child obesity. In order to analyze this complex relationship, I created an interaction term that I would add to my model (Child Obesity v.2). As shown in Table 4, the interaction term has a p-value greater than 10%, so I cannot say that it is significant. The adjusted R-squared also decreased so I have decided to leave this out of my finalized model.

The same regression had a whopping mean VIF of 156.06. In hopes of reducing the variance inflation factor, not only did I leave out the interaction term but I also decided to omit the variable foodinsec. Upon re-reading the USDA documentation, I learned that the food environment index was actually comprised of data relevant to food insecurity. Leaving this variable out (Child Obesity v. 3) actually resulted in a slightly higher adjusted R-squared as shown in the last column of Table 4. Upon further examination, I found this regression to maximize the explanatory power of the model. An F-test between the 3 insignificant variables low access, metro2, and adult obesity result in no joint significance.

IV. Tests

A. Heteroskedasticity

To test whether or not error term of my final model varied across samples, I implemented a simple Breusch-Pagan test. With a p-value well above the 10% significance level, I can confirm that my data is indeed homoscedastic. The results of the tests are shown in Figure 2 below.

Figure 2. *Breusch-Pagan Test for the final model.*

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of childobesity

         chi2(1)      =     0.02
         Prob > chi2  =   0.8965
```

B. Endogeneity

For my final model, there are a few variables I believe are truly exogenous. Let us recall that a variable is truly exogenous if it is not systematically affected by the changes in other variables of the model (no correlation with the error term) and is purely determined by factors outside of the model. Such variables would be 1) The metropolitan dummy variable 2) population density 3) A college education and 4) Income.

Likewise, my final model contains truly endogenous variables as well. To review, an endogenous variable is one that is determined by the effects of other variables in the model. They are not predetermined, nor fixed in repeated samples. The causes may be simultaneity within the model, measurement errors, or omitted variable bias. These variables are 1) The adult obesity rate 2) The food environment index 3) The soda price 4) Fast food 6) Recreational facilities and 7) The school graduation rate.

In order to see the true effect of an endogenous variable on the dependent term, an instrumental variable can be used in the model. A good instrumental variable should be relevant, exist outside of the model, be highly correlated with the variable that is to be replaced, and have no correlation with the error term. Outside of the model, it would be interesting to perhaps add SAT scores as an instrumental variable since it is highly correlated with high school graduate rate.

V. Conclusion

By developing an econometric model (Child Obesity v.3) that approaches child obesity at a more local level, I wish to help others identify characteristics that are statistically relevant to this health issue. By beginning to debunk its influential factors, a conversation can be started amongst parents and policy makers alike. While the initial model was based on adult obesity, we instead focused on the outcomes pertaining to child obesity since they provided much more relevant results. We saw at a 1% significance level that a person's income and education level negatively impacted the rate of child obesity. Similarly, we also saw a high significance in the number of recreational facilities. Acknowledging the negative relationship with the child obesity rate, the data supports programs that keep students in school and involved in afterschool activities. On the other hand, factors like the percentage of people that have low access to superstores or the difference between residents of metropolitan and rural counties are not significant in understanding the child obesity rate in California.

Decreasing the child obesity rate begins by understanding its causal factors and dynamics. As medical entities strive to continue providing new research on the topic, our understanding of child obesity and its prevention will be improved.