Exploratory data analysis of health metrics using SAS; Benjamin Cox, Dean English, Mathias Gerle, Nicholas Hanson; Spring 2024

# 

# Datasets Merged

* 2017-2020 Demographic Variables and Sample Weights – P\_DEMO
* 2017-2020 Body Measures – P\_BMX
* 2017-2020 Complete Blood Count with 5-Part Differential in Whole Blood – P\_CBC

# Data source

NHANES (National Health and Nutrition Examination Survey) – CDC

# Definitions of Variables

## Demographics

**RIDAGEYR – Age in years**

**RIDRETH3 – Race/Ethnicity**

DMDEDUC2 - Education level - Adults 20+

# Variables

BMXWT - Weight (kg)

BMXHT - Standing Height (cm)

BMXWAIST - Waist Circumference (cm)

LBXWBCSI – White blood cell count

LBXRBCSI – Red blood cell count

LBXHGB – Hemoglobin

BMXWT - Weight (kg) \* BMXHT - Standing Height (cm) - BMI

# Introduction

We are interested in seeing how an American’s average health and build changes by looking through the scope of several variables such as age, race/ethnicity, and education level to see how their weight, height, waist circumference, white blood cell, and red blood cell count. We will use SAS programming software to analysis and compare variables within datasets. Additionally, we will produce graphics and statistics and then discuss them within this report.

# Exploratory analyses:

## Data Preparation and manipulation steps

Our preparation started with discussion of variables and topics we wanted to look at from the NHANES (National Health and Nutrition Examination Survey) database. After deciding on our sample and variables of interest we downloaded the data and uploaded it into the SAS software. Then merged the data into one data set to compare linked by sequence number (SEQN).

The datasets we merged for the data were the 2020 Demographic Variables and Sample “Weights – P\_DEMO”, “2017-2020 Body Measures – P\_BMX”, and “2017-2020 Complete Blood Count with 5-Part Differential in Whole Blood – P\_CBC” Using these datasets, we looked at how Age in years, Race/Ethnicity, and the education level in adults ages 20 and above these variables.

# Variables

BMXWT - Weight (kg)

BMXHT - Standing Height (cm)

BMXWAIST - Waist Circumference (cm)

LBXWBCSI – White blood cell count

LBXRBCSI – Red blood cell count

LBXHGB – Hemoglobin

BMXWT - Weight (kg) \* BMXHT - Standing Height (cm) - BMI

A grid of numbers and symbols

Description automatically generated with medium confidence

Merged Data Screenshot, the data was merged using PROC SQL

A screenshot of a computer code

Description automatically generated

# Descriptive Statistics

### Figure [1] - Education Level-BMI

This graph depicts distribution of BMI for specific education levels using PROC GLM.

Reminder:

1. Less than 9th Grade
2. 9-11th grade (includes 12th grade with no diploma)
3. Highschool Graduate/GED or equivalent
4. Some college or AA degree
5. College Graduate or above

The graph depicts a box and whisker plot for the chosen variables. X-axis including education level, and Y-axis showing BMI distributions in comparison to each other.

A screenshot of a computer screen

Description automatically generated

### Figure [2] - Education Level-Waist Circumference

This graph depicts distribution of waist circumference for specific education levels using PROC GLM.

Reminder:

1. Less than 9th Grade
2. 9-11th grade (includes 12th grade with no diploma)
3. Highschool Graduate/GED or equivalent
4. Some college of AA degree
5. College Graduate or above

The graph is a box and whisker plot for the chosen variables. X-axis education level (1-5), and waist circumference on the Y-axis. Objectively, education level 3 and 4 have the highest mean waist circumference. Level 4 especially has many more outliers shown than any other as well. However, level 3 has the biggest range of values.

A diagram of a distribution of bmi

Description automatically generated with medium confidence

### Figure [3] - Education Level-Hemoglobin

This graph depicts distribution of waist circumference for specific education levels using PROC GLM.

Reminder:

1. Less than 9th Grade
2. 9-11th grade (includes 12th grade with no diploma)
3. Highschool Graduate/GED or equivalent
4. Some college of AA degree
5. College Graduate or above

This graph includes summary statistics for hemoglobin data, and a box and whisker plot including the variables as well. Overall, the graph isn’t particularly varied. Averages remain relatively constant for each level of education as well. The R squared values is also very low which supports the trends shown on the graph.

A screenshot of a computer screen

Description automatically generated

### Figure [4] - Education Level-Red Blood Cell Count

This graph depicts distribution of waist circumference for specific education levels using PROC GLM.

Reminder:

1. Less than 9th Grade
2. 9-11th grade (includes 12th grade with no diploma)
3. Highschool Graduate/GED or equivalent
4. Some college of AA degree
5. College Graduate or above

This graph includes summary statistics for red blood cell data (LBXRBCSI), and a box and whisker plot including the variables as well. Overall, the graph doesn’t show a huge variance between the education levels, similarly to hemoglobin counts. The R squared values are also very low which supports the trends shown on the data.

A screenshot of a computer

Description automatically generated

### Table [1] - Education Level-BMI Summary Statistics

Found using PROC GLM

### A screenshot of a computer screen Description automatically generated

### Table [2] Education Level-Waist Circumference

Found using PROC GLM

A screenshot of a computer

Description automatically generated

## Table [3] -Education Level-BMI Summary Statistics

Found using PROC GLM

A screenshot of a computer

Description automatically generated

### Table [4] – Education Level-Red Blood Cell Count

Found using PROC GLM

A screenshot of a data analysis

Description automatically generated

### Figure [5] – Education Level-Red Blood Cell Count \*Residual plots\*

### A screenshot of a screenshot of a computer Description automatically generated

Found using PROC REG

### Figure [6] – Education Level-Hemoglobin \*Residual plots\*

A screenshot of a graph

Description automatically generatedFound using PROC REG

### Figure [7] – Education Level-Body Waist Circumference \*Residual plots\*

A screenshot of a graph

Description automatically generated

Found using PROC REG

### Figure [8] – Education Level-White Blood Cell Count \*Residual and best fit plot\*

A screenshot of a graph

Description automatically generated

Found using PROC REG

## Table [5] - Education Level-Waist Circumference Summary Statistics

A screenshot of a computer

Description automatically generated

**Found using PROC REG**

## Table [6] - Education Level-White Blood Cell Count

A screenshot of a computer

Description automatically generated

**Found using PROC REG**

### Figure [9] - Education Level-BMI \*Residual Plots\*

A screenshot of a computer

Description automatically generated

**Found using PROC REG**

## Figure [9] – Average Weight, Height, Waist Circumference, and BMI by Education Level

A screenshot of a graph

Description automatically generated

A graph of a scatter plot

Description automatically generated

**Found using PROC REPORT and PROC SGPLOT**

## Figure [10] – Average Weight, Height, Waist Circumference, and BMI by Race/Ethnicity

A screenshot of a table

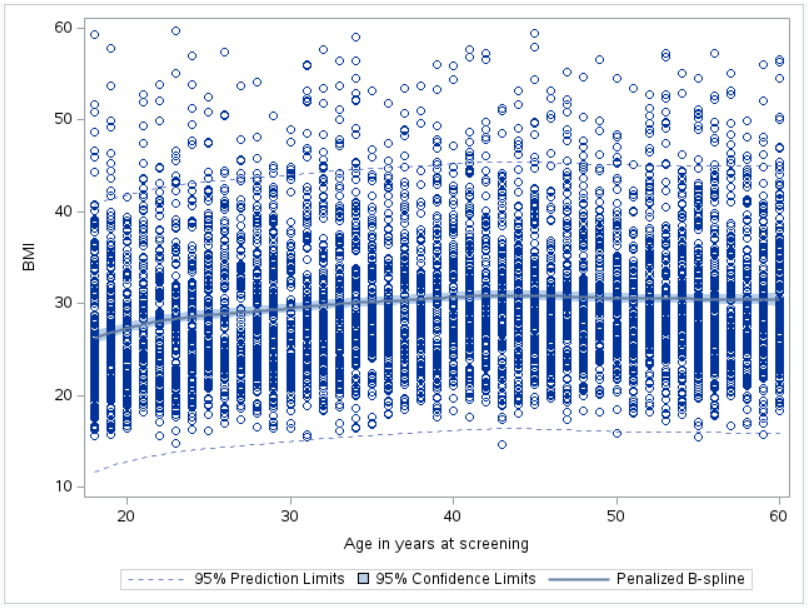
Description automatically generated

A diagram of a scatter plot

Description automatically generated

**Found using PROC REPORT and PROC SGPLOT**

### Figure [11] - Age in Years at screening – BMI with spline



**Found with PROC SGplot**

# Discussions

## Education Level and Overall Health

Our analysis is looking for a correlation between education level and overall health within adults. Since the variable of education (DMDEDUC2), only accounts for adults 20+ within the demographic data set. We used multiple methods of analyzing the data including box and whisker plots, residual plots, r-squared correlation, summary statistics, etc.

With our data we discovered there was no significant correlation between the variables of overall health. R-squared values produced with all variables ranged from .01 or lower which depicts a low connection, (Figure 1,3,4). In general, BMI and waist circumference had a higher correlation than blood related variables (red blood cell count and hemoglobin) but still not close to significant, (Table 1-4). While some of the graphs may show variance, it is expected that each group will have outliers and typical differences, but on average are related.

This shows that within the population selected by NHANES, there was no correlation between education level and overall health. (Figure 9) Meaning, an education level of 1, a less than high school educated individual was not less or more likely to have lower health related statistics. Before seeing our findings, we asserted that education level would impact health-related outcomes. Hypothesizing income from higher education levels would allow for more food, better shelter, and overall health treatment. However, as our data shows it did not influence overall outcomes.

# Concluding remarks

In our study investigating the relationship between education level and overall health among adults aged 20 and above, we conducted a comprehensive analysis using various statistical methods, including box and whisker plots, residual plots, R-Square correlation, and summary statistics. Despite employing these analytical approaches, our findings consistently indicated a lack of significant correlation between education level (represented by the variable DMDEDUC2) and overall health indicators. The R-Square values calculated for all variables were notably low, generally ranging from 0.01 or lower, suggesting a weak association between education level and health outcomes. Notably, while body mass index (BMI) and waist circumference showed relatively stronger correlations compared to blood-related variables such as red blood cell count and hemoglobin, these relationships were still not statistically significant.

Furthermore, our analysis within the sampled population from NHANES revealed that education level did not demonstrate a discernible correlation with overall health outcomes. Specifically, individuals with lower education levels (less than high school) were neither more nor less likely to exhibit poorer health-related statistics. This finding contradicts our initial hypothesis, which proposed that higher education levels might lead to improved health outcomes through increased income, enabling better access to food, shelter, and healthcare services. The complexity of the relationship between education and health within this demographic underscores the need for further investigation into additional factors that may influence health outcomes, beyond the scope of education level alone.

# References

*2017-March 2020 Pre-Pandemic Demographics Data - Continuous NHANES*. (2017). Cdc.gov. <https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&Cycle=2017-2020>

*2017-March 2020 Pre-Pandemic Examination Data - Continuous NHANES*. (2017). Cdc.gov. <https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination&Cycle=2017-2020>

Cody, R. (2018). Learning SAS by Example (2nd ed.). SAS Institute Inc.. <https://online.vitalsource.com/books/9781635266580>

SAS® OnDemand for Academics (2024). Statistical Analysis System [Computer software] [SASoda@sas.com](mailto:SASoda@sas.com)‌.

# Appendix

A computer code with text

Description automatically generatedA computer screen shot of a code

Description automatically generatedA white background with black text

Description automatically generatedA computer code with text

Description automatically generatedA computer code with text

Description automatically generatedA computer code with text

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated