# Lab 2 - A1C vs Waist-To-Height Ratio Descriptive Analysis

**Github Repository: https://github.com/mids-w203/lab2-Dominguez-Fardaev-Gorur-Habib**

Mukhammadali Fardaev, Umair Habib, Vishnu Gorur, Samuel Dominguez

April 17, 2025

## Introduction

Diabetes is a disease, approximately affecting 38.4 million Americans as of 2021[1], that occurs when the body cannot control blood glucose (sugar) levels due to a lack of Insulin damaging organs[1]. One of the important factors to look at in patients is their A1C level as it represents the patient's average blood glucose levels for the past 3 months[2]. While measurements related to mass and obesity are useful, as used by ADA guidelines[2], numerous existing studies have shown that waist to height ratio is a better metric when screening for diabetes and other diseases[3,4,5,6]. As this topic is at the forefront of public health, our team of data scientists represents a shared interest in better understanding the relationship between A1C levels and a variety of extrinsic factors as well as intrinsic health indicators.

The main focus of our study is to answer the following question: Is there a significant relationship between waist-to-height ratio and A1C levels? Our initial hypothesis is that waist-to-height ratio is positively correlated with A1C levels. To answer this question and test our hypothesis, we will explore the following related questions: (1) How descriptive is a simple linear regression model as compared to a multiple linear regression model that accounts for differences in human biological systems and external socio-economic variables? (2) Is there an effect of ethnicity in the relationship between waist-to-height ratio and A1C levels, particularly when comparing Non-Hispanic Whites to other ethnic groups? (3) Is there a significant relationship between Insulin and A1C levels? (4) Do we have an omitted variable bias confounding the interpretability of our models?

This descriptive study aims to provide insights for healthcare professionals to better identify diabetes risk factors.

## Description of the Data Source

This study utilizes data from the National Health and Nutrition Examination Survey (NHANES). NHANES combines interviews and physical examinations to collect data on a wide range of health indicators[7]. Our analysis specifically draws from the NHANES 2017-2018 dataset which includes a sample of 9,254 U.S. adults aged 12-80+, collected via trained interviewers using the Computer-Assisted Personal Interview system (Appendix A.3). This survey is cross-sectional as it contains self-reported measurements from the participants at one point in time during this 2017-2018 window.

However, there are limitations in the dataset as it was self-reported meaning that participants had the choice of not disclosing personal information as well as participants who did not know how to answer certain questions.

## Data Wrangling

We begin by merging 6 different NHANES datasets, namely diabetes questionnaires (DIQ_J), demographics (DEMO_J), glycemic markers (GHB_J), body measurements (BMX_J), triglycerides (TRIGLY_J), and insulin (INS_J). The joined dataset contains 9,254 observations with 133 variables. Then, to examine our research question, we transform the data to narrow down to 9 key variables: A1C (dependent variable), waist circumference, height, income-to-poverty ratio, age, pregnancy, triglycerides , insulin, gender, ethnicity, and family history of diabetes. The latter 3 variables, we consider them as covariates to investigate more in depth subgroup differences.

After cleaning the data to remove participants who are pregnant, under 18 years of age, have missing predictor variable values, or have missing A1C values, we rename variable names for clarity and recode the categorical variables into factor variables. Next, we split the cleaned dataset into 2 parts for development (exploratory_df) and validation (confirmatory_df) purposes using 3:7 ratio respectively.

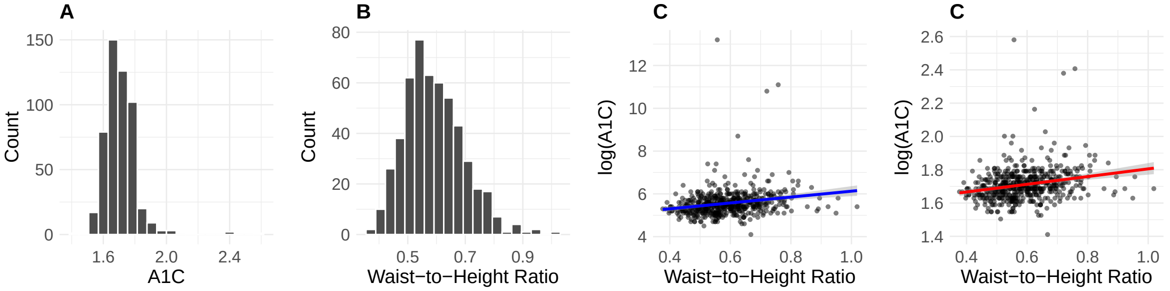
## Operationalization

Broadly, the research question examines the relationship between various predictors and A1C. We combined height and waist circumference, both measured in centimeters, into a single variable (Waist circumference / height ratio). Family history is the response to the question of whether the respondent has been told they have a health condition or family history that increases their risk of diabetes, either Yes or No (Appendix A.2). Triglyceride levels are measured as milligrams per deciliter (mg/dL) (Appendix A.6) and insulin levels are measured in microunits per milliliter ($\mu$U/mL) (Appendix A.7). Income-to-poverty is measured as a continuous ratio variable by "dividing family (or individual) income by the poverty guidelines specific to the survey year" (Appendix A.3). Poverty could also be operationalized by comparing family income against the national poverty line[8] for a binary impoverished or not impoverished value. However, the ratio provides a wider range of values that provide more granular information on a respondents income than a binary variable. Age is measured in years and gender is either Male or Female (Appendix A.3). Ethnicity is represented as numbers 1-7, with labels Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, Other Race - Including Multi-Racial, and Missing, respectively (Appendix A.3). Finally, according to the ADA, an A1C level below 5.7% is considered normal, 5.7-6.4% indicates pre-diabetic, and 6.5% or higher marks diabetes[2].

Starting with 9,254 observations, 47 observations from pregnant participants and 3,398 observations from participants under 18 years of age were removed to reduce confounding due to altered metabolic states during pregnancy and adolescent development. Next, 595 observations with missing A1C values were removed. Finally, 3,503 observations with a missing predictor variable were removed (this can introduce potential biases: selection bias and attrition bias). This left 1,711 total observations and 11 variables for our descriptive analysis. Model specifications were tested against the explanatory dataset, which contained 30% of the observations and then tested against the confirmatory dataset containing the other 70% of observations.

## Model Specification

In order to examine the relationship between A1C and waist-to-height ratio along with other predictor variables, we developed 2 linear regression models on our exploratory dataset (515 observations). After, we applied the final features to the confirmatory dataset (1196 observations) to guarantee consistent description patterns.



Before proceeding with model design, we examined the distribution of A1C (Figure A), and waist-to-height ratio (Figure B). Our simple linear regression model was first built off untransformed A1C vs Waist-to-Height ratio (Figure C). However, as we added more variables, our model improved as we transformed A1C to log(A1C). We applied the log transformation as A1C was highly right-skewed. The transformation allowed us to maintain a log-linear model instead of adding more complicated transformations. Moreover, after reviewing the positively skewed distributions of triglycerides and insulin (Appendix C), we applied logarithmic transformation to these variables as well. Finally, we began by fitting the simple model to examine the relationship between the waist-to-height ratio and log(A1C) only.

$$\log(\text{A1C}) = \beta_0 + \beta_1 \cdot \text{WaistHeightRatio} + \varepsilon$$

We then, expanded the simple model by including more predictor variables, such as age, gender, ethnicity, income-to-poverty ratio, family history, triglycerides, and insulin together with the waist-to-height ratio.

$$\begin{aligned}\log(\text{A1C}) = \beta_0 &+ \beta_1 \cdot \text{WaistHeightRatio} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Gender} \\ &+ \beta_4 \cdot \text{Ethnicity} + \beta_5 \cdot \text{IncomePovertyRatio} + \beta_6 \cdot \text{FamilyHistory} \\ &+ \beta_7 \cdot \log(\text{Triglycerides}) + \beta_8 \cdot \log(\text{Insulin}) + \varepsilon\end{aligned}$$

## Model Assumptions

Our model meets most of the key assumptions for our regression model. The residual versus fitted plot shows that the residuals are randomly distributed around 0 and supports our second assumption of zero expectation conditional mean, as well as assumption of linearity (Appendix C). The variance of residuals is constant based on the Breusch-Pagan test ($\text{BP} = 13.561, \quad \text{df} = 12, \quad \text{p-value} = 0.3296$), revealing the homoscedasticity. However, the Shapiro test rejects the null hypothesis of normality of residual distribution with p-value of $3.455 \times 10^{-25}$. For I.I.D, NHANES data are collected through sampling design that can cause geographic clustering and oversampling of subgroups[7], violating the assumption to some extent.

## Model Results and Interpretation

Table 1: Regression Model for A1C

| | *Dependent variable:* | |
|---|---|---|
| | log(A1C) | |
| | (1) | (2) |
| Waist-Height Ratio | $0.23^{***}$ (0.03) | 0.04 (0.03) |
| Age | | $0.002^{***}$ (0.0001) |
| Male | | 0.01 (0.005) |
| Mexican American | | $0.03^{***}$ (0.01) |
| Other Hispanic | | $0.02^{**}$ (0.01) |
| Non-Hispanic Black | | $0.04^{***}$ (0.01) |
| Non-Hispanic Asian | | $0.03^{***}$ (0.01) |
| Other/Multi-Racial | | $0.03^{**}$ (0.01) |
| Income-to-Poverty Ratio | | −0.001 (0.002) |
| Family History: Yes | | $0.03^{***}$ (0.01) |
| Triglycerides (log) | | $0.01^{***}$ (0.005) |
| Insulin (log) | | $0.03^{***}$ (0.004) |
| Observations | 1,196 | 1,196 |
| $R^2$ | 0.06 | 0.24 |
| Adjusted $R^2$ | 0.06 | 0.23 |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Firstly, we conducted an Anova test and $R^2$ analysis to see if adding more predictors created statistically significant differences between our single and multiple linear regression models.

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$$
$$H_a : \text{At least one } \beta_i \neq 0 \text{ for } i \in \{2, 3, \dots, 8\}$$

The Analysis of Variance table (anova) indicates a p-value of $6.91 \times 10^{-48}$ which is less than our $\alpha = 0.05$, making it statistically significant and that the null hypothesis should be rejected. In broader terms, this anova test confirms that the additional predictors added improve the model fit and is able to explain more variation in A1C than waist-to-height ratio alone. As the highest value in the computed VIF values for our model (1.73) is below 4, there is no multicollinearity in our model. Additionally, the $R^2$ for our model is 24% meaning that 24% of the variation in the outcome variable can be explained by the variation in our predictor variables. Lastly, we applied Holm's sequential method to avoid an issue with Family Wise Error Rates (FWER) and to ensure that the probability of making a Type I error is no greater than (0.05) (Appendix E).

In the stargazer table above, the waist-to-height ratio coefficient in the simple regression model decreases from 0.23 to 0.04 in the multiple regression model. Additionally the adjusted p-value (see Appendix E) of p = 0.53>0.05 threshold indicates that in the extended model, the waist-to-height ratio is not a statistically significant predictor variable. This is a notable result as it would imply the presence of confounder variables in the simple linear model inflating the significance of the coefficient. This is an example of omitted variable bias where 1 (The coefficient of waist-to-height ratio) should move closer to 0.

The multiple linear regression when clustering for different ethnicities (in particular Black and Asian) shows a statistically significant difference in predicting A1C levels when compared to the base group - Non-Hispanic Whites. A 1 unit increase in the waist-to-height ratio in Non-Hispanic Blacks and Asians are linked to 4.08% and 3.05% increase respectively in A1C levels relative to the reference group while keeping other variables fixed. This is also an interesting result because it aligns with studies indicating that Asian countries tend to have the highest number of diabetic patients[9].

One of the most important variables that is tied to diabetes and A1C levels is insulin. Our adjusted p-value $= 2 \times 10^{-8}$ indicates that the log(insulin) is a highly significant variable in our analysis. Given a proportional change in log(insulin) concentration, the log(insulin) coefficient (0.03) describes the percent change in A1C.

A limitation of our study stems from the presence of NA values in our surveyed dataset. Filtering out NA values puts us at risk of a selection bias where our sample values used in our model are no longer representative of the larger sampled population. Going forward, one possible solution to avoid this is to use predictor variables without missing values. Another limitation of this study is the violation of the normality assumption which brings into question our calculated p-values and its respective significance. Going forward, further work can be done to use different predictor variables or transformations to improve the model. Interaction between predictor variables can be added to see if there are multiplicative effects of the predictor variables on the outcome variable.

In conclusion, our study shows that there is not a significant relationship between waist-to-height ratio and A1C levels. There is a weak association such that an increase in waist-to-height ratio is associated with an increase in A1C. Interestingly, there are a multitude of biological and socio-economic factors that are potentially more influential than waist-to-height ratio based on statistical significance such as log(insulin), ethnicity etc. With this study, clinicians and researchers have another statistical analysis looking at the relationship between A1C and health indicators that will help improve identifying patients at higher risk for diabetes.

## References

1. Centers for Disease Control and Prevention. (2023). National Diabetes Statistics Report: Estimates of diabetes and its burden in the United States. U.S. Department of Health and Human Services. https://www.cdc.gov/diabetes/php/data-research/index.html

2. American Diabetes Association; 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2020. Diabetes Care 1 January 2020; 43 (Supplement_1): S14–S31.

3. Kuerban, A. (2020). Beyond Asian-specific cutoffs: Gender effects on the predictability of body mass index, waist circumference, and waist circumference-to-height ratio on hemoglobin A1c. PLOS ONE, 15(8), e0237095.

4. Ashwell M, Gunn P, Gibson S. Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: systematic review and meta-analysis. Obes Rev. 2012;13(3):275–86.

5. Browning LM, Hsieh SD, Ashwell M. A systematic review of waist-to-height ratio as a screening tool for the prediction of cardiovascular disease and diabetes: $0 \cdot 5$ could be a suitable global boundary value. Nutr Res Rev. 2010;23(2):247–69.

6. Yang H, Xin Z, Feng J-P, Yang J-K. Waist-to-height ratio is better than body mass index and waist circumference as a screening criterion for metabolic syndrome in Han Chinese adults. Medicine (Baltimore). 2017;96(39).

7. Centers for Disease Control and Prevention. (2023). NHANES analytic guidelines: Sample design. National Center for Health Statistics.

8. U.S Department of Health and Human Services. (n.d.). Poverty guidelines. Office of the Assistant Secretary for Planning and Evaluation. https://aspe.hhs.gov/topics/poverty-economic-mobility/poverty-guidelines

9. International Diabetes Federation. IDF diabetes atlas. International Diabetes Federation; 2017

## Appendix

### Appendix A. Dataset Links

1. NHANES datasets: Link
2. DIQ_J - Diabetes questionnaire: (Family-History): Link
3. DEMO_J - Demographics (Age, Gender, Ethnicity, Income-to-Poverty Ratio): Link
4. GHB_J - Glycemic indicators (A1C): Link
5. BMX_J - Body measurements (Waist and Height Measurements): Link
6. TRIGLY_J - Triglyceride (Triglycerides): Link
7. INS_J- Insulin (Insulin): Link

### Appendix B. Model Specifications

1. $\text{A1C} = \beta_0 + \beta_1 \cdot \text{WaistHeightRatio} + \varepsilon$

   Base simple linear regression model.

2. $\text{A1C} = \beta_0 + \beta_1 \cdot \text{WaistHeightRatio} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Gender} + \beta_4 \cdot \text{Ethnicity} + \beta_5 \cdot \text{IncomePovertyRatio} + \beta_6 \cdot \text{FamilyHistory} + \beta_7 \cdot \text{Triglycerides} + \beta_8 \cdot \text{Insulin} + \varepsilon$

   Base multiple linear regression model.

3. $\log(\text{A1C}) = \beta_0 + \beta_1 \cdot \text{WaistHeightRatio} + \varepsilon$

   Log transformation on dependent variable (A1C) improved model 1.

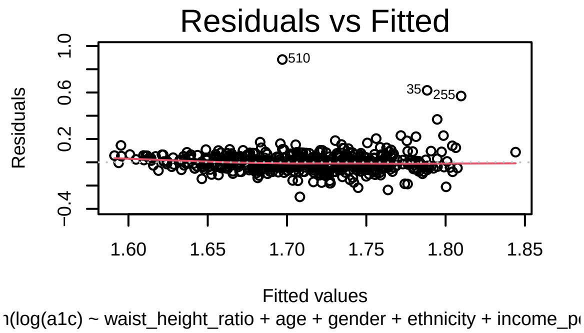
4. $\log(\text{A1C}) = \beta_0 + \beta_1 \cdot \text{WaistHeightRatio} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Gender} + \beta_4 \cdot \text{Ethnicity} + \beta_5 \cdot \text{IncomePovertyRatio} + \beta_6 \cdot \text{FamilyHistory} + \beta_7 \cdot \text{Triglycerides} + \beta_8 \cdot \text{Insulin} + \varepsilon$

   Log transformation on dependent variable (A1C) improved model 2.

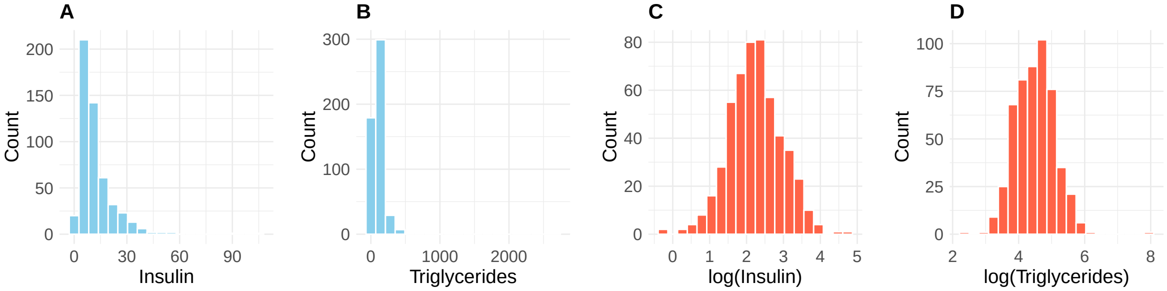
5. $\log(\text{A1C}) = \beta_0 + \beta_1 \cdot \text{WaistHeightRatio} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Gender} + \beta_4 \cdot \text{Ethnicity} + \beta_5 \cdot \text{IncomePovertyRatio} + \beta_6 \cdot \text{FamilyHistory} + \beta_7 \cdot \log(\text{Triglycerides}) + \beta_8 \cdot \log(\text{Insulin}) + \varepsilon$

   Log transformation on triglycerides and insulin variables as they show a linear relationship with A1C in residuals vs component plots.

### Appendix C. Residuals vs Fitted Plot



## Appendix D. Distribution of Insulin and Triglycerides



## Appendix E. Holm-Adjusted P-Values

Table 2: Regression Coefficients with Holm-Adjusted P-Values

| | Raw.P.Value | Adjusted.P |
|---|---|---|
| (Intercept) | 0.00e+00 | 0.00e+00 |
| waist_height_ratio | 2.09e-01 | 5.35e-01 |
| age | 1.55e-35 | 1.86e-34 |
| genderMale | 1.78e-01 | 5.35e-01 |
| ethnicityMexican American | 6.36e-04 | 4.45e-03 |
| ethnicityOther Hispanic | 1.76e-02 | 8.81e-02 |
| ethnicityNon-Hispanic Black | 1.36e-10 | 1.50e-09 |
| ethnicityNon-Hispanic Asian | 5.98e-06 | 5.38e-05 |
| ethnicityOther/Multi-Racial | 1.85e-02 | 8.81e-02 |
| income_poverty_ratio | 4.95e-01 | 5.35e-01 |
| family_historyYes | 1.03e-04 | 8.25e-04 |
| log(triglycerides) | 8.85e-03 | 5.31e-02 |
| log(insulin) | 2.03e-09 | 2.03e-08 |