**I. Introduction**

Investigators are interested in examining Type II Diabetes in 403 black participants residing in rural counties of Virginia. The data was combined from two published studies characterizing cardiovascular risk factors in Virginia for African Americans. The first study was a community-based study measuring the prevalence of coronary heart disease among Black population in Virginia. The second study measured the effectiveness of church-based smoking cessation interventions for rural Black populations. A smoking cessation program was implemented by church groups to target smoking behaviors. This program included one-on-one counseling with two counselors present, self-help materials, and community activities. Investigators followed-up with participants after 18 months to measure the effectiveness of the intervention. We now use this data to conduct studies to observe the relationship between the onset of Type II Diabetes and factors associated with the disease. Type II diabetes is known to be strongly associated with obesity and hypertension. The waist-to-hip ratio may also be a predictor of both diabetes and heart disease. Glycosylated hemoglobin (HbA1C) > 7.0 is usually indicative of diabetes. The goal of this paper is to describe the relationship between BMI and hemoglobin A1C and determine other factors that may impact this relationship. Researchers postulate that the higher the BMI the more likely an individual will have been diagnosed with diabetes.

**II. Methods**

We conducted statistical analyses of 403 Black participants in two cohorts using the SAS statistical package with a significance level of 0.05. The two cohorts were interviewed to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans. Participant characteristics of interest were collected such as height, weight, age, gender, blood glucose levels, total cholesterol, systolic and diastolic blood pressure, waist and hip circumferences. We created several variables using patient characteristics to carry out analyses such as BMI, BMI status (normal, overweight, obese), natural logarithmic hemoglobin A1C, waist-to-hip circumference ratio, and mean systolic blood pressure.

**III. Results**

1. **Descriptive Statistics (Question 3)**

Descriptive statistics were produced for a sample size of 386 participants due to missing data (17 participants excluded from analysis). We used a PROC FREQ function to obtain frequencies and percentages for categorical variables, displayed below in Table 1. The sample was observed to be predominantly female with 226 participants representing 58.66 % of the sample. The most represented BMI category in the sample was those identified to be obese (39.38%) followed by those identified to be overweight (31.97%) and, finally, participants identified to be normal as the least represented in the sample (28.76%).

Table 1. Descriptive Statistics for Participant Characteristics

|  |  |
| --- | --- |
| Variable | N = 386 |
| Male:  Female  Male | 226 (58.66%)  160 (41.45%) |
| BMI Category:  Normal  Overweight  Obese | 111 (28.76%)  123 (31.87%)  152 (39.38%) |

A PROC MEANS function was utilized to obtain descriptive statistics for continuous variables including the sample size, mean, median, standard deviation, and minimum and maximum values (see Table 2). Various characteristics in relation to cardiovascular health were measured and varied greatly amongst participants. The mean age of the sample was observed to be 46.71 years with participant age ranging between 19-92 years. Mean BMI was observed to be 29.13 kg/m2 with participant BMI ranging between 19.1 – 55.8 kg/m2. 137 observations for mean systolic blood pressure were used as 249 participants were missing 2nd measurements for diastolic and systolic blood pressure. Mean systolic blood pressure was observed to be 230.66 mmHgThe outcomes of this study are hemoglobin A1C and natural logarithmic hemoglobin A1C. Mean hemoglobin A1C was observed to be 5.60% and mean natural logarithmic hemoglobin A1C was observed to be 1.67 units.

Table.2 Descriptive Statistics for Participant Characteristics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | N | Mean | Median | Std.Dev | Minimum | Maximum |
| Age (yrs) | 386 | 46.71 | 44 | 16.43 | 19 | 92 |
| Height (in) | 386 | 65.96 | 66 | 3.92 | 52 | 76 |
| Weight (lbs) | 386 | 179.57 | 174 | 39.43 | 102 | 325 |
| BMI (kg/m2) | 386 | 29.13 | 29.98 | 6.42 | 19.1 | 55.8 |
| Total Cholesterol (mg/dL) | 386 | 208.25 | 204 | 43.95 | 118 | 443 |
| Blood Glucose Level (mg/dL) | 386 | 107.16 | 90 | 53.37 | 48 | 385 |
| Systolic Blood Pressure, 1st measure (mmHg) | 382 | 136.92 | 136 | 22.95 | 90 | 250 |
| Systolic Blood Pressure, 2nd measure (mmHg) | 137 | 152.12 | 149 | 21.87 | 110 | 238 |
| Diastolic Blood Pressure, 1st measure (mmHg) | 382 | 83.33 | 82 | 13.33 | 48 | 124 |
| Diastolic Blood Pressure, 2nd measure (mmHg) | 137 | 92.32 | 92 | 11.61 | 60 | 124 |
| Waist Circumference (in) | 384 | 38.15 | 38 | 5.66 | 26 | 56 |
| Hip Circumference (in) | 384 | 43.28 | 42 | 5.55 | 30 | 64 |
| Waist-to-Hip ratio | 384 | 0.88 | 0.88 | 0.07 | 0.68 | 1.14 |
| Hemoglobin A1C (%) | 373 | 5.60 | 4.86 | 2.23 | 2.68 | 16.1 |
| Natural Log of Hemoglobin A1C | 373 | 1.67 | 1.58 | 0.31 | 0.99 | 2.78 |
| Mean Systolic Blood Pressure (mmHg) | 137 | 230.66 | 223 | 31.02 | 155 | 347 |

1. **Association between BMI categories and Hemoglobin A1C (Question 4)**

We observed the association between BMI categories (normal, overweight, and obese) and hemoglobin A1C and the natural log of hemoglobin A1C in two separate one-way ANOVA models utilizing a PROC GLM statement. The normal category was used as the reference group.

**One-Way ANOVA between BMI study categories and non-logarithmic Hemoglobin A1C**

In order to test common means of hemoglobin A1C in the study sample of 386 participants versus an alternative global hypothesis of unequal means by BMI category, we performed a one-way ANOVA with post-hoc testing using Tukey’s procedure applying a two-sided alpha of 0.05. We did not observe a statistically significant difference in means in mean hemoglobin A1C levels across the three different BMI categories (F-statistic = 2.14, d.f. = (2, 370), R2 = 0.0114, p-value = 0.1191). The model explains 1.14% of variance between BMI categories and hemoglobin A1C. Due to the difference in means of hemoglobin A1C not being statistically significant the linear associations of the main effects with hemoglobin A1C are unreliable. We observed that, on average, hemoglobin A1C increased 0.58329 units as obesity status increased. We also observed that, on average, hemoglobin A1C increased 0.32494 units (summarized in table 4A).

Table 4A. Slope Estimates between BMI categories and Hemoglobin A1C

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Slope (Beta Estimate) | Standard Error | t-value | p-value |
| Intercept | 5.26439 | 0.21454 | 24.54 | <0.0001 |
| obese | 0.58329 | 0.28201 | 2.07 | 0.0393 |
| overweight | 0.32494 | 0.29566 | 1.10 | 0.2725 |
| normal | reference |  |  |  |

**One-Way ANOVA between BMI categories and natural logarithmic Hemoglobin A1C**

In order to test for differences in mean natural logarithmic hemoglobin A1C by BMI category, we performed a one-way ANOVA with post hoc testing using Tukey’s procedure applying a two-sided alpha of 0.05. According to Table 4B, we observed a statistically significant difference in mean natural logarithmic hemoglobin A1C levels across the three different BMI categories (F-statistic = 3.36, d.f. = (2, 370), R2 = 0.0179, p-value = 0.0357). The model explains 1.79% of variance between mean natural logarithmic hemoglobin A1C and BMI categories. Due to the difference in mean natural logarithmic hemoglobin A1C by BMI category being statistically significant, we observed the association between the main effects and mean natural logarithmic hemoglobin A1C. The normal BMI category was determined to be the reference group for this analysis. We observed a statistically significant association between the obese BMI category and mean logarithmic hemoglobin A1C (t-statistic = 2.59, d.f. = 1, p-value = 0.0099). On average, mean logarithmic hemoglobin A1C increased 0.10273 units as the obese category increased. In contrast, we did not observe a statistically significant association between participants identified as overweight and mean logarithmic hemoglobin A1C (t-statistic = 1.33, d.f = 1, p-value = 0.1829). According to Table 4C, we observed that on average, mean logarithmic hemoglobin A1C increased 0.0555 units as the overweight BMI category increased. In post hoc testing using Tukey’s procedure to compare pairs of group means, we found statistically significant differences only between normal and obese study groups (p-value = 0.0268). Mean logarithmic hemoglobin A1C was observed to be higher in the obese study group in comparison to the normal study group (see Appendix A). We determined that that the rest of the statistical analyses in this paper was more appropriately conducted with logarithmic hemoglobin A1C based on our conclusions in part B.

Table 4B. Slope Estimates between BMI categories and Natural Logarithmic Hemoglobin A1C

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Slope (Beta Estimates) | Standard Error | t-value | p-value |
| Intercept | 1.60738 | 0.03016 | 53.29 | <.00010 |
| Obese | 0.10273 | 0.03965 | 2.59 | 0.0099 |
| overweight | 0.05546 | 0.04156 | 1.33 | 0.1829 |
| normal | reference |  |  |  |

Table 4C. Descriptive Statistics of Natural Logarithmic Hemoglobin A1C by BMI category

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Natural Logarithmic Hemoglobin A1C | |
| Variable | N | Mean | Std.dev |
| Obese | 147 | 1.71 | 0.31 |
| overweight | 119 | 1.66 | 0.33 |
| normal | 107 | 1.61 | 0.29 |

1. **Linear Regression Models Describing the Linear Association between BMI and Logarithmic Hemoglobin A1C (Question 5)**

We used a PROC REG function producing a simple linear regression model to observe the linear association between various created variables of BMI and logarithmic hemoglobin A1C.

**Linear Regression Model Between Logarithmic Hemoglobin A1C and BMI Dummy Variables: Overweight and Obese (Question 5A)**

In order to observe the linear association between logarithmic hemoglobin A1C and BMI dummy variables (obese and overweight) we used a simple linear regression model. We discerned that there is a statistically significant linear association between logarithmic hemoglobin A1C and obese and overweight (F-statistic = 0.0357, d.f. = (2, 370), adj-R2 = 0.0125, p-value = 0.0357). The linear regression model explains 1.25% of variance between logarithmic hemoglobin A1C and obese and overweight. Due to the linear association between logarithmic hemoglobin A1C and obese and overweight dummy variables being statistically significant, we observed the linear association between the logarithmic hemoglobin A1C and the main effects. We first examined the linear association between logarithmic hemoglobin A1C and the dummy variable overweight and observed no statistically significant linear association between logarithmic hemoglobin A1C and overweight (t-statistic = 1.33, d.f. = 1, p-value = 0.1829). We then examined the linear association between logarithmic hemoglobin A1C and the dummy variable obese and observed a statistically significant linear association between logarithmic hemoglobin A1C and obese (t-statistic =2.59, d.f. = 1, p-value = 0.0099). The linear associations between natural logarithmic hemoglobin A1C and BMI dummy variables overweight and obese are summarized in table 5A.

Table 5A. Slope Estimates between BMI Dummy Variables: Overweight and Obese

and Natural Logarithmic Hemoglobin A1C

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slopes (Beta Estimates) | Standard Error | t-statistic | P-value |
| Intercept | 1 | 1.60738 | 0.03016 | 53.29 | <.0001 |
| Overweight | 1 | 0.05546 | 0.04156 | 1.33 | 0.1829 |
| Obese | 1 | 0.10273 | 0.03965 | 2.59 | 0.0099 |

Figure 5.1. Fit Diagnostics between Natural Logarithmic Hemoglobin A1C

and Dummy Variables: Obese and Overweight

**A screenshot of a computer screen

Description automatically generated**

**Linear Regression Model Between Natural Logarithmic Hemoglobin A1C and Ordinal BMI (Question 5B)**

In order to observe the linear association between natural logarithmic hemoglobin A1C and ordinal BMI we utilized a simple linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and ordinal BMI (F-statistic = 6.73, d.f. = (1, 371), R2 = 0.0178, p-value = 0.0098). The model explains 1.78% of variance between natural logarithmic hemoglobin A1C and ordinal BMI. Due to there being a statistically significant linear association, we examined the association of the main effects with natural logarithmic hemoglobin A1C. We observed the linear association between ordinal BMI and natural logarithmic hemoglobin A1C and discerned a statistically significant linear association (t-statistic = 2.59, d.f. = 1, p-value = 0.0098). On average, natural logarithmic hemoglobin A1C increased 0.05116 units as ordinal BMI increased. The linear association between ordinal BMI and natural logarithmic hemoglobin A1C is summarized in table 5B.

Table 5B. Slope Estimates between Ordinal BMI and Natural Logarithmic

Hemoglobin A1C

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slopes (Beta Estimates) | Standard Error | t-statistic | P-value |
| Intercept | 1 | 1.55776 | 0.04457 | 34.95 | <.0001 |
| BMI Categories | 1 | 0.05116 | 0.01972 | 2.59 | 0.0098 |

Figure 5.2. Fit Diagnostics between Natural Logarithmic Hemoglobin A1C

and Ordinal BMI

A screenshot of a computer screen

Description automatically generated

**Linear Regression Model Between Natural Logarithmic Hemoglobin A1C and Continuous BMI (Question 5C)**

We observed the linear association between natural logarithmic hemoglobin A1C and continuous BMI utilizing a simple linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and continuous BMI (F-statistic = 7.55, d.f. = (1, 371), R2 = 0.0199, p-value = 0.0063). The model explains 1.99% of variance between natural logarithmic hemoglobin A1C and continuous BMI. Due to an observed statistically significant linear association, the association of the main effects with natural logarithmic hemoglobin A1C were examined. We observed the linear association between natural logarithmic hemoglobin A1C and continuous BMI and discerned a statistically significant linear association between the two variables (t-statistic = 2.75, d.f. = 1, p-value = 0.0063). On average, natural logarithmic hemoglobin A1C increased 0.00689 units as continuous BMI increased. The linear association between continuous BMI and natural logarithmic hemoglobin A1C is summarized in table 5C.

Table 5C. Slope Estimates between Continuous BMI and Natural Logarithmic

Hemoglobin A1C

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slopes (Beta Estimates) | Standard Error | t-statistic | P-value |
| Intercept | 1 | 1.46487 | 0.07479 | 19.59 | <.0001 |
| BMI | 1 | 0.00689 | 0.00251 | 2.75 | 0.0063 |

Figure 5.3. Fit Diagnostics between Natural Logarithmic Hemoglobin A1C

and Continuous BMI

A screenshot of a computer screen

Description automatically generated

**Piecewise Linear Regression Model Between Natural Logarithmic Hemoglobin A1C and Piecewise BMI variables: BMI1 BMI2 BMI3 (Question 5D)**

In order to test the linear association between natural logarithmic hemoglobin A1C and piecewise BMI variables: BMI1 BMI2 and BMI3 we utilized a piecewise linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and the piecewise BMI variables (F-statistic = 3.66, d.f. = (3, 369), adj-R2 = 0.0210, p-value = 0.0127). The model explains 2.10% of the variance between natural logarithmic hemoglobin A1C and the piecewise BMI variables: BMI1, BMI2, and BMI3. Due to the observation of a statistically significant linear association between natural logarithmic hemoglobin A1C and piecewise BMI we examined the association between the main effects and natural logarithmic hemoglobin A1C. We first examined the association between natural logarithmic hemoglobin A1C and BMI1 and discerned that there was no statistically significant linear association (t-statistic = 1.93, d.f. = 1, p-value = 0.0542). We discerned that, on average, natural logarithmic hemoglobin A1C increased 0.12641 units as BMI1 increased. Next, we examined the linear association between natural logarithmic hemoglobin A1C and BMI2 and observed no statistically significant linear association (t-statistic = 0.55, d.f. = 1, p-value = 0.5570). We discerned that, on average, natural logarithmic hemoglobin A1C increased 0.04101 units as BMI2 increased. Finally, we examined the linear association between natural logarithmic hemoglobin A1C and BMI3 and observed no statistically significant linear association (t-statistic =0.56, d.f. = 1, p-value = 0.5770). We observed that, on average, natural logarithmic hemoglobin A1C increased 0.03401 units as BMI3 increased. The standardized slope estimates describing the linear association between natural logarithmic hemoglobin A1C and BMI piecewise variables: BMI1, BMI2, BMI3 are summarized in table 5D.

Table 5D. Slope Estimates between Natural Logarithmic Hemoglobin A1C and

Piecewise BMI Variables: BMI1, BMI2, BMI3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slopes (Beta Estimate) | Standard Error | t-value | p-value | Standardized Estimate |
| Intercept | 1 | 0.77981 | 0.26898 | 2.90 | 0.0040 | 0 |
| BMI1 | 1 | 0.02643 | 0.01369 | 1.93 | 0.0542 | 0.12641 |
| BMI2 | 1 | 0.00584 | 0.01064 | 0.55 | 0.5834 | 0.04101 |
| BMI3 | 1 | 0.00257 | 0.00460 | 0.56 | 0.5770 | 0.03401 |

Figure 5.4. Fit Diagnostics between Natural Logarithmic Hemoglobin A1C

and Piecewise BMI Variables: BMI1, BMI2, and BMI3

A screenshot of a graph

Description automatically generated

The piecewise model provides the best fit to data for modeling the linear association between natural logarithmic hemoglobin A1C and BMI. Of the four models, the piecewise linear regression model explains the largest amount of variance between BMI and natural logarithmic hemoglobin A1C (adj-R2 = 0.0210). Furthermore, the piecewise linear model is relatively normally distributed as the distribution of the data is shaped as a bell curve and is distributed around the line (see in Figure 5.4 row 3 plot 1 and row 2 plot 2).

1. **ANCOVA Models Between Natural Logarithmic Hemoglobin A1C and Age and Sex with Interaction (Question 6)**

We used a PROC GLM function to conduct two ANCOVA models observing the linear association between logarithmic hemoglobin A1C, BMI categories, and age and sex with interaction, separately.

**ANCOVA Model Between Natural Logarithmic Hemoglobin A1C, BMI Categories, and Age with Interaction**

In order to test the linear association between natural logarithmic hemoglobin A1C, BMI categories, and age with interaction, we conducted an ANCOVA test. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C, BMI categories, and age with interaction (F-statistic = 15.68, d.f. = (5, 367), R2 = 0.1760, p-value < 0.0001). The ANCOVA model explains 17.60% of the variance between natural logarithmic hemoglobin A1C, BMI categories, and age with interaction. We then determined to observe the association between the main effects and natural logarithmic hemoglobin A1C. We observed the linear association between natural logarithmic hemoglobin A1C and BMI categories, and age with interaction and discerned that there was no statistically significant linear association between natural logarithmic hemoglobin A1C and the interaction term (F-statistic = 2.04, d.f. = 2, p-value = 0.1309). Due to the linear association between natural logarithmic hemoglobin A1C and the interaction between BMI categories and age being statistically not significant, we recommend conducting the analysis removing the interaction term to conduct further analysis of the main effects (summarized in table 6A). Furthermore, due to the interaction between BMI categories and age not being statistically significant we can safely conclude that age is not an effect measure modifier.

Table 6A. Measuring Effect Measure Modification Between Natural Logarithmic Hemoglobin A1C and BMI Categories and Age with Interaction

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Degrees of Freedom | F-value | p-value |
| BMI categories | 2 | -0.21 | 0.8326 |
| age | 1 | 66.37 | <.0001 |
| Age\*BMI categories | 2 | 2.04 | 0.1309 |

**ANCOVA Model Between Natural Logarithmic Hemoglobin A1C, BMI Categories, and Male variable with Interaction**

In order to test the linear association between natural logarithmic hemoglobin A1C, BMI categories, and male with interaction, we conducted an ANCOVA test. We observed no statistically significant linear association between natural logarithmic hemoglobin A1C, BMI categories, and male with interaction (F-statistic = 1.96, d.f. = (5, 367), R2 = 0.0260, p-value = 0.0844). The model explains 2.60% of variance between natural logarithmic hemoglobin A1C, BMI categories, and male with interaction. Since we discerned that there was no statistically significant linear association between natural logarithmic hemoglobin A1C, BMI categories, and male with interaction, we recommend removing the interaction term from the model to conduct further analysis of the main effects. Furthermore, we can safely assume that male is not an effect measure modifier based on these facts.

1. **Linear Regression Models Between Natural Logarithmic Hemoglobin A1C and Waist Circumference, Hip Circumference, and Waist to Hip Ratio (Question 7)**

We used a PROC REG function to conduct multiple simple linear regression models to observe the linear association between natural logarithmic hemoglobin A1C and Waist, Hip, and Waist to Hip Ratio, separately. Then, we conducted a multivariable linear regression model, using PROC REG again, to observe collinearity between natural logarithmic H1AC, Waist, Hip, and Waist to Hip Ratio. In addition, we also applied residuals test to observe any outliers and influence points in the multivariable model.

**Simple Linear Regression Models Natural Logarithmic Hemoglobin A1C and Waist, Hip, and Waist to Hip Ratio (Question 7A)**

**Simple Linear Regression Model between Natural Logarithmic Hemoglobin 1AC and Waist Circumference**

In order to test the linear association between natural logarithmic hemoglobin A1C and waist, we conducted a simple linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and waist (F-statistic = 28.50, d.f. = (1, 369), R2 = 0.0717, p-value < 0.0001). The model explains 7.17% of variance between natural logarithmic hemoglobin A1C and waist. Due to an observed statistically significant linear association in the model, we observed the linear association between the main effects and natural logarithmic hemoglobin A1C. We observed the linear association between natural logarithmic hemoglobin A1C and the main effect (waist) and discerned a statistically significant linear association (t-statistic = 5.34, d.f. = 1, p-value < 0.0001). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.01471 units as waist circumference increased (summarized in table 7A).

Table 7A. Slope Estimates between Natural Logarithmic Hemoglobin A1C and Waist Circumference

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slopes (Beta Estimates) | Standard Error | t-value | P-value |
| Intercept | 1 | 1.10646 | 0.10633 | 10.41 | <.0001 |
| Waist Circumference | 1 | 0.01471 | 0.00276 | 5.34 | <.0001 |

In order to test the linear association between natural logarithmic hemoglobin A1C and hip, we conducted a simple linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and hip (F-statistic = 9.42, d.f. = (1, 369), R2 = 0.0249, p-value = 0.0023). Due to an observed statistically significant linear association in the model, we examined the linear association between natural logarithmic hemoglobin A1C and the main effect (hip). We discerned a statistically significant linear association between natural logarithmic hemoglobin A1C and hip (t-statistic = 3.07, d.f. = 1, p-value = 0.0023). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.00892 units as hip circumference increased (summarized in Table 7B).

Table 7B. Slope Estimates between Natural Logarithmic Hemoglobin A1C and Hip Circumference

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slopes (Beta Estimates) | Standard Error | t-value | P-value |
| Intercept | 1 | 1.28195 | 0.12679 | 10.11 | <.0001 |
| Hip Circumference | 1 | 0.00892 | 0.00291 | 3.07 | 0.0023 |

**Simple Linear Regression Models Natural Logarithmic Hemoglobin A1C and Waist-to-Hip Circumference Ratio**

In order to test the linear association between natural logarithmic hemoglobin A1C and waist-to-hip ratio, we conducted a simple linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and waist-to-hip circumference ratio (F-statistic = 22.63, d.f. = (1, 369), R2 = 0.0578, p-value < 0.0001). The model explains 5.78% of the variance between natural logarithmic hemoglobin A1C and waist-to-hip circumference ratio. Since we observed a statistically significant linear association between in the model, we examined the linear association between natural logarithmic hemoglobin A1C and the main effect (waist-to-hip circumference ratio). We discerned a statistically significant linear association between natural logarithmic A1C and the main effect (t-statistic = 4.76, d.f. = 1, p-value <.0001). We also observed that, on average, natural logarithmic hemoglobin A1C increased 1.01961 units as waist-to-hip circumference ratio increased (summarized in Table 7C).

Table 7C. Slope Estimates between Natural Logarithmic Hemoglobin A1C and Hip Circumference

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slopes (Beta Estimates) | Standard Error | t-value | P-value |
| Intercept | 1 | 0.76882 | 0.18963 | 4.05 | <.0001 |
| Waist-to-Hip Circumference | 1 | 1.01961 | 0.21431 | 4.76 | 0.0023 |

**Observing Collinearity, Outliers, and Influence Points in the Predictors of Natural Logarithmic Hemoglobin A1C**

We used a PROC REG function to perform a multivariable linear regression model to observe the linear association between natural logarithmic hemoglobin A1C, hip circumference, waist circumference, and waist-to-hip circumference ratio. We applied a collinearity test to observe the validity of the model. We also applied residuals test to observe any outliers and influence points in the model.

In order to observe the linear association between natural logarithmic hemoglobin A1C, hip circumference, waist circumference, and waist-to-hip circumference ratio, we performed a multivariable linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C, hip circumference, waist circumference, and waist-to-hip circumference ratio (F-statistic = 11.40, d.f. = (3, 367), adj-R2 = 0.0777, p-value < .0001). The model explains 7.77% of variance between natural logarithmic hemoglobin A1C and hip circumference, waist circumference, and waist-to-hip circumference ratio. Due to a statistically significant linear association observed in the model, we examined the linear association of natural logarithmic hemoglobin A1C with the main effects. We observed the linear association between natural logarithmic hemoglobin A1C and waist circumference and discerned there was no statistically significant linear association (t-statistic = -0.04, d.f. = 1, p-value = 0.9656). Due to the there being no statistically significant linear association the slope describing the linear association between natural logarithmic hemoglobin A1C and waist circumference is deemed unreliable (ß = -0.00193) (see Table 7D). Next, we examined the linear association between natural logarithmic hemoglobin A1C and hip circumference. We observed no statistically significant linear association between natural logarithmic hemoglobin A1C and hip circumference (t-statistic = 0.28, d.f. = 1, p-value = 0.7765). Since there was no statistically significant linear observation between natural logarithmic hemoglobin A1C and hip circumference, we have deemed the regression coefficient describing the linear association between the two variables as unreliable (ß = 0.01104) (summarized in table 7D). Then, we examined the linear association between natural logarithmic hemoglobin A1Cand waist-to-hip circumference ratio (t-statistic = 0.58, d.f. = 1, p-value = 0.5623). Due to the linear association between not being statistically significant, we have deemed the regression coefficient describing the linear association between natural logarithmic hemoglobin A1Cand waist-to-hip circumference ratio as unreliable (ß = 1.12569) (see Table 7D). We observed collinearity in each predictor of the model as the variance influence factors for each predictor was greater than 10 (summarized in Table 7D). We also did not observe outliers in the model as the predicted values were observed to be in the range of 1.4-1.8. There were also no influence points as Cook’s D values were not observed to be >1 (see Appendix 7.7A). The best variable to utilize with natural logarithmic hemoglobin A1C was observed to be waist-to-hip circumference ratio as the variance inflation factor was observed to be the lowest among the predictors (see Table 7D). In addition, the model explained 99.1% of variance between natural logarithmic hemoglobin A1C and waist-to-hip circumference making waist-to-hip circumference ratio the best predictor to observe in a linear model with natural hemoglobin logarithmic A1C (see Table 7E).

Table 7D. Parameter Estimates between Natural Logarithmic Hemoglobin A1C and Hip and Waist Circumference, and Waist-to-Hip Circumference Ratio

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slopes (Beta Estimates) | Standard Error | t-value | P-value | Tolerance | Variance Inflation |
| Intercept | 1 | 0.76882 | 0.18963 | 4.05 | <.0001 |  |  |
| Waist Circumference | 1 | -0.00193 | 0.04462 | -0.04 | 0.9656 | 0.00378 | 264.51 |
| Hip Circumference | 1 | 0.01104 | 0.03885 | 0.28 | 0.7765 | 0.00527 | 189.60 |
| Waist-to-Hip Circumference Ratio | 1 | 1.12569 | 1.94095 | 0.58 | 0.5623 | 0.01190 | 84.02 |

Table 7E. Collinearity Diagnostics of Predictors

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | Proportion of Variance | | |
| Number | Eigenvalue | Condition Index | Waist Circumference | Hip Circumference | Waist-to-Hip Circumference |
| 1 | 1.96863 | 1.00000 | 0.00097441 | 0.00094905 | 0.00083651 |
| 2 | 1.02950 | 1.38283 | 6.803884E-7 | 0.00150 | 0.00817 |
| 3 | 0.00186 | 32.50343 | 0.99902 | 0.99755 | 0.99100 |

**Multivariable Linear Regression Model Between Natural Logarithmic Hemoglobin A1C and BMI Categories, adjusting for Age, Male, Cholesterol, Mean Systolic Blood Pressure, and Waist-to-Hip Ratio (Question 7B)**

We used a series of PROC REG functions to construct various multivariable linear regression models to observe the linear association between natural logarithmic hemoglobin A1C and BMI categories adjusted for various predictors. We also applied a standardization method of the slopes.

In order to observe the linear association between natural logarithmic hemoglobin A1C and BMI Categories, adjusting for Age, Male, Cholesterol, Mean Systolic Blood Pressure, and Waist-to-Hip Ratio, we conducted a multivariable linear regression analysis. We observed that there was a statistically significant linear association in the model (F-statistic = 8.96, d.f. = (6, 126), adj-R2 = 0.2086, p-value <0.0001). Due to there being a statistically linear association in the model, we observed the linear association between natural logarithmic hemoglobin A1C and the main effects. First, we observed the linear association between natural logarithmic hemoglobin A1C and BMI categories. We discerned that there was a statistically significant linear association between natural logarithmic hemoglobin A1C and BMI categories (t-statistic = 2.19, d.f. = 1, p-value = 0.0307). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.17090 units as BMI categories increased (see Table 7F). Then, we observed the linear association between natural logarithmic hemoglobin A1C and age. We discerned that there was a statistically significant linear association between natural logarithmic hemoglobin A1C and age (t-statistic = 3.96, d.f. = 1, p-value < 0.0001). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.33624 units as age increased. Fourth, we observed the linear association between natural logarithmic hemoglobin A1C and cholesterol. We discerned that there was a statistically significant linear association between natural logarithmic hemoglobin A1C and cholesterol (t-statistic = 3.84, d.f. = 1, p-value = 0.0002). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.30229 units as cholesterol increased (see table 7F). Next, we observed the linear association between natural logarithmic hemoglobin A1C and sex of the participant (male variable). We discerned that there was no statistically significant linear association between natural logarithmic hemoglobin A1C and the sex of the participant (t-statistic = 0.68, d.f. = 1, p-value = 0.4989). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.05757 units as sex increased (see Table 7F). Then, we observed the linear association between natural logarithmic hemoglobin A1C and mean systolic blood pressure. We discerned that there was no statistically significant linear association between natural logarithmic hemoglobin A1C and mean systolic blood pressure (t-statistic = -0.42, d.f. = 1, p-value = 0.6775). We also observed that, on average, natural logarithmic hemoglobin A1C decreased -0.03373 units as mean systolic blood pressure increased (see Table 7F). Finally, we observed the linear association between natural logarithmic hemoglobin A1C and waist-to-hip circumference ratio. We discerned that there was a statistically significant linear association between natural logarithmic hemoglobin A1C and waist-to-hip circumference ratio (t-statistic = 1.36, d.f. = 1, p-value = 0.1753). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.11278 units as waist-to-hip circumference ratio increased.

Table 7F. Parameter Estimates of The Multivariable Linear Regression Model Between Natural Logarithmic Hemoglobin A1C and BMI Categories, adjusting for Age, Male, Cholesterol, Mean Systolic Blood Pressure, and Waist-to-Hip Ratio

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slope (Beta Estimates) | Standard Error | t-value | p-value | Standardized Slope Estimates |
| Intercept | 1 | 0.29921 | 0.36678 | 0.82 | 0.4162 | 0 |
| BMI categories | 1 | 0.07709 | 0.03527 | 2.19 | 0.0307 | 0.17090 |
| Age | 1 | 0.00748 | 0.00189 | 3.96 | 0.0001 | 0.33624 |
| Cholesterol | 1 | 0.00216 | 0.000563 | 3.84 | 0.0002 | 0.30229 |
| Male | 1 | 0.04017 | 0.05922 | 0.68 | 0.4989 | 0.05757 |
| Mean Systolic Blood Pressure | 1 | -0.000377 | 0.000904 | -0.42 | 0.6775 | -0.03373 |
| Waist-to-Hip Circumference Ratio | 1 | 0.52083 | 0.38214 | 1.36 | 0.1753 | 0.11278 |

The strongest observed linear association with natural logarithmic hemoglobin A1C was discerned to be age (Standardize Beta = 0.33624). The second strongest linear association with natural logarithmic hemoglobin A1C was cholesterol (Standardized Beta = 0.30229). The next strongest observed linear association with natural logarithmic hemoglobin A1C was BMI categories (Standardized Beta = 0.17090). The fourth strongest linear association with natural logarithmic hemoglobin A1C was waist-to-hip circumference ratio (Standardized Beta = 0.01641). The fifth strongest linear association with natural logarithmic hemoglobin A1C was observed to be the male variable (Standardized Beta = 0.05757). The sixth strongest linear association with natural logarithmic hemoglobin A1C was observed to be mean systolic blood pressure (Standardized Beta = -0.03373).

**Observing the Effect of Age on BMI Categories**

In order to observe the linear association between natural logarithmic hemoglobin A1C and BMI categories adjusting for age, we conducted a multivariable linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and BMI categories adjusting for age (F-statistic = 36.98, d.f. = (2, 370), adj-R2 = 0.1621, p-value <0.0001). The model explains 16.21% of variance between natural logarithmic hemoglobin A1C and BMI categories, adjusting for age. Due to there being a statistically significant linear association in the model, we examined the linear association of natural logarithmic hemoglobin A1C with the main effects. The first of these analyses observed the linear association between natural logarithmic hemoglobin A1C and BMI categories. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and BMI categories (t-statistic = 2.48, d.f. = 1, p-value = 0.0138). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.11758 units as BMI categories increased. The second analysis we conducted observed the linear association between natural logarithmic hemoglobin A1C and age. We discerned a statistically significant linear association between natural logarithmic hemoglobin A1C and age (t-statistic = 8.12, d.f. = 1, p-value <0.0001). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.38624 units as age increased. These linear association estimates between natural logarithmic hemoglobin A1C and BMI categories adjusting for age, are summarized in Table 7G. We also observed if age was a confounder on the linear association between natural logarithmic hemoglobin A1C and BMI categories. According to Figure 7.1, the observed confounding estimate did not pass the 10% (<10%), therefore, age confounds the linear relationship between natural logarithm hemoglobin AlC and BMI categories. Also, age was observed to be a covariate as there was a statistically significant linear association between age and natural logarithmic hemoglobin A1C (summarized in Table 7G).

Table 7G. Parameter Estimates of The Multivariable Linear Regression Model Between Natural Logarithmic Hemoglobin A1C and BMI Categories, adjusting for Age

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slope (Beta Estimates) | Standard Error | t-value | p-value | Standardized Slope Estimates |
| Intercept | 1 | 1.22934 | 0.05765 | 21.32 | <.0001 | 0 |
| BMI categories | 1 | 0.04506 | 0.01820 | 2.48 | 0.0138 | 0.11758 |
| Age | 1 | 0.00732 | 0.00090 | 8.13 | <.0001 | 0.38603 |

Figure 7.1 Confounding Estimate of Age on BMI Categories

**Effect of Cholesterol on BMI Status**

In order to observe the linear association between natural logarithmic hemoglobin A1C and BMI categories adjusting for cholesterol, we conducted a multivariable linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and BMI status adjusting for cholesterol (F-statistic = 17.91, d.f. = (2, 370), adj-R2 = 0.0833, p-value <0.0001). The model explains 8.33% of variance between natural logarithmic hemoglobin A1C and BMI status adjusting for cholesterol. Due to a statistically significant linear association being observed in the model, we examined the linear association of natural logarithmic hemoglobin A1C with the main effects. The first of these analyses observed the linear association between natural logarithmic hemoglobin A1C and obesity status. We discerned that there was a statistically significant linear association between natural logarithmic hemoglobin A1C and BMI categories (t-statistic = 2.08, d.f. = 1, p-value = 0.0381). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.10395 units as BMI categories increased (summarized in table 7H). The second analysis we performed observed the linear association between natural logarithmic hemoglobin A1C and cholesterol. We discerned that there was a statistically significant linear association between natural logarithmic hemoglobin A1C and cholesterol (t-statistic = 5.35, d.f. = 1, p-value <0.0001). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.26701 units as cholesterol increased (summarized in table 7H). According to Figure 7.2, we observed that cholesterol confounds the linear relationship between natural logarithmic hemoglobin A1C and BMI categories as the confounding estimate does pass the 10% rule (<10%).

Table 7H. Parameter Estimates of The Multivariable Linear Regression Model Between Natural Logarithmic Hemoglobin A1C and BMI Categories, adjusting for Cholesterol

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slope (Beta Estimates) | Standard Error | t-value | p-value | Standardized Slope Estimates |
| Intercept | 1 | 1.18796 | 0.08145 | 14.59 | <.0001 | 0 |
| BMI categories | 1 | 0.03983 | 0.01914 | 2.08 | 0.0381 | 0.10395 |
| Cholesterol | 1 | 0.00190 | 0.00035460 | 5.35 | <.0001 | 0.26701 |

Figure 7.2 Confounding Estimates of Cholesterol on BMI Categories

**Effect of Sex (Male variable) on BMI status**

In order to observe the linear association between natural logarithmic hemoglobin A1C and BMI status, adjusting for sex of the participant, we conducted a multivariable linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and BMI status, adjusting for sex of the participant (F-statistic = 4.90, d.f. = (2, 370), adj-R2 = 0.0205, p-value = 0.0079). The model explains 2.05% of the variance between natural logarithmic hemoglobin A1C and BMI status, adjusting for sex of the participant. Due to there being an observed statistically significant linear association in the model, we examined the linear association between natural logarithmic hemoglobin A1C and the main effects. The first of these analyses examined the linear association between natural logarithmic hemoglobin A1C and bmi categories. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and obesity (t-statistic = 2.94, d.f. = 1, p-value = 0.0035). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.15518 units as obesity increased. The final analysis conducted observed the linear association between natural logarithmic hemoglobin A1C and sex of the participant. We observed no statistically significant linear association between natural logarithmic hemoglobin A1C and male (t-statistic = 1.74, d.f. = 1, p-value = 0.0826). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.09190 units as male increased. According to figure 7.3, sex of the participant (male variable) did not confound the linear relationship between natural logarithmic hemoglobin A1C and BMI categories as the confounding estimate (9.20%) did meet the criterion of the 10% rule (<10%). Furthermore, male was observed not to be a covariate as the linear association between natural logarithmic hemoglobin A1C and male was not statistically significant (summarized in Table 7I).

Table 7I. Parameter Estimates of The Multivariable Linear Regression Model Between Natural Logarithmic Hemoglobin A1C and BMI Categories, adjusting for Male variable

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slope (Beta Estimates) | Standard Error | t-value | p-value | Standardized Slope Estimates |
| Intercept | 1 | 1.51623 | 0.05045 | 30.05 | <.0001 | 0 |
| BMI categories | 1 | 0.05946 | 0.02023 | 2.94 | 0.0035 | 0.15518 |
| Male | 1 | 0.05858 | 0.03366 | 1.74 | 0.0826 | 0.09190 |

Figure 7.3 Confounding Estimate of Male Variable on BMI Categories

**Effect of Mean Systolic Blood Pressure on BMI Categories**

In order to observe the linear association between natural logarithmic hemoglobin A1C and BMI status, adjusting for mean systolic blood pressure, we conducted a multivariable linear regression model. We observed no statistically linear association between natural logarithmic hemoglobin A1C and BMI categories, adjusting for mean systolic blood pressure (F-statistic = 2.55, d.f. = (2, 131), adj-R2 = 0.0228, p-value = 0.0821). The model explains 2.28% of the variance between natural logarithmic hemoglobin A1C and BMI status, adjusting for mean systolic blood pressure. Since we did not observe a statistically significant linear association in the model, the following analyses were deemed to be unreliable in describing the true linear relationship between natural logarithmic hemoglobin A1C and BMI categories, adjusting for mean systolic blood pressure. We observed the linear association between natural logarithmic hemoglobin A1C and the main effects. The first analysis examined the linear association between natural logarithmic hemoglobin A1C and BMI categories. We observed no statistically significant linear association between natural logarithmic hemoglobin A1C and BMI categories (t-statistic = 1.59, d.f. = 1, p-value = 0.1134). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.13661 units as BMI categories increased. The final analysis we performed examined the linear association between natural logarithmic hemoglobin A1C and mean systolic blood pressure. We observed no statistically significant linear association between natural logarithmic hemoglobin A1C and mean systolic blood pressure (t-statistic = 1.62, d.f. = 1, p-value 0.1068). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.00295 units as mean systolic blood pressure increased. Mean systolic blood pressure was not a confounder but was observed to be a covariate due to the linear association between natural logarithmic hemoglobin A1C not being statistically significant (see Table 7J).

Table 7J. Parameter Estimates of The Multivariable Linear Regression Model Between Natural Logarithmic Hemoglobin A1C and BMI Categories, adjusting for Mean Systolic Blood Pressure

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slope (Beta Estimates) | Standard Error | t-value | p-value | Standardized Slope Estimates |
| Intercept | 1 | 1.21887 | 0.24122 | 30.05 | <.0001 | 0 |
| BMI categories | 1 | 0.06112 | 0.03835 | 1.59 | 0.1134 | 0.13661 |
| Mean SBP | 1 | 0.00156 | 0.00095861 | 1.62 | 0.1068 | 0.13923 |

**Effect of Waist-to-Hip Circumference Ratio on BMI Categories**

In order to observe the linear association between natural logarithmic hemoglobin A1C and BMI status, adjusting for waist-to-hip circumference ratio, we conducted a multivariable linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and BMI status, adjusting for waist-to-hip circumference ratio (F-statistic = 13.26, d.f. = (2, 368), adj-R2 = 0.0621, p-value <0.0001). The model explains 6.21% of variance between natural logarithmic hemoglobin A1C and BMI categories, adjusting for waist-to-hip circumference ratio. Due to an observed statistically significant linear association in the model, we examined the linear association between natural logarithmic hemoglobin A1C and the main effects. The first analysis examined the linear natural logarithmic hemoglobin A1C and BMI categories. We observed no statistically significant linear association between natural logarithmic hemoglobin A1C and BMI categories (t-statistic = 1.93, d.f. = 1, p-value = 0.0548). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.09809 units as BMI categories increased. The final analysis examined the linear association between natural logarithmic hemoglobin A1C and waist-to-hip circumference ratio. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and waist-to-hip circumference ratio (t-statistic = 4.43, d.f. = 1, p-value <0.0001). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.22579 units as waist-to-hip circumference ratio increased. According to Figure 7.4, waist-to-hip circumference ratio was observed to confound the linear relationship between natural logarithmic hemoglobin A1C and BMI categories as it did not pass the 10% rule (<10%).

Table 7K. Parameter Estimates of The Multivariable Linear Regression Model Between Natural Logarithmic Hemoglobin A1C and BMI Categories, adjusting for Waist-to-Hip Circumference Ratio

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slope (Beta Estimates) | Standard Error | t-value | p-value | Standardized Slope Estimates |
| Intercept | 1 | 0.74446 | 0.18936 | 3.93 | 0.0001 | 0 |
| BMI categories | 1 | 0.03744 | 0.01943 | 1.93 | 0.0548 | 0.09809 |
| Waist-to-Hip Circumference Ratio | 1 | 0.95763 | 0.21594 | 4.43 | <.0001 | 0.22579 |

Figure 7.4 Confounding Estimate of Waist-to-Hip Circumference Ratio on BMI categories

1. **LASSO and Backwards Tests for Multivariable Linear Regression Model in 7B (Question 8)**

We used a PROC GLMSELECT function to observe the linear association between natural logarithmic hemoglobin A1C and BMI status, adjusting for age, male, cholesterol, mean systolic blood pressure, and waist-to-hip ratio. We applied a LASSO procedure to control for the overestimation of the slopes describing linear association between the outcome and the predictors. We also performed a backwards procedure to the model to compare raw slope estimates with the LASSO procedure slope estimates.

**LASSO Procedure for Multivariable Linear Regression Model Between Natural Logarithmic Hemoglobin A1C and BMI Status, Adjusting for Various Predictors**

In order to observe the standardized linear association between natural logarithmic hemoglobin A1C and BMI status, adjusting for age, male, cholesterol, mean systolic blood pressure, and waist-to-hip ratio, we conducted a LASSO multivariable linear regression with an AIC of best fit test to observe underfitting and overfitting of the model. (According to Figure 8.1, we observed 4 variables that were selected by the model which were identified to be age (std Beta = 0.006385), cholesterol (std Beta = 0.001893), waist-to-hip circumference ratio (std Beta = 0.468527), and categorical BMI (std Beta = 0.050726). According to Figure 8.2, male variable and mean systolic blood pressure were removed from the model as selection by AIC stopped at mean systolic blood pressure, indicating that the model was overfitted. Also, BMI categories was observed to be the best criterion value for the LASSO linear regression model and was the selected step by AIC. The parameter estimates describing the linear association of the selected variables are summarized in Table 7L.

Table 7L. LASSO Regression Parameter Estimates of Natural Logarithmic Hemoglobin A1C

|  |  |  |
| --- | --- | --- |
| Variable | Degrees of Freedom | Slope Estimates |
| Intercept | 1 | 0.449709 |
| BMI Categories | 1 | 0.050726 |
| Age | 1 | 0.006385 |
| Cholesterol | 1 | 0.001893 |
| Waist-to-Hip Circumference Ratio | 1 | 0.468527 |

**Applying the Multivariable Linear Regression Model Assumptions Test**

According to Figure 8.3 row 1 plot 1, the model does not meet the assumption of homoscedasticity as the variances between the predicted values and the residuals are tightly packed at lower levels of x but become spread out at higher levels of x such as at x = 2.0, indicating unequal variances. Furthermore, the residuals are not scattered across the Residual = 0 horizontal line as the residuals are tightly packed along the line, indicating that the assumption of linearity is not met (see Figure 8.3 row 1 plot 1). The assumption of independence is met as the data was collected from two different rural counties in Virginia. According to Figure 8.3, the assumption of normality was met as the histogram displayed in row 3 plot 1 is distributed in a bell-shaped curve. In addition, the residuals are normally distributed in the Q-Q plot displayed in row 2 plot 1, indicating normality in the distribution of the residuals. Despite 2 assumptions meeting the assumption of multivariable linear regression, not all assumptions were satisfied, therefore, the multivariable linear regression model does not meet the assumptions of multivariable linear regression.

Figure 8.1 LASSO Linear Association between Natural Logarithmic Hemoglobin A1C and BMI Status, adjusting for Age, Cholesterol, Waist-to-Hip Circumference Ratio and Mean Systolic Blood Pressure

**A graph of a number of steps

Description automatically generated**

Figure 8.2 Test of Best Fit for Natural Logarithmic Hemoglobin A1C and Predictors

**A graph of a diagram

Description automatically generated with medium confidence**

Figure 8.3 Diagnostics of Natural Logarithmic Hemoglobin A1C and Predictors

A screenshot of a graph

Description automatically generated

**Backwards Linear Regression Model of Natural Logarithmic Hemoglobin A1C**

**and BMI Categories adjusted for Various Predictors**

In order to observe the linear association between natural logarithmic hemoglobin A1C and BMI categories, adjusting for age, male, cholesterol, mean systolic blood pressure, and waist-to-hip ratio, we performed a backwards elimination multivariable linear regression model. We observed the removal of the male variable (F-statistic = 0.17, p-value = 0.6775) and mean systolic blood pressure (F-statistic = 0.49, p-value = 0.4852) due not meeting the criterion significance level of p-value < 0.15. The backwards elimination procedure was stopped at BMI categories as it was set as the next candidate for removal but was observed to meet the criterion significance level of p-value < 0.15 (p-value = 0.0799). Four variables were retained within the model and observed for the linear association with natural logarithmic hemoglobin A1C (F-statistic = 13.41, d.f. = (4, 128), adj-R2 = 0.2733). The model explains 27.33% of variance between natural logarithmic hemoglobin A1C and BMI categories, adjusted for age, cholesterol, and waist-to-hip circumference ratio. On average, Natural logarithmic hemoglobin A1C increased 0.070091 units as BMI categories increased. We observed that, on average, natural logarithmic hemoglobin A1C increased 0.007064 units as age increased. We also discerned that, on average, natural logarithmic hemoglobin A1C increased 0.002099 units as cholesterol increased. We observed that, on average, natural logarithmic hemoglobin A1C increased 0.615375 units as waist-to-hip circumference ratio (results summarized in Table 8B). The backward elimination linear regression model agrees with the LASSO linear regression model in that both models retained BMI categories, age, cholesterol, and waist-to-hip ratio and removed mean systolic blood pressure and male from the analysis. The main difference between the two models is that the backwards elimination regression model deemed the male variable to be the optimal criterion in shrinking variance whereas the LASSO regression model deemed BMI categories to be the optimal criterion for shrinking variance.

Table 8B. Multivariable Linear Regression Slope Estimates of Dependent Variable: Natural Logarithmic Hemoglobin A1C

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Degrees of Freedom | Slope (Beta estimate) | Standard Error | t-value |
| Intercept | 1 | 0.195868 | 0.325931 | 0.60 |
| BMI categories | 1 | 0.070091 | 0.033610 | 2.09 |
| Age | 1 | 0.007064 | 0.001748 | 4.04 |
| Cholesterol | 1 | 0.002099 | 0.000553 | 3.79 |
| Waist-to-Hip Circumference Ratio | 1 | 0.615375 | 0.348562 | 1.77 |

**IV. Conclusion**

We determined that there is a significant linear association between hemoglobin A1C and BMI. A PROC MEANS function was utilized to obtain descriptive statistics for continuous variables including the sample size, mean, median, standard deviation, and minimum and maximum values (see Table 2). Various characteristics in relation to cardiovascular health were measured and varied greatly amongst participants. The mean age of the sample was observed to be 46.71 years with participant age ranging between 19-92 years. Mean BMI was observed to be 29.13 kg/m2 with participant BMI ranging between 19.1 – 55.8 kg/m2. 137 observations for mean systolic blood pressure were used as 249 participants were missing 2nd measurements for diastolic and systolic blood pressure. Mean systolic blood pressure was observed to be 230.66 mmHgThe outcomes of this study are hemoglobin A1C and natural logarithmic hemoglobin A1C. Mean hemoglobin A1C was observed to be 5.60% and mean natural logarithmic hemoglobin A1C was observed to be 1.67 units.

We conducted two separate ANOVA models for hemoglobin A1C and natural logarithmic hemoglobin A1C. We utilized the natural logarithmic hemoglobin variable to continue our analysis based on the results. In order to test for differences in mean natural logarithmic hemoglobin A1C by BMI category, we performed a one-way ANOVA with post hoc testing using Tukey’s procedure applying a two-sided alpha of 0.05. According to Table 4B, we observed a statistically significant difference in mean natural logarithmic hemoglobin A1C levels across the three different BMI categories (F-statistic = 3.36, d.f. = (2, 370), R2 = 0.0179, p-value = 0.0357). The model explains 1.79% of variance between mean natural logarithmic hemoglobin A1C and BMI categories. Due to the difference in mean natural logarithmic hemoglobin A1C by BMI category being statistically significant, we observed the association between the main effects and mean natural logarithmic hemoglobin A1C. The normal BMI category was determined to be the reference group for this analysis. We observed a statistically significant association between the obese BMI category and mean logarithmic hemoglobin A1C (t-statistic = 2.59, d.f. = 1, p-value = 0.0099). On average, mean logarithmic hemoglobin A1C increased 0.10273 units as the obese category increased. In contrast, we did not observe a statistically significant association between participants identified as overweight and mean logarithmic hemoglobin A1C (t-statistic = 1.33, d.f = 1, p-value = 0.1829). According to Table 4C, we observed that on average, mean logarithmic hemoglobin A1C increased 0.0555 units as the overweight BMI category increased. In post hoc testing using Tukey’s procedure to compare pairs of group means, we found statistically significant differences only between normal and obese study groups (p-value = 0.0268). Mean logarithmic hemoglobin A1C was observed to be higher in the obese study group in comparison to the normal study group (see Appendix A). We determined that that the rest of the statistical analyses in this paper was more appropriately conducted with logarithmic hemoglobin A1C based on our conclusions in part B.

We determined that the piecewise linear regression model would be the best fit of the multivariable linear regression model. In order to test the linear association between natural logarithmic hemoglobin A1C and piecewise BMI variables: BMI1 BMI2 and BMI3, we utilized a piecewise linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C and the piecewise BMI variables (F-statistic = 3.66, d.f. = (3, 369), adj-R2 = 0.0210, p-value = 0.0127). The model explains 2.10% of the variance between natural logarithmic hemoglobin A1C and the piecewise BMI variables: BMI1, BMI2, and BMI3. Due to the observation of a statistically significant linear association between natural logarithmic hemoglobin A1C and piecewise BMI we examined the association between the main effects and natural logarithmic hemoglobin A1C. We first examined the association between natural logarithmic hemoglobin A1C and BMI1 and discerned that there was no statistically significant linear association (t-statistic = 1.93, d.f. = 1, p-value = 0.0542). We discerned that, on average, natural logarithmic hemoglobin A1C increased 0.12641 units as BMI1 increased. Next, we examined the linear association between natural logarithmic hemoglobin A1C and BMI2 and observed no statistically significant linear association (t-statistic = 0.55, d.f. = 1, p-value = 0.5570). We discerned that, on average, natural logarithmic hemoglobin A1C increased 0.04101 units as BMI2 increased. Finally, we examined the linear association between natural logarithmic hemoglobin A1C and BMI3 and observed no statistically significant linear association (t-statistic =0.56, d.f. = 1, p-value = 0.5770). We observed that, on average, natural logarithmic hemoglobin A1C increased 0.03401 units as BMI3 increased. The standardized slope estimates describing the linear association between natural logarithmic hemoglobin A1C and BMI piecewise variables: BMI1, BMI2, BMI3 are summarized in table 5D.

In order to test the linear association between natural logarithmic hemoglobin A1C, BMI categories, and age with interaction, we conducted an ANCOVA test. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C, BMI categories, and age with interaction (F-statistic = 15.68, d.f. = (5, 367), R2 = 0.1760, p-value < 0.0001). The ANCOVA model explains 17.60% of the variance between natural logarithmic hemoglobin A1C, BMI categories, and age with interaction. We then determined to observe the association between the main effects and natural logarithmic hemoglobin A1C. We observed the linear association between natural logarithmic hemoglobin A1C and BMI categories, and age with interaction and discerned that there was no statistically significant linear association between natural logarithmic hemoglobin A1C and the interaction term (F-statistic = 2.04, d.f. = 2, p-value = 0.1309). Due to the linear association between natural logarithmic hemoglobin A1C and the interaction between BMI categories and age being statistically not significant, we recommend conducting the analysis removing the interaction term to conduct further analysis of the main effects (summarized in table 6A). Furthermore, due to the interaction between BMI categories and age not being statistically significant we can safely conclude that age is not an effect measure modifier. In order to test the linear association between natural logarithmic hemoglobin A1C, BMI categories, and male with interaction, we conducted an ANCOVA test. We observed no statistically significant linear association between natural logarithmic hemoglobin A1C, BMI categories, and male with interaction (F-statistic = 1.96, d.f. = (5, 367), R2 = 0.0260, p-value = 0.0844). The model explains 2.60% of variance between natural logarithmic hemoglobin A1C, BMI categories, and male with interaction. Since we discerned that there was no statistically significant linear association between natural logarithmic hemoglobin A1C, BMI categories, and male with interaction, we recommend removing the interaction term from the model to conduct further analysis of the main effects. Furthermore, we can safely assume that male is not an effect measure modifier based on these facts.

In order to observe the linear association between natural logarithmic hemoglobin A1C, hip circumference, waist circumference, and waist-to-hip circumference ratio, we performed a multivariable linear regression model. We observed a statistically significant linear association between natural logarithmic hemoglobin A1C, hip circumference, waist circumference, and waist-to-hip circumference ratio (F-statistic = 11.40, d.f. = (3, 367), adj-R2 = 0.0777, p-value < .0001). The model explains 7.77% of variance between natural logarithmic hemoglobin A1C and hip circumference, waist circumference, and waist-to-hip circumference ratio. Due to a statistically significant linear association observed in the model, we examined the linear association of natural logarithmic hemoglobin A1C with the main effects. We observed the linear association between natural logarithmic hemoglobin A1C and waist circumference and discerned there was no statistically significant linear association (t-statistic = -0.04, d.f. = 1, p-value = 0.9656). Due to the there being no statistically significant linear association the slope describing the linear association between natural logarithmic hemoglobin A1C and waist circumference is deemed unreliable (ß = -0.00193) (see Table 7D). Next, we examined the linear association between natural logarithmic hemoglobin A1C and hip circumference. We observed no statistically significant linear association between natural logarithmic hemoglobin A1C and hip circumference (t-statistic = 0.28, d.f. = 1, p-value = 0.7765). Since there was no statistically significant linear observation between natural logarithmic hemoglobin A1C and hip circumference, we have deemed the regression coefficient describing the linear association between the two variables as unreliable (ß = 0.01104) (summarized in table 7D). Then, we examined the linear association between natural logarithmic hemoglobin A1Cand waist-to-hip circumference ratio (t-statistic = 0.58, d.f. = 1, p-value = 0.5623). Due to the linear association between not being statistically significant, we have deemed the regression coefficient describing the linear association between natural logarithmic hemoglobin A1Cand waist-to-hip circumference ratio as unreliable (ß = 1.12569) (see Table 7D). We observed collinearity in each predictor of the model as the variance influence factors for each predictor was greater than 10 (summarized in Table 7D). We also did not observe outliers in the model as the predicted values were observed to be in the range of 1.4-1.8. There were also no influence points as Cook’s D values were not observed to be >1 (see Appendix 7.7A). The best variable to utilize with natural logarithmic hemoglobin A1C was observed to be waist-to-hip circumference ratio as the variance inflation factor was observed to be the lowest among the predictors (see Table 7D). In addition, the model explained 99.1% of variance between natural logarithmic hemoglobin A1C and waist-to-hip circumference making waist-to-hip circumference ratio the best predictor to observe in a linear model with natural hemoglobin logarithmic A1C (see Table 7E).

We used a series of PROC REG functions to construct various multivariable linear regression models to observe the linear association between natural logarithmic hemoglobin A1C and BMI categories adjusted for various predictors. We also applied a standardization method of the slopes. In order to observe the linear association between natural logarithmic hemoglobin A1C and BMI Categories, adjusting for Age, Male, Cholesterol, Mean Systolic Blood Pressure, and Waist-to-Hip Ratio, we conducted a multivariable linear regression analysis. We observed that there was a statistically significant linear association in the model (F-statistic = 8.96, d.f. = (6, 126), adj-R2 = 0.2086, p-value <0.0001). Due to there being a statistically linear association in the model, we observed the linear association between natural logarithmic hemoglobin A1C and the main effects. First, we observed the linear association between natural logarithmic hemoglobin A1C and BMI categories. We discerned that there was a statistically significant linear association between natural logarithmic hemoglobin A1C and BMI categories (t-statistic = 2.19, d.f. = 1, p-value = 0.0307). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.17090 units as BMI categories increased (see Table 7F). Then, we observed the linear association between natural logarithmic hemoglobin A1C and age. We discerned that there was a statistically significant linear association between natural logarithmic hemoglobin A1C and age (t-statistic = 3.96, d.f. = 1, p-value < 0.0001). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.33624 units as age increased. Fourth, we observed the linear association between natural logarithmic hemoglobin A1C and cholesterol. We discerned that there was a statistically significant linear association between natural logarithmic hemoglobin A1C and cholesterol (t-statistic = 3.84, d.f. = 1, p-value = 0.0002). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.30229 units as cholesterol increased (see table 7F). Next, we observed the linear association between natural logarithmic hemoglobin A1C and sex of the participant (male variable). We discerned that there was no statistically significant linear association between natural logarithmic hemoglobin A1C and the sex of the participant (t-statistic = 0.68, d.f. = 1, p-value = 0.4989). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.05757 units as sex increased (see Table 7F). Then, we observed the linear association between natural logarithmic hemoglobin A1C and mean systolic blood pressure. We discerned that there was no statistically significant linear association between natural logarithmic hemoglobin A1C and mean systolic blood pressure (t-statistic = -0.42, d.f. = 1, p-value = 0.6775). We also observed that, on average, natural logarithmic hemoglobin A1C decreased -0.03373 units as mean systolic blood pressure increased (see Table 7F). Finally, we observed the linear association between natural logarithmic hemoglobin A1C and waist-to-hip circumference ratio. We discerned that there was a statistically significant linear association between natural logarithmic hemoglobin A1C and waist-to-hip circumference ratio (t-statistic = 1.36, d.f. = 1, p-value = 0.1753). We also observed that, on average, natural logarithmic hemoglobin A1C increased 0.11278 units as waist-to-hip circumference ratio increased. According to Figure 7.1, the observed confounding estimate did not pass the 10% (<10%), therefore, age confounds the linear relationship between natural logarithm hemoglobin AlC and BMI categories. Also, age was observed to be a covariate as there was a statistically significant linear association between age and natural logarithmic hemoglobin A1C (summarized in Table 7G). According to Figure 7.2, we observed that cholesterol confounds the linear relationship between natural logarithmic hemoglobin A1C and BMI categories as the confounding estimate does pass the 10% rule (<10%). Mean systolic blood pressure was not a confounder but was observed to be a covariate due to the linear association between natural logarithmic hemoglobin A1C not being statistically significant (see Table 7J). According to Figure 7.4, waist-to-hip circumference ratio was observed to confound the linear relationship between natural logarithmic hemoglobin A1C and BMI categories as it did not pass the 10% rule (<10%).

In order to observe the standardized linear association between natural logarithmic hemoglobin A1C and BMI status, adjusting for age, male, cholesterol, mean systolic blood pressure, and waist-to-hip ratio, we conducted a LASSO multivariable linear regression with an AIC of best fit test to observe underfitting and overfitting of the model. (According to Figure 8.1, we observed 4 variables that were selected by the model which were identified to be age (std Beta = 0.006385), cholesterol (std Beta = 0.001893), waist-to-hip circumference ratio (std Beta = 0.468527), and categorical BMI (std Beta = 0.050726). According to Figure 8.2, male variable and mean systolic blood pressure were removed from the model as selection by AIC stopped at mean systolic blood pressure, indicating that the model was overfitted. Also, BMI categories was observed to be the best criterion value for the LASSO linear regression model and was the selected step by AIC. The parameter estimates describing the linear association of the selected variables are summarized in Table 7L. According to Figure 8.3 row 1 plot 1, the model does not meet the assumption of homoscedasticity as the variances between the predicted values and the residuals are tightly packed at lower levels of x but become spread out at higher levels of x such as at x = 2.0, indicating unequal variances. Furthermore, the residuals are not scattered across the Residual = 0 horizontal line as the residuals are tightly packed along the line, indicating that the assumption of linearity is not met (see Figure 8.3 row 1 plot 1). The assumption of independence is met as the data was collected from two different rural counties in Virginia. According to Figure 8.3, the assumption of normality was met as the histogram displayed in row 3 plot 1 is distributed in a bell-shaped curve. In addition, the residuals are normally distributed in the Q-Q plot displayed in row 2 plot 1, indicating normality in the distribution of the residuals. Despite 2 assumptions meeting the assumption of multivariable linear regression, not all assumptions were satisfied, therefore, the multivariable linear regression model does not meet the assumptions of multivariable linear regression.

We then confirmed the LASSO regression results by conducting a backward elimination linear regression model. In order to observe the linear association between natural logarithmic hemoglobin A1C and BMI categories, adjusting for age, male, cholesterol, mean systolic blood pressure, and waist-to-hip ratio, we performed a backwards elimination multivariable linear regression model. We observed the removal of the male variable (F-statistic = 0.17, p-value = 0.6775) and mean systolic blood pressure (F-statistic = 0.49, p-value = 0.4852) due not meeting the criterion significance level of p-value < 0.15. The backwards elimination procedure was stopped at BMI categories as it was set as the next candidate for removal but was observed to meet the criterion significance level of p-value < 0.15 (p-value = 0.0799). Four variables were retained within the model and observed for the linear association with natural logarithmic hemoglobin A1C (F-statistic = 13.41, d.f. = (4, 128), adj-R2 = 0.2733). The model explains 27.33% of variance between natural logarithmic hemoglobin A1C and BMI categories, adjusted for age, cholesterol, and waist-to-hip circumference ratio. On average, Natural logarithmic hemoglobin A1C increased 0.070091 units as BMI categories increased. We observed that, on average, natural logarithmic hemoglobin A1C increased 0.007064 units as age increased. We also discerned that, on average, natural logarithmic hemoglobin A1C increased 0.002099 units as cholesterol increased. We observed that, on average, natural logarithmic hemoglobin A1C increased 0.615375 units as waist-to-hip circumference ratio (results summarized in Table 8B). The backward elimination linear regression model is in agreement with the LASSO linear regression model in that both models retained BMI categories, age, cholesterol, and waist-to-hip ratio and removed mean systolic blood pressure and male from the analysis. The main difference between the two models is that the backwards elimination regression model deemed the male variable to be the optimal criterion in shrinking variance whereas the LASSO regression model deemed BMI categories to be the optimal criterion for shrinking variance.

**Appendix B**

/\*Merging Baseline and Clinical data for both cohorts into combined Data Set\*/

libname Project '/home/u63114430/BS 805/Project';

run;

Data new\_baseline;

set project.baseline\_cohort1 project.baseline\_cohort2;

run;

data new\_clinical;

set project.clinical\_cohort1 project.clinical\_cohort2;

run;

data combined;

merge new\_baseline new\_clinical;

by ID;

run;

proc print data = combined;

run;

/\*Temporary dataset for analyses \*/

proc format;

value Cat\_BMIf 1 = 'normal' 2 = 'overweight' 3 = 'obese';

value malef 0 = 'female' 1 = 'male';

run;

data new\_combined;

set combined;

label age = 'Age in years'

height = 'height in inches'

weight = 'weight in pounds'

BMI = 'BMI in kg/m^2'

chol = 'Total cholesterol (mg/dL)'

stab\_glu = 'Blood Glucose Level (mg/dL)'

glyhb = 'Hemoglobin A1C (percent)'

bp\_1s = 'Systolic Blood Pressure, 1st measure (mmHg)'

bp\_2s = 'Systolic Blood Pressure, 2nd measure (mmHg)'

bp\_1d = 'Diastolic Blood Pressure, 1st measure (mmHg)'

bp\_2d = 'Diastolic Blood Pressure, 2nd measure (mmHg)'

waist = 'Waist circumference (inches)'

hip = 'hip circumference (inches)'

Ratio\_WH = 'Waist-to-hip ratio'

Mean\_SBP = 'Mean Systolic Blood Pressure (mmHg)'

ln\_H1AC = 'Natural Log of Hemoglobin A1C';

/\*Accounting for missing variables\*/

/\*2A Creating BMI variable excluding those with BMI <19\*/

BMI = (weight\*0.454)/((height\*0.0254)\*\*2);

if BMI<19 then DELETE;

/\*2B Computing Categorical BMI variable \*/

if BMI = . then Cat\_BMI = .;

else if BMI<25 then Cat\_BMI = 1;

else if 25<= BMI <30 then Cat\_BMI = 2;

else if BMI>=30 then Cat\_BMI = 3;

/\*2C creating Piecewise Variables with BMI \*/

if 0<=BMI<25 then BMI1 = BMI;

else if BMI>=25 then BMI1 = 25;

if 0<=BMI<25 then BMI2 = 25;

else if 25<=BMI<30 then BMI2 = BMI;

else if BMI>=30 then BMI2 = 30;

if 0<=BMI<30 then BMI3 = 30;

else if BMI>=30 then BMI3 = BMI;

if 0<=BMI<25 then BMIgroup = 'Group 1';

else if 25<=BMI<30 then BMIgroup = 'Group 2';

else if BMI>=30 then BMIgroup = 'Group 3';

/\* Creating dummy variables for BMI in 5A \*/

if Cat\_BMI = 2 then overweight = 1; else overweight = 0;

if Cat\_BMI = 3 then obese = 1; else obese = 0;

/\*Question 2D Creating Waist to Hip Ratio \*/

Ratio\_WH = waist/hip;

/\* Question 2E Creating Mean SBP variable \*/

Mean\_SBP = (bp\_1s + bp\_2s/2);

/\* Question 2F Natural Log of H1AC \*/

ln\_H1AC = LOG(glyhb);

/\* Question 2G Creating dummy variable MALE \*/

if gender = 'female' then male = 0;

else if gender = 'male' then male = 1;

format Cat\_BMI Cat\_BMIf. male malef.;

run;

proc print data = new\_combined;

run;

/\* Question 3 descriptive statistics \*/

proc means data= new\_combined N mean median std min max;

var age height weight BMI chol stab\_glu bp\_1s bp\_2s bp\_1d bp\_2d waist hip Ratio\_WH glyhb ln\_H1AC Mean\_SBP;

run;

proc freq data = new\_combined;

tables Cat\_BMI male;

run;

/\*Question 4 One-Way ANOVA analysis between Natural Log and normal H1AC \*/

proc glm data = new\_combined;

class Cat\_BMI (ref = 'normal');

model glyhb = Cat\_BMI/ solution;

means Cat\_BMI;

lsmeans Cat\_BMI / pdiff adjust=tukey;

title 'One Way ANOVA of Hemoglobin and BMI categories';

run;

proc glm data = new\_combined;

class Cat\_BMI (ref = 'normal');

model ln\_H1AC = Cat\_BMI/ solution;

means Cat\_BMI;

lsmeans Cat\_BMI / pdiff adjust=tukey;

title 'One-Way ANOVA of LN Hemoglobin and BMI categories';

run;

/\*Question 5A Using the Natural Log of Hemoglobin for further analysis \*/

Data combined2;

set new\_combined;

label Cat\_BMI = 'BMI categories'

male = 'Sex of participant';

/\*Creating dummy variables for BMI in 5A \*/

if Cat\_BMI = 2 then overweight = 1; else overweight = 0;

if Cat\_BMI = 3 then obese = 1; else obese = 0;

run;

proc reg data = combined2;

model ln\_H1AC = overweight obese;

title 'Simple linear regression of Natural Log H1AC with BMI Dummy Variables';

run;

/\* Question 5B Linear Regression with Ordinal BMI \*/

proc reg data = combined2;

model ln\_H1AC = Cat\_BMI;

title 'Simple Linear Regression of Natural Log H1AC with Ordinal BMI';

run;

/\* Question 5C Linear Regression with Continuous BMI \*/

proc reg data = combined2;

model ln\_H1AC = BMI;

title 'Simple Linear Regression of Natural Log H1AC with Continuous BMI';

run;

/\* Question 5D Linear Regression of Natural Log H1AC with Dummy Variables \*/

proc reg data = combined2;

model ln\_H1AC = BMI1 BMI2 BMI3/ stb;

output out = BMI pred = yhat;

test BMI1 = BMI2;

test BMI2 = BMI3;

run;

title 'Piecewise Linear Regression of Natural Log H1AC with BMI1 BMI2 BMI3 Variables';

run;

/\* Question 6 Testing for Effect Measure Modification of Age and Sex using interaction \*/

proc glm data = combined2;

class Cat\_BMI (ref = 'normal');

model ln\_H1AC = Cat\_BMI|age/ solution;

title 'ANCOVA model of ln\_H1AC with Categorical BMI and Age with Interaction';

run;

proc glm data = combined2;

class Cat\_BMI (ref = 'normal');

model ln\_H1AC = Cat\_BMI|male/ solution;

title 'ANCOVA model of ln\_H1AC with Categorical BMI and Sex with Interaction';

run;

/\*Question 7A Observing Confounding or Covariates between LN H1AC and BMI with Waist, Hip, Waist to Hip Ratio \*/

proc reg data = combined2;

model ln\_H1AC = waist;

title 'Simple Linear Regression of Ln H1AC and Waist';

run;

proc reg data = combined2;

model ln\_H1AC = hip;

title 'Simple Linear Regression of Ln H1AC and Hip';

run;

proc reg data = combined2;

model ln\_H1AC = Ratio\_WH;

title 'Simple Linear Regression of Ln H1AC and Waist to Hip Ratio';

run;

/\* Waist to Hip ratio should be used further in the analysis as it had the least observed VIF value and still explains 99% of

the variance in the model \*/

proc reg data = combined2;

model ln\_H1AC = waist hip Ratio\_WH/ tol VIF collinoint r;

title 'Simple Linear Regression of Ln H1AC and Waist, Hip, and Waist to Hip Ratio';

run;

/\* Question 7b Running a linear regression model with ln H1AC, chol, age, male, Cat BMI, mean SBP, and Waist to Hip ratio

will do confounding for each variable using BMI slope and every other slope \*/

proc reg data = combined2;

model ln\_H1AC = Cat\_BMI age chol male Mean\_SBP Ratio\_WH/r stb;

title 'Multivariable Linear Regression model between ln\_H1AC and Cat BMI with Age Chol Male Mean\_SBP and Waist to Hip Ratio';

run;

/\*Checking for confounding for age when unadjusted \*/

proc reg data = combined2;

model ln\_H1AC = Cat\_BMI age/stb;

title 'Linear Regression model between ln\_H1AC and Cat BMI adjusting for Age ';

run;

/\*checking for confounding for cholesterol when unadjusted \*/

proc reg data = combined2;

model ln\_H1AC = Cat\_BMI chol/ stb;

title 'Linear Regression model between ln\_H1AC and Cat BMI adjusting for Cholesterol';

run;

/\*checking for confounding for male when removed when unadjusted \*/

proc reg data = combined2;

model ln\_H1AC = Cat\_BMI male/ stb;

title 'Multivariable Linear Regression model between ln\_H1AC and Cat BMI with Age Chol Male Mean\_SBP and Waist to Hip Ratio';

run;

/\*checking for confounding for mean SBP when unadjusted \*/

proc reg data = combined2;

model ln\_H1AC = Cat\_BMI Mean\_SBP/stb;

title 'Multivariable Linear Regression model between ln\_H1AC and Cat BMI with Age Chol Male Mean\_SBP and Waist to Hip Ratio';

run;

proc reg data = combined2;

model ln\_H1AC = Cat\_BMI Ratio\_WH/ stb;

title 'Multivariable Linear Regression model between ln\_H1AC and Cat BMI with Age Chol Male Mean\_SBP and Waist to Hip Ratio';

run;

/\* Question 8 Running a LASSO based Linear Regression Model followed by backward regression selection \*/

/\*Compare the Beta estimates of LASSO and Backwards to determine if they are different \*/

proc glmselect data = combined2 plots=all;

model ln\_H1AC = Cat\_BMI age chol male Mean\_SBP Ratio\_WH/ selection=lasso (stop=none choose=aic);

title 'Lasso linear Regression model of ln H1AC';

run;

proc glmselect data = combined2;

model ln\_H1AC = Cat\_BMI age chol male Mean\_SBP Ratio\_WH/ selection=backward (stop=sl choose=aic) sle=0.15 sls=0.15 select=aic;

title 'Backward linear Regression model of ln H1AC';

run;