BS805 On-Line Course Project – Jessy Yang

Predictors of Hemoglobin A1C

Introduction

Type II Diabetes (T2DM) is known to have strong association to obesity and hypertension, both of which are significant contributors to cardiovascular risk factors. Glycosylated hemoglobin (HbA1C), a measure of long-term blood glucose control, serves as a well-established marker for diabetes diagnosis (HbA1c > 7.0) and a crucial indicator of glycemic control. Elevated body mass index (BMI) is a known risk factor for T2DM, and the waist-to-hip ratio is considered associated with both diabetes and heart disease. With the aim of understanding the connections, the primary objective of this project is to explore the relationship between BMI and HbA1c, and the secondary objective involves assessing other clinical factors such as the waist-to-hip ratio that may influence or modify the relationship BMI and HbA1c.

Data

This project uses data sets from two published studies conducted by Dr. John Schorling from the University of Virginia School of Medicine that examined cardiovascular risk factors in central Virginia for African Americans. The first study was conducted on a rural black community in Virginia to evaluate the prevalence of coronary heart disease (CHD) risk factors. The second study was conducted among two rural counties in Virginia to measure the effects of a smoking cessation intervention. Data for the forementioned two different cohorts was combined, containing information on subject ID, demographics (gender and age), physical measurements (e.g. weight, height, waist and hip circumferences, etc.), and lab tests (e.g. total cholesterol, blood glucose level, hemoglobin A1C, blood pressure, etc.). Subjects with BMI lower than 19 were excluded.

Descriptive Statistics

Among 386 subjects, 58.55 were females (n=226) and 41.45% were males (n=160). Average age was 47 years. Subjects had a mean BMI of 29.11, with 28.76% categorized as normal (n=111), 31.87% categorized as overweight (n=123), and 39.38% categorized as obese (n=152). On average, subjects reported a 5.60% for HbA1c. Subjects had a mean total cholesterol of 208.25 mg/dL, a mean systolic blood pressure of 136.47 mm-Hg, and a mean waist-to-hip ratio of 136.47. (Table 1)

Table 1: Sample Characteristics

Variable	n (%)	Mean (SD)
Male (yes)	160 (41.45%)	-
BMI category		
Normal	111 (28.76%)	-
Overweight	123 (31.87%)	-
Obese	152 (39.38%)	-

Age (year)	386	46.71 (16.43)
Height (inch)	386	65.96 (3.92)
Weight (pound)	386	179.57 (39.43)
Total cholesterol (mg/dL)	386	208.25 (43.95)
Blood glucose level (mg/dL)	386	107.16 (53.37)
Hemoglobin A1C (%)	373	5.60 (2.23)
Systolic blood pressure, 1st measure (mm Hg)	382	136.92 (22.95)
Systolic blood pressure, 2nd measure (mm Hg)	137	152.12 (21.87)
Diastolic blood pressure, 1st measure (mm Hg)	382	83.22 (13.34)
Diastolic blood pressure, 2nd measure (mm Hg)	137	92.32 (11.61)
Waist circumference (inch)	384	38.15 (5.66)
Hip circumference (inch)	384	43.28 (5.55)
BMI (kg/m2)	386	29.11 (6.41)
Waist-to-hip ratio	384	0.88 (0.07)
Mean systolic blood pressure (mm Hg)	382	136.47 (22.51)
The natural logarithm of HbA1c	373	1.67 (0.31)

*Total number of samples = 386

Relationship between HbA1c and BMI

One-way ANOVA with post hoc testing using Tukey's procedure was conducted to assess if mean HbA1c is the same across three BMI categories. Two separate models were performed using the original HbA1c and the log transformed HbA1c as outcome variables. BMI was categorized as three groups: normal (BMI< 25), overweight (25 <=BMI < 30), and obese (BMI>= 30).

In the model testing the original HbA1c, we found a statistically non-significant difference at the 0.05 level in the means of HbA1c across the three BMI categories (global F statistic (2, 370) =2.14; p=0.1191; R-squared=0.0114; n=373). In post hoc testing using Tukey's procedure to compare pairs of group means, we found no statistically significant differences at the 0.05 level in the mean HbA1c for any groups (p-value for each pair of groups: normal- overweight, p=0.6127; overweight- obese: p=0.5153; normal- overweight, p=0.0979).

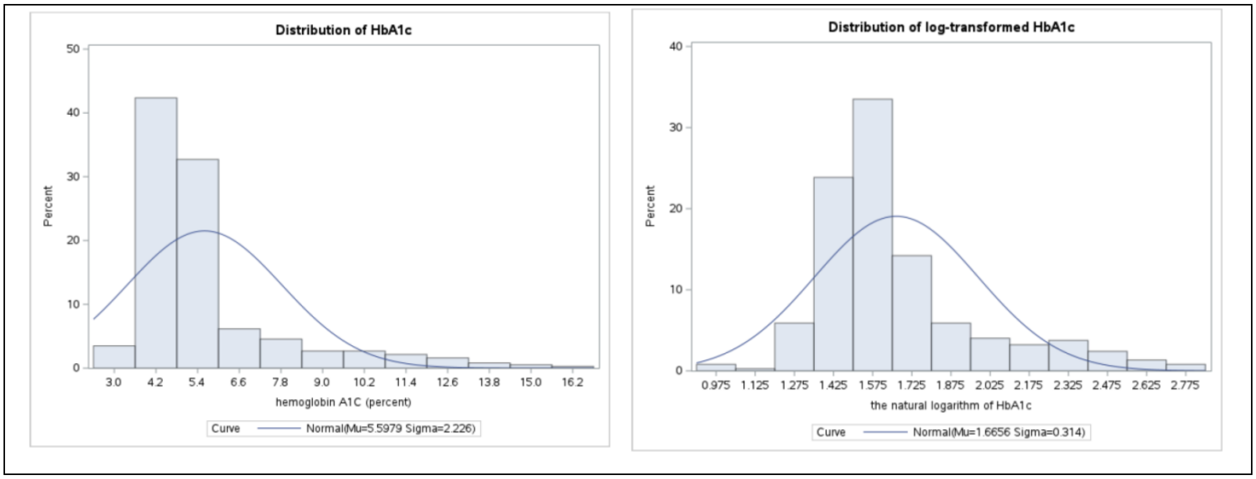
The mean HbA1c for the three BMI categories were: normal (n=107), 5.26%; overweight (n=119), 5.59%; obese (n=147), 5.85%. Overweight subjects have a non-significant 0.32 higher mean HbA1c compared to normal subjects (standard error=0.296; 95%CI=-0.26, 0.91); obese subjects have a significant 0.58 higher mean HbA1c compared to normal subjects (standard error=0.282; 95%CI=0.03, 1.14).

In the model testing the log-transformed HbA1c, we found a statistically significant difference at the 0.05 level in the means of log-transformed HbA1c across the three BMI categories (global F statistic (2, 370) =3.36; p=0.0357; R-squared=0.0179; n=373). In post hoc testing using Tukey's procedure to compare pairs of group means, we found a statistically significant differences at the 0.05 level in the mean log-transformed HbA1c for normal and obese groups, with obese subjects showing greater log-transformed HbA1c on average than normal subjects. (t statistic =2.59; p =0.0268)

The mean log-transformed HbA1c for the three BMI categories were: normal (n=107), 1.61; overweight (n=119), 1.66; obese (n=147), 1.71. Overweight subjects have a non-significant 0.06 higher mean log-transformed HbA1c compared to normal subjects (standard error=0.04; 95%CI=-0.03, 0.14); obese subjects have a significant 0.10 higher mean log-transformed HbA1c compared to normal subjects (standard error=0.04; 95%CI=0.02, 0.18).

By comparing the results, we can see that the three BMI categories do not have significantly different means of HbA1c, but after transforming the HbA1c values to a nature log version, there appears to be a significant difference in the means of log-transformed HbA1c across the three BMI categories. The BMI explains 0.65% more of the variability in log-transformed HbA1c than in the original HbA1c. Moreover, by comparing the histograms, log-transformed HbA1c is more normally distributed than the original HbA1c (Figure 1). Based on these differences we conclude that the log version of HbA1c is more appropriate as the outcome for linear models using BMI categories as the independent variables.

Figure 1. Histogram of HbA1c (left) vs. log-transformed HbA1c (right)



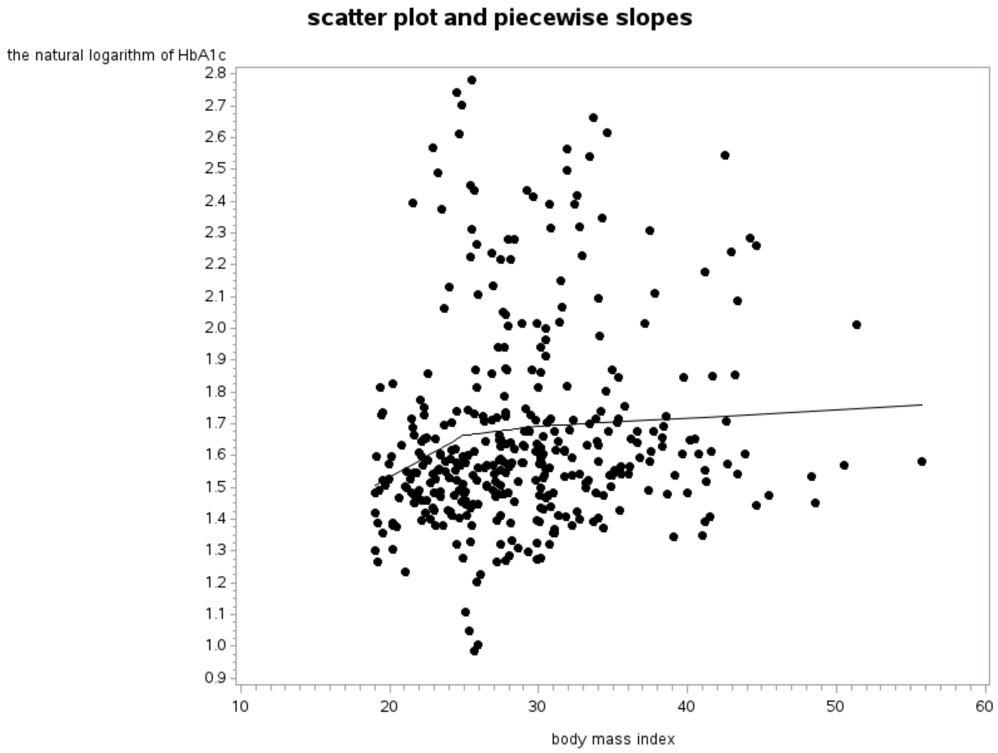
We further tested the association with different forms representing BMI.

Simple linear regression model was performed to examine the association between nature log of HbA1c and BMI category. Although test results showed a statistically significant linear association (t-statistic=2.59; p=0.0098; R-squared=0.0178; n=373), this is not meaningful and not appropriate because the model assumes that there is a linear relationship between the BMI category and nature log of HbA1c, which is not expected to exist across three BMI categories.

We then performed another simple linear regression to examine if there is a linear relationship between nature log of HbA1c and BMI. Results showed that for every unit increase in BMI, nature log of HbA1c increases by 0.0069, suggesting a significant linear association at the 0.05 level. (standard error=0.0025; t-statistic=2.75; p=0.0063; R-squared=0.01995; n=373)

We also conducted a piecewise linear model by dividing BMI into three intervals, using the same cutoffs as BMI categories, to predict nature log of HbA1c. Test results observed a significant linear association between nature log of HbA1c and the three BMI intervals (global F statistic (3, 369) =3.65; p=0.0127; R-squared=0.0289; n=373). However, all three BMI intervals were not statistically significant predictors of nature log of HbA1c. When comparing the effect estimates, model suggested that BMI less than 25 and BMI between 25 and 30 have the same effects on nature log of HbA1c (global F statistic (1, 369) =0.89; p=0.3456), and that BMI between 25 and 30 and BMI greater than or equal to 30 have the same effects on nature log of HbA1c (global F statistic (1, 369) =0.06; p=0.8070). This can be seen from the slopes in the scatter plot, the nature log of HbA1c increases slightly across the BMI range of 20 to 25 and then stays about the same for BMI greater than 25 (Figure 2). Since the piecewise model suggested that the adjacent slopes are statistically the same, the piecewise model is not a better fit than the regular regression model.

Figure 2. BMI vs. Nature log of HbA1c



For modeling the association between BMI and nature log of HbA1c, BMI is preferred over BMI category since it provides more precise effect estimates, whereas model using BMI category only comparing differences between groups. The variability in nature log of HbA1c explained by BMI and BMI category are similar, considering the different number of parameters used in the models (R-squared=0.01995 vs. 0.0179), but BMI has a stronger association with nature log of HbA1c (p=0.0063 vs. p=0.0357).

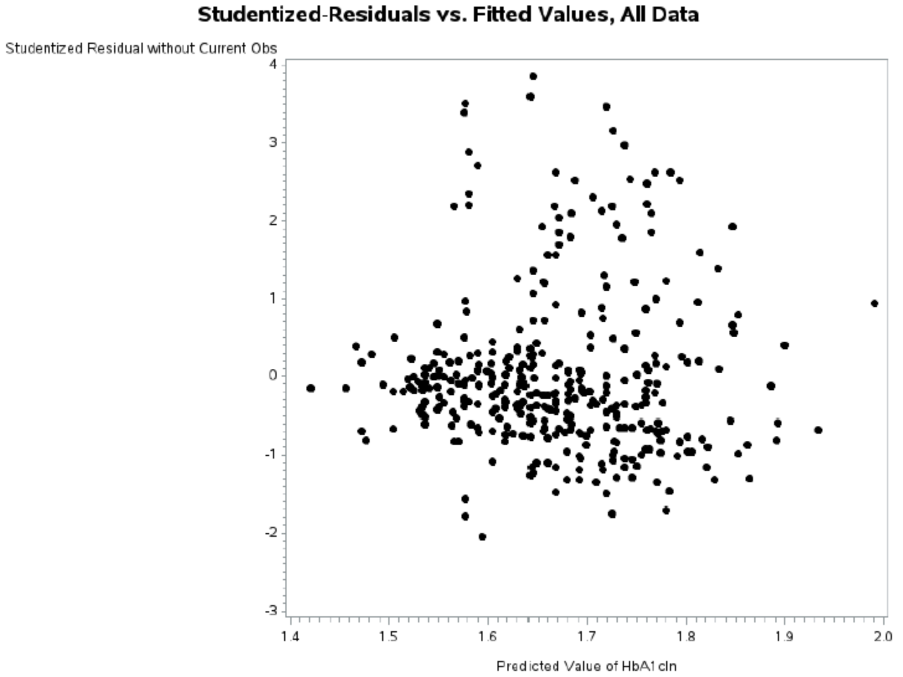
Effect Modifiers, Confounders, and Covariates

Two-factor ANOVA was performed to test the potential effect modifications by age and sex, and results showed no interaction either between age and BMI category (F statistic (5, 367) =2.04; p=0.1309) or between sex and BMI category (F statistic (5, 367) =0.02; p=0.9773). We conclude that age and sex are not effect modifiers of the relationship between nature log of HbA1c and BMI category.

Simple linear regression was conducted for waist, hip, waist-to-hip ratio separately to assess their relationship with nature log of HbA1C. Test results suggested that waist, hip, waist-to-hip ratio all have a statistically significant association with nature log of HbA1c. (p<0.0001, =0.0023, <0.0001, respectively) Multiple linear regression was further performed for all three variables as predictors of nature log of HbA1C. We found large variance inflation factor for waist, hip, waist-to-hip ratio (VIF=264.5, 189.6, 84.0, respectively), suggesting there is potential collinearity between the three predictors. In the principal components-based analysis we also observed a large condition index of 32.5 for principal component 3, by which over 99% of the variance of waist, hip, and waist-to-hip ratio was explained. Since multi-collinearity is present, we cannot include all three variables in a linear model with nature log of HbA1c. Variable waist is preferred over hip and waist-to-hip because it explains the most variance of nature log of HbA1c. (R-squared=0.0717, 0.0249, 0.0578, respectively)

In the plot of studentized residuals by predicted values (Figure 3), we see that almost all the datapoints have studentized residuals within the range of (-2, 2), and only a few has studentized residual outside the range, and thus we consider these datapoints outliers in this model. Several influence points was identified using the rule-of-thumb of Cook's D values greater than 4/n. (see Appendix for the full list for outliers and influential points)

Figure 3. Plot studentized residuals vs. predicted values



Combined the above findings, A multiple linear regression model was fitted for BMI category, age, sex, total cholesterol, mean systolic blood pressure (SBP), and waist all as predictors of nature log of HbA1c. The model observed significant evidence of an association between chest deceleration score and at least one of the predictors (global F statistic (7, 359) =14.86; p<0.0001; Adjusted R-squared=0.2095; n=367).

Among the three BMI categories, overweight subjects have a non-significant 0.019 lower mean nature log of HbA1c compared to normal subjects (standard error=0.041; 95%CI=-0.099, 0.062); obese subjects have a non-significant 0.047 lower mean nature log of HbA1c compared to normal subjects (standard error=0.054; 95%CI=-0.153, 0.058), after adjusting for age, sex, total cholesterol, mean SBP, and waist. The mean nature log of HbA1c for the three BMI categories were: normal, 1.70; overweight, 1.68; obese, 1.65. In post hoc testing using Tukey's procedure to compare pairs of group means, we found no statistically significant differences in the mean nature log of HbA1c for any of the BMI category pairs. Normal subjects had non-significant greater nature log of HbA1c than overweight subjects (t statistic =0.46; p =0.8896); overweight subjects had non-significant greater nature log of HbA1c than obese subjects (t statistic =0.67; p =0.7785); normal subjects had non-significant greater nature log of HbA1c than obese subjects (t statistic =0.88; p =0.6513).

A significant linear association with nature log of HbA1c was observed for age (t statistic=5.48; p<0.0001); overall, one-year increase in age results in a 0.0056-unit increase in nature log of HbA1c, after adjusting for BMI category, sex, total cholesterol, mean SBP, and waist (standard error=0.001).

No significant difference in mean nature log of HbA1c was found between males and females (F statistic (1, 359) =0.61; p<0.4351); male subjects have a 0.024 greater nature log of HbA1c than female subjects, after adjusting for BMI category, age, total cholesterol, mean SBP, and waist (standard error=0.031).

A significant linear association with nature log of HbA1c was observed for total cholesterol (t statistic=3.52; p=0.0005); overall, one-mg/dL increase in total cholesterol results in a 0.0012-unit increase in nature log of HbA1c, after adjusting for BMI category, age, sex, mean SBP, and waist (standard error=0.0003).

No significant linear association with nature log of HbA1c was observed for mean SBP (t statistic=0.17; p=0.8645), one-mmHg increase in mean SBP results in a 0.0001-unit increase in nature log of HbA1c, after adjusting for BMI category, age, sex, total cholesterol, and waist (standard error=0.0007).

A significant linear association with nature log of HbA1c was observed for waist (t statistic=3.52; p=0.0005); overall, one-inch increase in waist circumference results in a 0.0135-unit increase in nature log of HbA1c, after adjusting for BMI category, age, sex, total cholesterol, and mean SBP (standard error=0.0038).

Evidence of joint confounding was found in the association between categorical BMI and nature log of HbA1c. Compared to the crude effect estimates, there was a 133.9% change for the effect of overweight BMI compared to normal BMI and a 145.6% change for the effect of obese BMI compared to normal BM. We conclude that the effect of BMI category is jointly confounded by age, sex, total cholesterol, mean SBP, and waist.

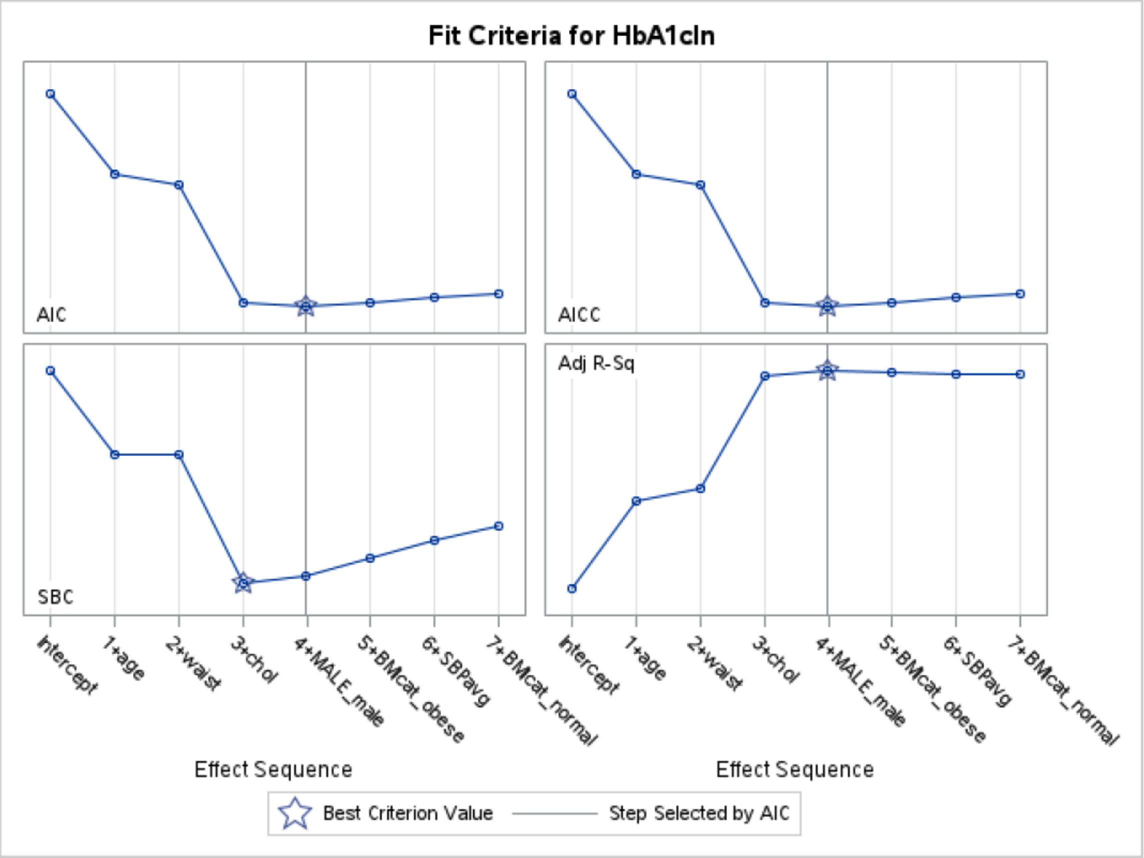
Since predictors age, total cholesterol, and waist are found significantly associated with outcome nature log of HbA1c, after adjusting for other variables in the model (p<0.0001, p=0.0005, p=0.0005, respectively), they are considered covariates of the relationship between nature log of HbA1c and BMI category.

We compare the absolute value of standardized estimates to rank the strength of association among the predictors. Variable age has the greatest absolute value of standardized β=0.297, indicating that age has the strongest linear association with nature log of HbA1c, followed by variable waist (|β|=0.245), total cholesterol (|β|=0.171), obese BMI (|β|=0.074), sex (|β|=0.038), overweight BMI (|β|=0.028), and mean SBP (|β|=0.009), with mean SBP having the weakest linear association with nature log of HbA1c

Model Selection - Goodness of Fit Statistics

LASSO-based linear regression was run to test the goodness of fit and to find the best model for predicting nature log of HbA1c. Variables age, sex, total cholesterol, and waist were selected for the best model based on the lowest AIC of -566.0835 found at step 4. The highest adjusted R-squared of 0.2133 was also found at selection. (Figure 4)

Figure 4. LASSO selection model based on AIC



A different model was selected from backward selection. Variables age, total cholesterol, and waist were selected for the best model. The lowest AIC of -567.32 was found at step 3, after removing BMI category, mean SBP, and sex from the model. The adjusted R-squared for the selected model was 0.2138.

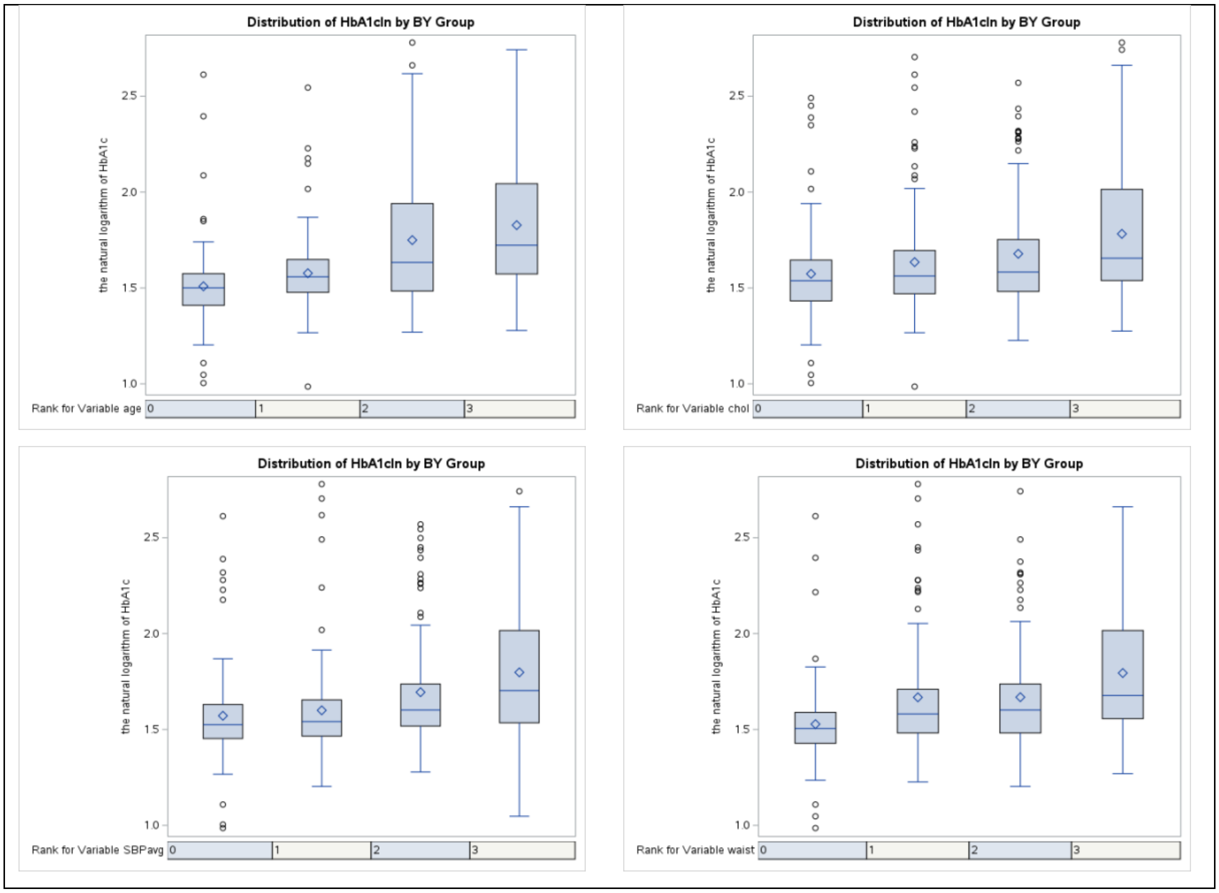
Independent variables selected in the models support the findings we observed in ranking the strength of association among the predictors, as variables age, total cholesterol, and waist have a higher absolute value of standardized estimates and are considered the more important predictors. This means the previous multiple linear regression model was overfitting.

Linear Regression Assumptions

Linear regression assumptions of linearity, homoscedasticity, and normality were checked for continuous variables age, total cholesterol, mean SBP, and waist.

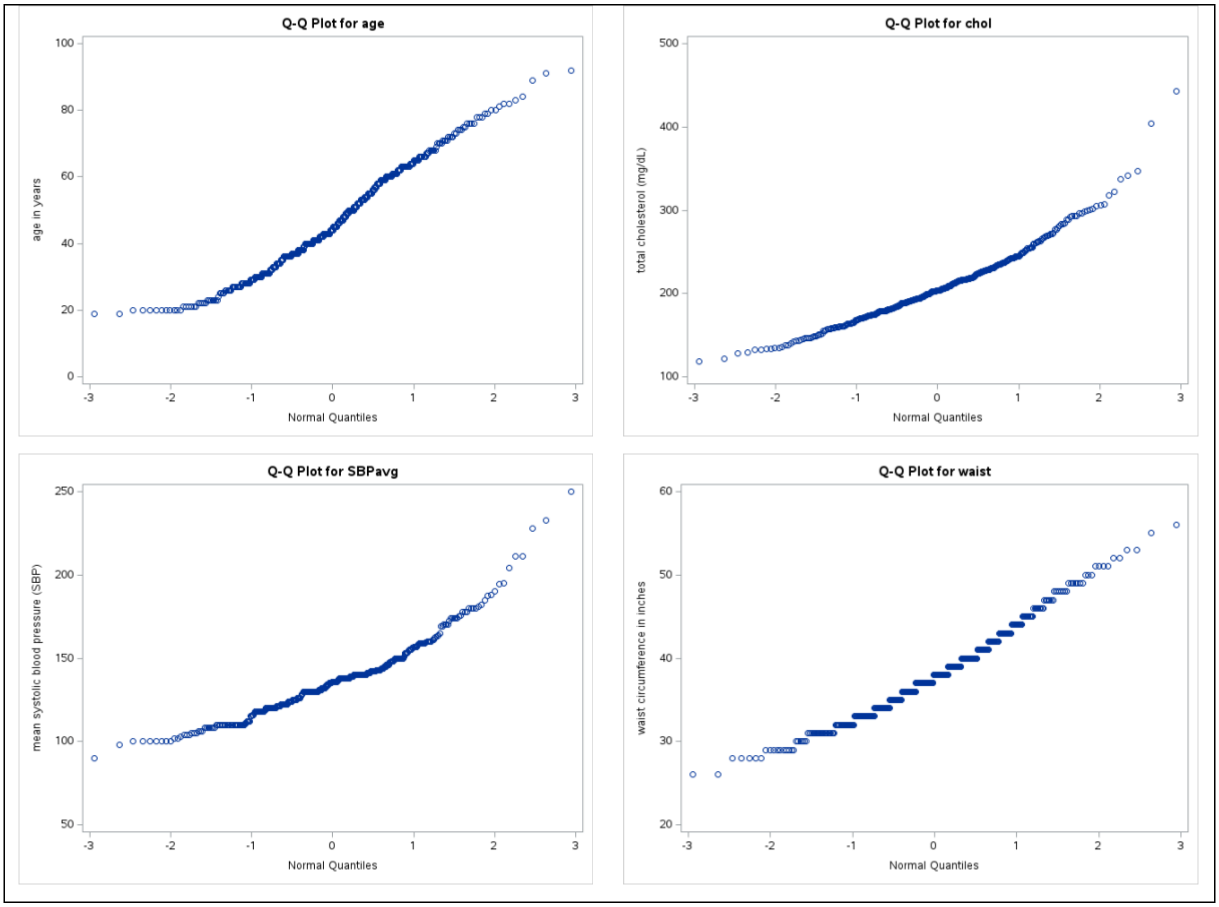
Linearity and homoscedasticity assumptions were checked by looking at the side-by-side boxplots. In the plot there is a trend in both mean and median nature log of HbA1c over each predictor. Therefore, the Linearity assumption was met for all four variables. However, change in variance across the inter-quartile range was observed in all plots. The non-constant variance suggests that the homoscedasticity assumption was not met for all four variables. (Figure 5)

Figure 5. Side-by-side boxplots for age (upper left), total cholesterol (upper right), mean SBP (lower left), and waist (lower right)



Normality assumption was checked by looking at the normal percentile plots (QQ plot). For all four variables the plot showed a straight line, suggesting a reasonable normality. (Figure 6)

Figure 6. Normal Quantile-Quantile Plot for age (upper left), total cholesterol (upper right), mean SBP (lower left), and waist (lower right)



Summary

In this report we evaluated the relationship between HbA1c and BMI. The nature log of HbA1c was found to be a more appropriate outcome than the original HbA1c for the linear association with BMI category. No effect modification was found by age and sex, and waist circumference was picked over hip and waist-hip-ratio. A multiple linear regression model was then fitted for BMI category, age, sex, total cholesterol, mean SBP, and waist all as predictors of nature log of HbA1c. Covariates age, total cholesterol, and waist were found associated with nature log of HbA1c and considered as important predictors. Joint confounding was observed by age, sex, total cholesterol, mean SBP, and waist in the association between categorical BMI and nature log of HbA1c. In conclusion, BMI has a significant linear association with nature log of HbA1c (p=0.0063), and BMI category is also significantly associated with nature log of HbA1c (p=0.0357), when they are a single predictor of nature log of HbA1c. No association between BMI category and nature log of HbA1c was found after adjusting for age, sex, total cholesterol, mean SBP, and waist as other predictors.

Appendix I: SAS code, log, and output (attached in a separate file)