Diabetes Factors Analysis

SDSC2102 STATISTICAL METHOD AND DATA ANALYSIS

Background and problem formulation

What is diabetes

- Diabetes mellitus (DM)
- Group of metabolic disorders characterized by a high blood sugar level over a prolonged period

Goal

- Assess the relationship between diabetic and it factors
- Finding out the which risk factors are more related to having diabetes
- Significance of that factor causing diabetes

Data processing method and Justification

LOGISTIC REGRESSION

- It is a non-linear model to predict binary class
- Use to predict the odds of occurrence
- Assumes that class attributes is linear in the coefficients of the predictive attributes

We use classification model instead of regression model because:

- Having diabetes or not is a binary variable
- 1 means the person has diabetes
- 0 means the person does not has diabetes
- If we use regression model instead of a classification model
  - Accuracy will be relatively lower than classification model

Data selected

- Source: Centers for Disease Control and Prevention of U.S. Department of Health & Human Services
- Duration: 2017-March 2020 Pre-Pandemic
- Survey Type:
  - Questionnaire Data
  - Examination Data

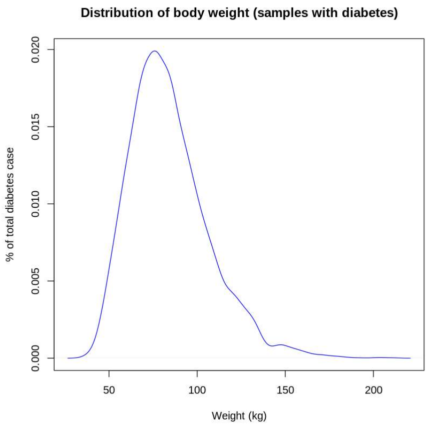
Data set		# of data	Data Code		Data Code	
P_ALQ	2017-March 2020 Pre-Pandemic Alcohol Use	8965	SEQN	Respondent sequence number	BMXBMI	Body Mass Index (kg/m**2)
P_BMX	2017-March 2020 Pre-Pandemic Body Measures	14300	DIQ010	Doctor told you have diabetes	BPQ020	Doctor told you - high blood pressure
P_BPQ	2017-March 2020 Pre-Pandemic Blood Pressure & Cholesterol	10195	BMXWT	Weight (kg)	BPQ080	Doctor told you - high cholesterol level
P_DIQ	2017-March 2020 Pre-Pandemic Diabetes	14986	BMXHT	Standing Height (cm)	ALQ121	Past 12 mo how often drink alcoholic bev
			BMXWAIST	Waist Circumference (cm)	ALQ130	Avg # alcoholic drinks/day - past 12 mos
			BMXHIP	Hip Circumference (cm)		

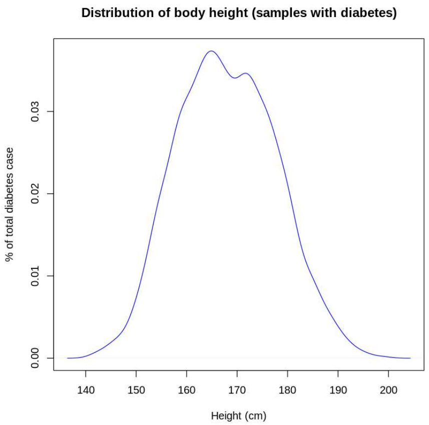
	Number of Instances	Number of Attributes
After merge	5428	12

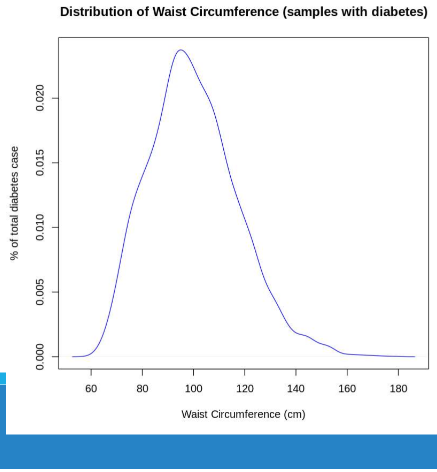
Data analysis procedure

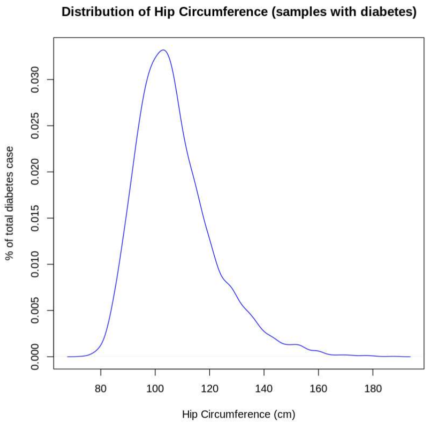
1. Download data from National Health and Nutrition Examination Survey
2. Merge data into data frame
3. Data transformation
    1. Yes-No answer into binary answers
    2. Data normalization
    3. Data filtering with answers such as “Don’t know”, “Refused to answer”, “Missing” etc.
4. Data pre-processing
    1. Calculate BMI base on weight and height

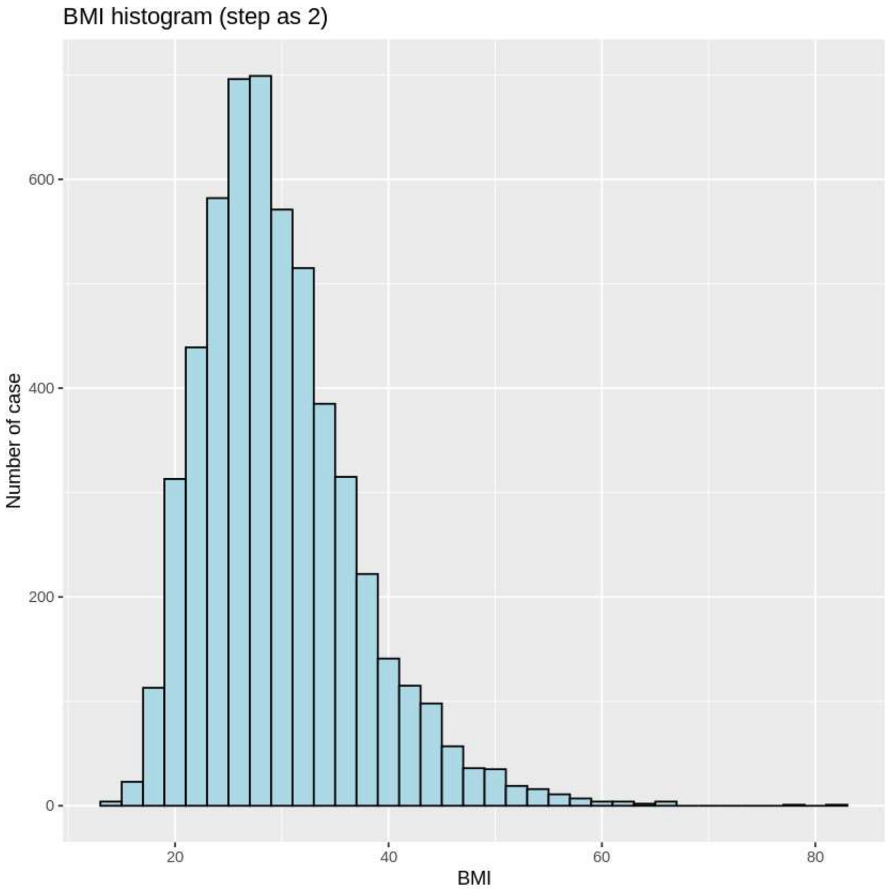
Data overview


Distribution of body weight (samples with diabetes)


Distribution of body height (samples with diabetes)


Distribution of Waist Circumference (samples with diabetes)

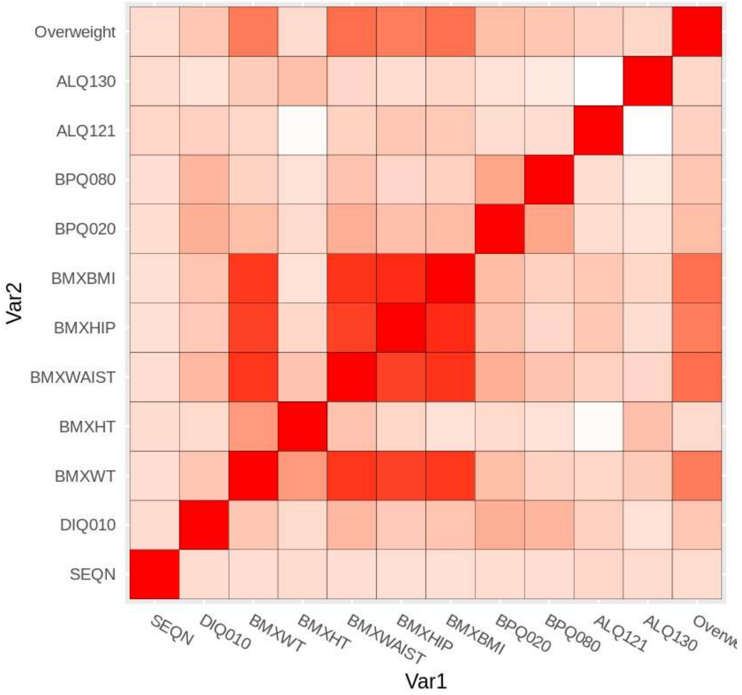

Distribution of Hip Circumference (samples with diabetes)


BMI histogram (step as 2)

Data overview

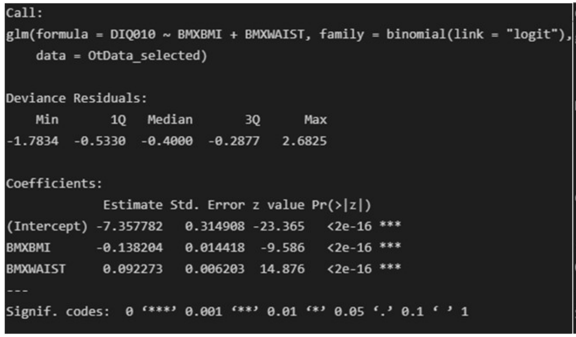
Highly correlated pair:

- BMI – Hip Circumference
- BMI – Waist Circumference
- BMI – Weight
- BMI – Height
- Diabetes – Waist Circumference
- Diabetes – BMI
- Diabetes – high blood pressure
- Diabetes – high cholesterol level

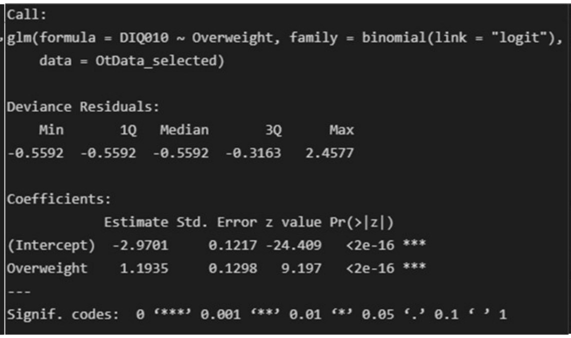


Correlation heatmap

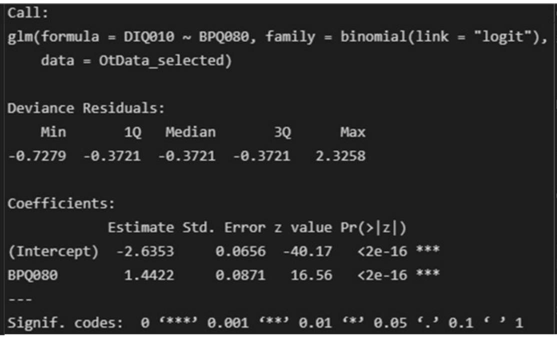
Results - Logistic regression model



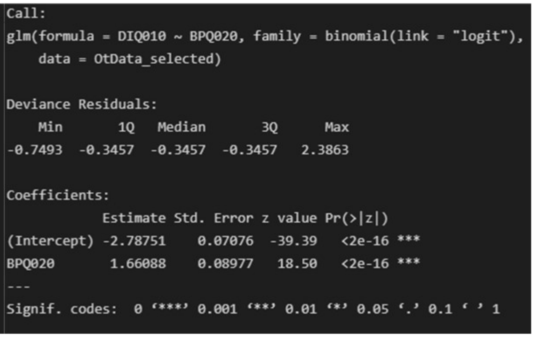
```
Call:
glm(formula = DIQ010 ~ BMXBMI + BMXWAIST, family = binomial(link = "logit"),
    data = OtData_selected)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7834  -0.5330  -0.4000  -0.2877   2.6825

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.357782   0.314908 -23.365   <2e-16 ***
BMXBMI      -0.138204   0.014418  -9.586   <2e-16 ***
BMXWAIST     0.092273   0.006203  14.876   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



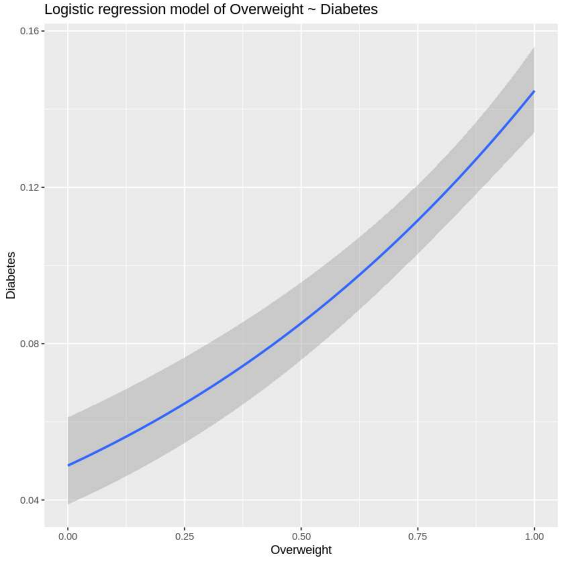
```
Call:
glm(formula = DIQ010 ~ Overweight, family = binomial(link = "logit"),
    data = OtData_selected)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5592  -0.5592  -0.5592  -0.3163   2.4577

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9701     0.1217 -24.409   <2e-16 ***
Overweight    1.1935     0.1298   9.197   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



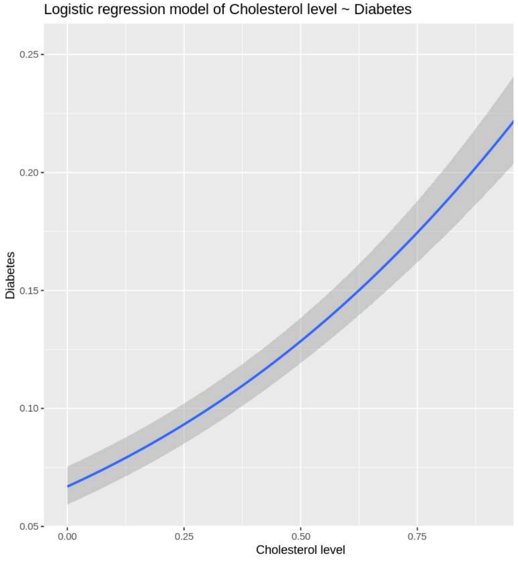
```
Call:
glm(formula = DIQ010 ~ BPQ080, family = binomial(link = "logit"),
    data = OtData_selected)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7279  -0.3721  -0.3721  -0.3721   2.3258

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6353     0.0656  -40.17   <2e-16 ***
BPQ080        1.4422     0.0871   16.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



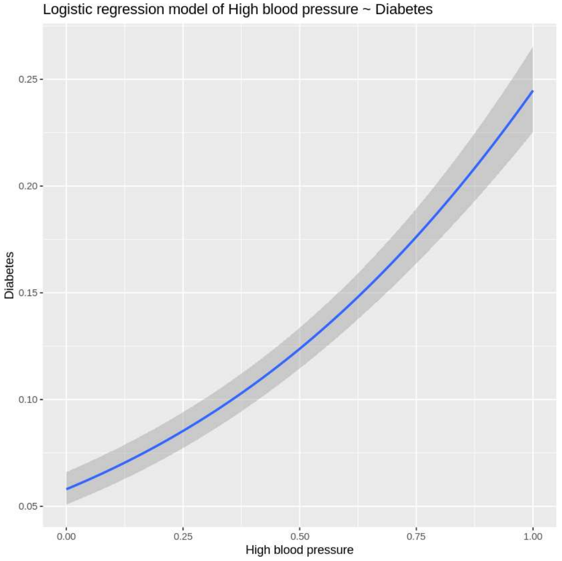
```
Call:
glm(formula = DIQ010 ~ BPQ020, family = binomial(link = "logit"),
    data = OtData_selected)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7493  -0.3457  -0.3457  -0.3457   2.3863

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.78751    0.07076  -39.39   <2e-16 ***
BPQ020       1.66088    0.08977   18.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Logistic regression model of Overweight ~ Diabetes

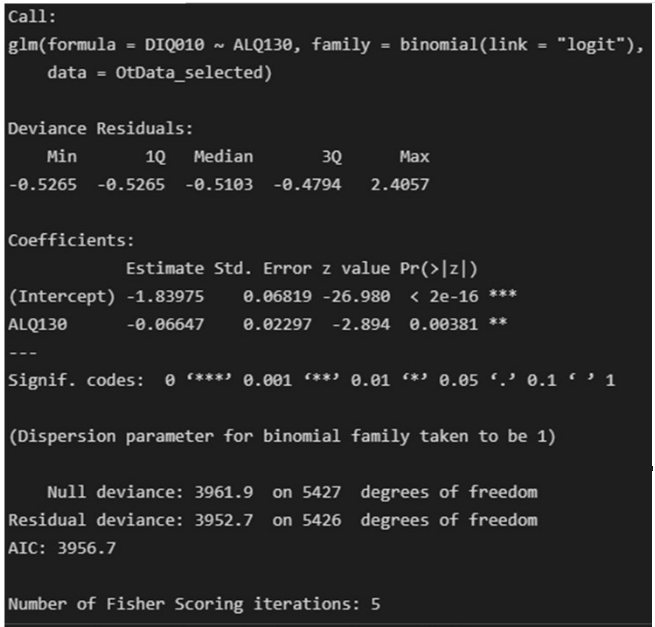


Logistic regression model of Cholesterol level ~ Diabetes

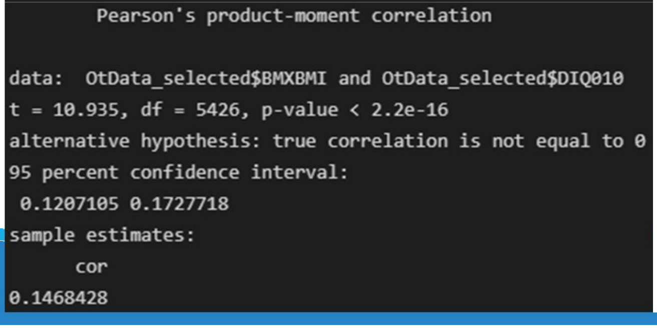


Logistic regression model of High blood pressure ~ Diabetes

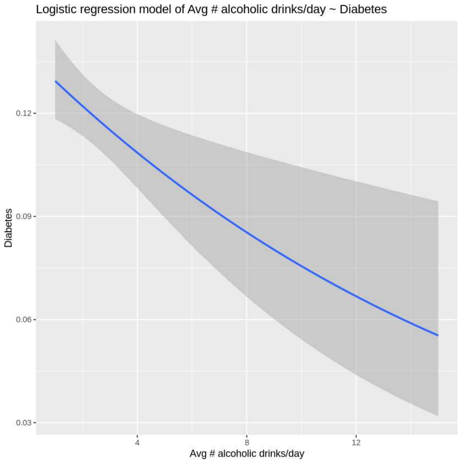
Result – Multiple Logistic regression model

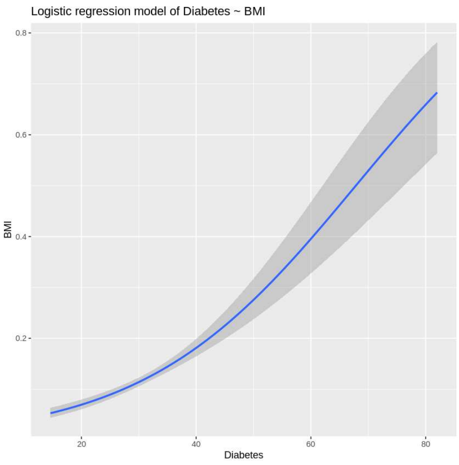


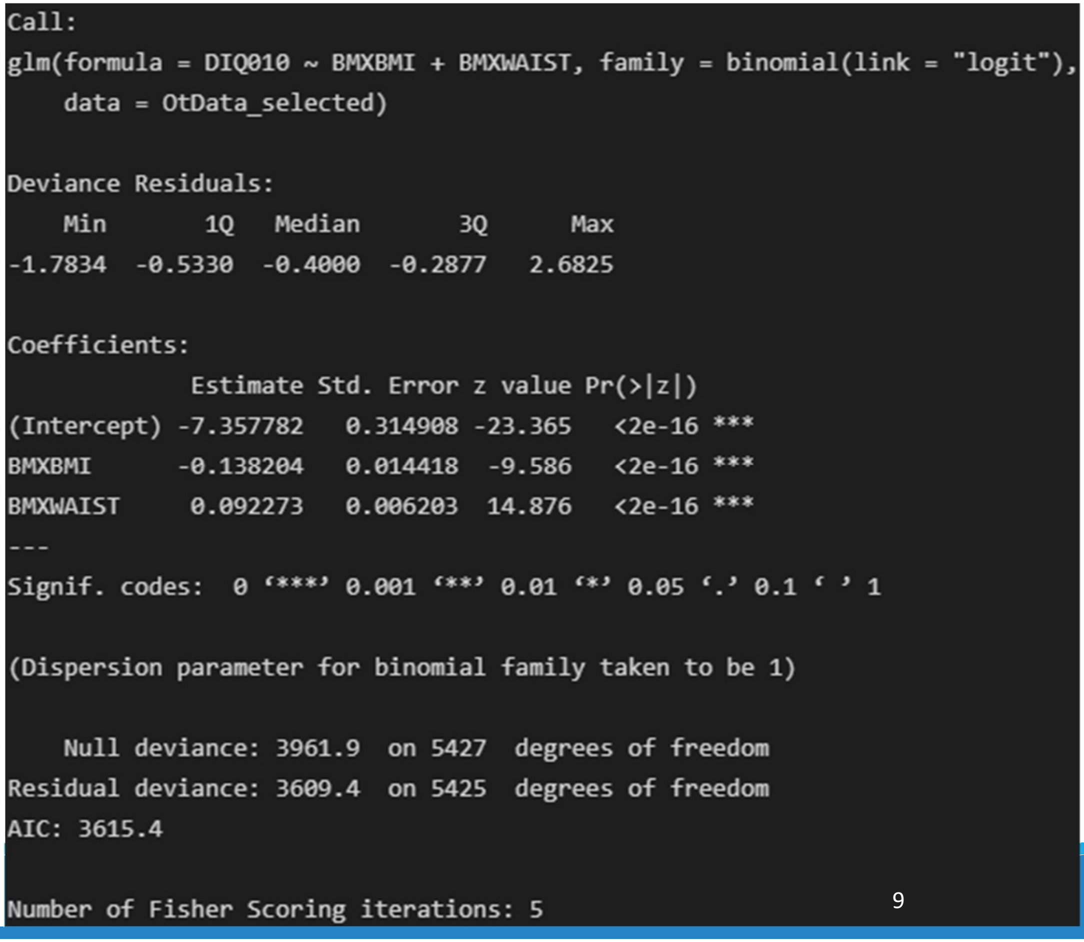
```
Call:
glm(formula = DIQ010 ~ ALQ130, family = binomial(link = "logit"),
    data = OtData_selected)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5265  -0.5265  -0.5103  -0.4794   2.4057

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.83975    0.06819 -26.980  < 2e-16 ***
ALQ130      -0.06647    0.02297  -2.894  0.00381 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3961.9  on 5427  degrees of freedom
Residual deviance: 3952.7  on 5426  degrees of freedom
AIC: 3956.7

Number of Fisher Scoring iterations: 5
```



```
         Pearson's product-moment correlation

data:  OtData_selected$BMXBMI and OtData_selected$DIQ010
t = 10.935, df = 5426, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1207105 0.1727718
sample estimates:
      cor
0.1468428
```

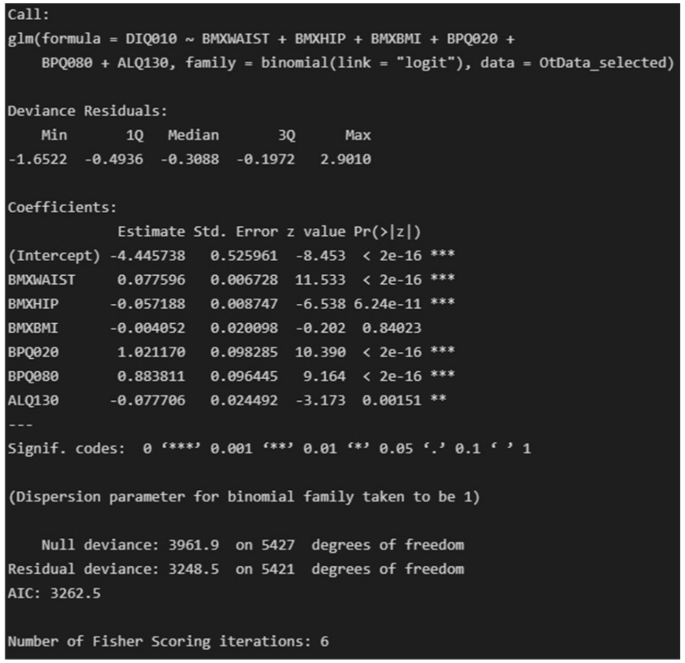






```
Call:
glm(formula = DIQ010 ~ BMXBMI + BMXWAIST, family = binomial(link = "logit"),
    data = OtData_selected)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7834  -0.5330  -0.4000  -0.2877   2.6825

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.357782   0.314908 -23.365   <2e-16 ***
BMXBMI      -0.138204   0.014418  -9.586   <2e-16 ***
BMXWAIST     0.092273   0.006203  14.876   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3961.9  on 5427  degrees of freedom
Residual deviance: 3609.4  on 5425  degrees of freedom
AIC: 3615.4

Number of Fisher Scoring iterations: 5
```

Result – Multiple Logistic regression model

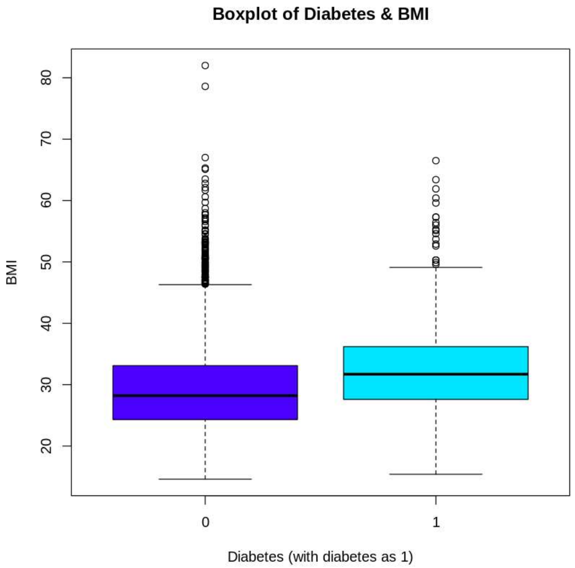


```
Call:
glm(formula = DIQ010 ~ BMXWAIST + BMXHIP + BMXBMI + BPQ020 + 
    BPQ080 + ALQ130, family = binomial(link = "logit"), data = OtData_selected)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.6522  -0.4936  -0.3088  -0.1972   2.9010  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.445738   0.525961  -8.453  < 2e-16 ***
BMXWAIST     0.077596   0.006728  11.533  < 2e-16 ***
BMXHIP      -0.057188   0.008747  -6.538 6.24e-11 ***
BMXBMI      -0.004052   0.020098  -0.202  0.84023    
BPQ020       1.021170   0.098285  10.390  < 2e-16 ***
BPQ080       0.883811   0.096445   9.164  < 2e-16 ***
ALQ130      -0.077706   0.024492  -3.173  0.00151 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3961.9  on 5427  degrees of freedom
Residual deviance: 3248.5  on 5421  degrees of freedom
AIC: 3262.5

Number of Fisher Scoring iterations: 6
```

Diabetes ~ WAIST+HIP+BMI+BPQ020+BPQ080+ALQ130

Boxplot analyzing

DIABETES & BMI



Conclusion and Discussion

- Significant relation with diabetes
  - Waist Circumference
  - Hip Circumference
  - BMI
  - High blood pressure
  - High cholesterol level
  - Heavy alcohol taking
- Having these factors
  - Tended to have higher chance of having diabetes
- Limitation
  - More variables can be used in this project
  - More datasets can be used in this project