1. In a diabetes study, 1123 subjects were recruited, and a number of clinical traits and information were collected, including (see data in the attached file “HW1\_data.txt”):

Sex: male/female Age: age of the study subject

bmi: body mass index fbg: fasting blood glucose

fins: fasting insulin hba1c: hemoglobin A1c

tg: total glyceride tcho: total cholesterol

hdl: high density lipoprotein ldl: low density lipoprotein

Particularly, the investigators are interested in the effects of fbg (X1) and tg (X2) on hba1c (Y). In addition, hba1c > 6.5 is considered to be diabetic. So another question is whether the same model can be used to characterize the relationship between predictors and hba1c for all individuals, or whether two different models are needed: one for people with diabetes, and the other for those without diabetes. Now, you are assigned to analyze the data, and are asked to complete the following tasks.

1. Fit a simple linear regression model between hba1c and fbg for diabetic patients.
2. Obtain the parameter estimations and the corresponding standard errors.

**Linear Model:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Residuals | | | | |
| Min | 1st Quartile | Median | 3rd Quartile | Max |
| -3.8952 | -1.1854 | -0.1871 | 1.1226 | 4.4051 |

Parameter Estimates using Least Squares Method:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | | | | |
|  | Estimate | Standard Error | t-value | Pr(>|t|) |
| Intercept | 5.54929 | 0.31580 | 17.57 | <2e-16 |
| fbg | 0.38403 | 0.03317 | 11.58 | <2e-16 |

Summary:

* Residual standard error: 1.615 on 389 degrees of freedom
* Multiple R-squared: 0.2562
* Adjusted R-squared: 0.2543
* F-statistic: 134 on 1 and 389 DF
* p-value: < 2.2e-16

**Confidence Intervals of Parameter Estimates:**

|  |  |  |
| --- | --- | --- |
|  | 2.5% | 97.5 |
| Intercept | 4.928406 | 6.1701748 |
| fbg | 0.318806 | 0.4492463 |

A graph with lines and numbers

Description automatically generated

**Estimated Model:**

Or

1. Check on the validity of the assumptions including normality and homoscedasticity, using the approaches you think appropriate.

***Note*** that in order to check whether variances are constant, you may need to put fbg values into several intervals, and consider all fbg value in a particular interval being the same. So use a software to bin the data with the following range: <4; (4, 5); (5, 6.5); (6.5, 8); (8, 10); >10. Then set the original X values to the mid-point of the corresponding interval.

For example, we have five data points with X={4.6, 5.4, 6.1, 7, 7.8}. After the process, we will change the X’s to 4.5, 5.75, 5.75, 7.25, and 7.25, as they fall into intervals (4,5), (5, 6.5), (5, 6.5), (6.5, 8), and (6.5, 8).

**Lack of Fit**

Plot to Examine LOF

A comparison of a graph

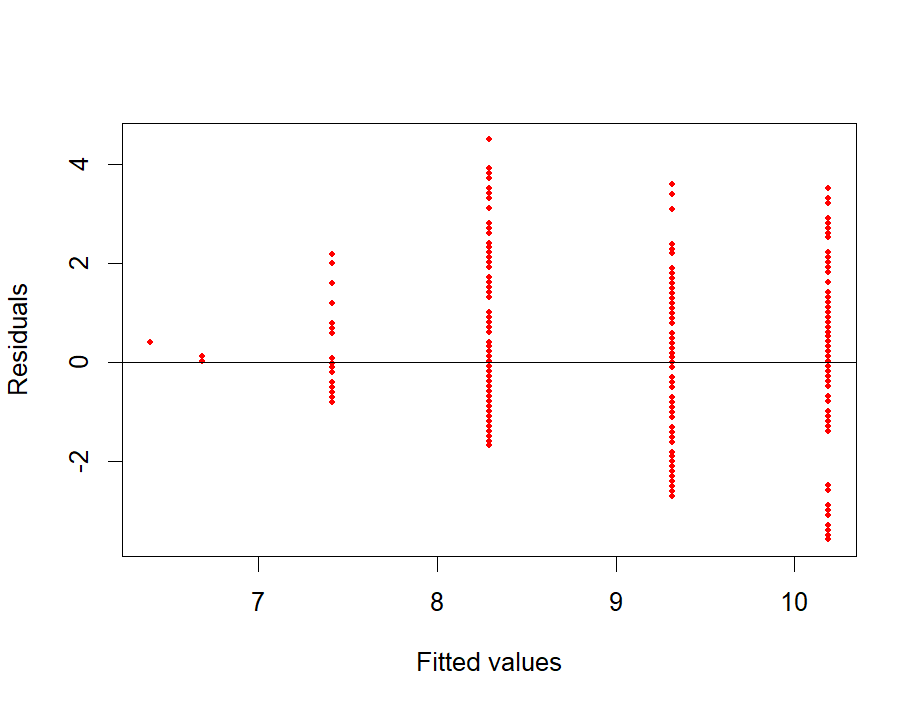
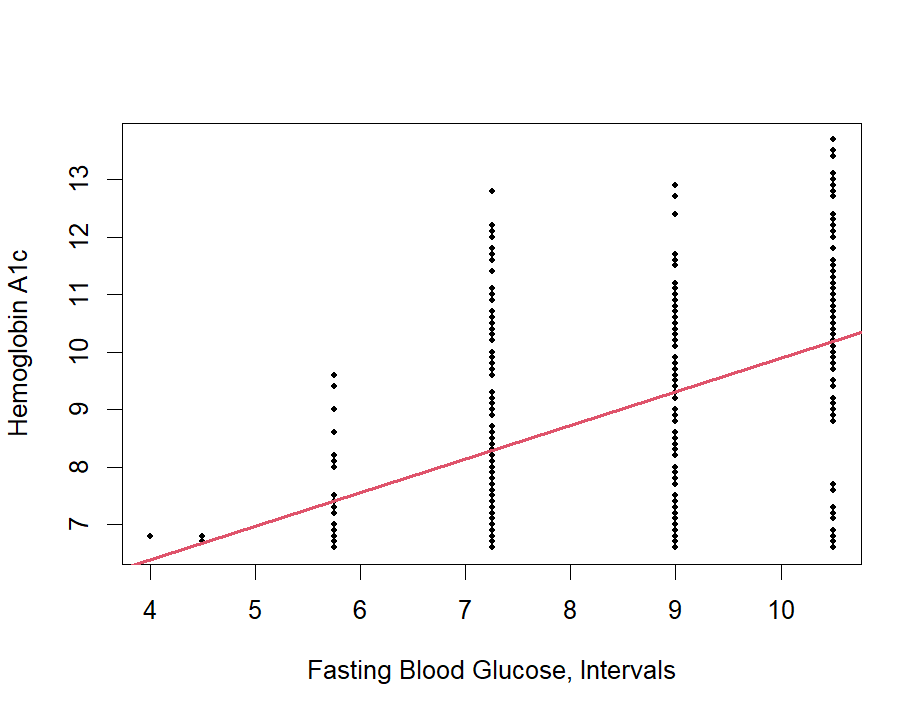
Description automatically generated with medium confidence

* The model plot follows linear trend relatively.
* The residual plot shows a funnel shape and slight negative curvature.
* This model could be missing predictors. May also need to be weighted or undergo a Y transformation.

**Evaluate Pure Errors**

Separate predictor values into intervals to use in place of the predictor variable.

Plots to represent the binned data:

****

Conduct an ANOVA of two models to evaluate pure errors:

Model 1:

Model 2:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA Table | | | | | | |
| Model | Res df | RSS | df | SS | F | Pr(>F) |
| 1 | 389 | 1038.3 |  |  |  |  |
| 2 | 389 | 1021.8 | 4 | 16.52 | 1.5561 | 0.1854 |

The ANOVA of the regular model and the model with the predictor as a factor is not statistically significant with a p-value of 0.1854. There appears to be no reason to doubt the adequacy of the model. There seems to be no significant difference among the groups’ pure errors.

* 1. Homoscedasticity of Errors

Examine Residual Plot

A comparison of a graph

Description automatically generated with medium confidence

Unsatisfactory residual plot

* Funnel-shaped
* Slight negative curvature

This model could be missing predictors. The model could also need to be weighted or undergo a Y transformation.

Levene’s Test for Homogeneity of Variance, Center=Median

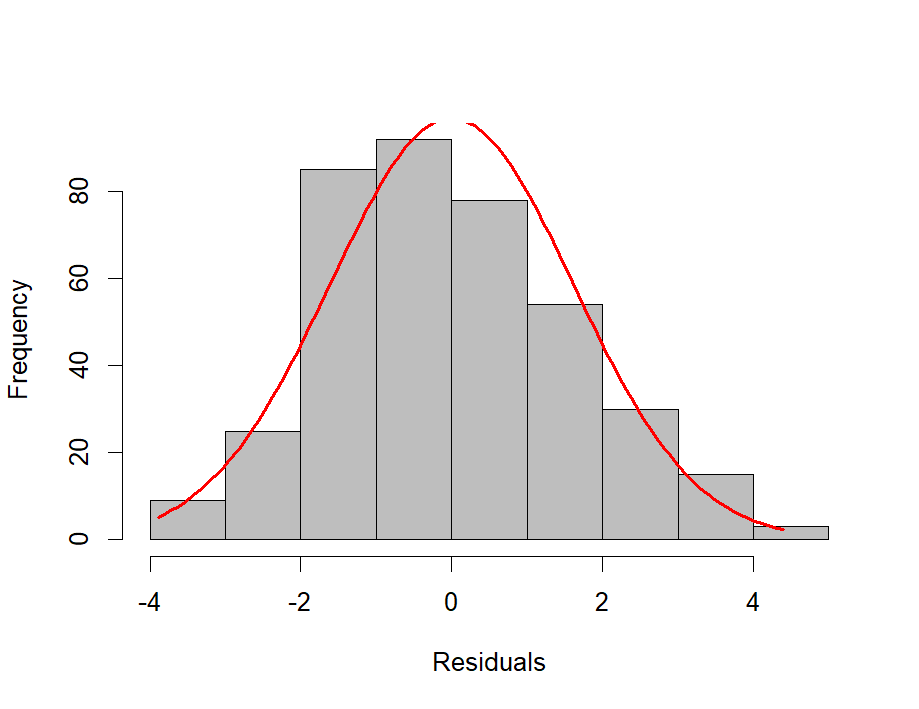
Using fbg intervals:

|  |  |  |  |
| --- | --- | --- | --- |
| Levene’s Test | | | |
|  | df | F-value | p-value |
| Group | 5 | 3.6507 | 0.003071 |
|  | 385 |  |  |

The F-value of 3.6507 is statistically significant at a p-level of 0.003071. This suggests that there are significant differences between the variances (heteroscedasticity) of the various groups. The groups were separated by fbg interval.

* 1. Normality of Errors

A graph and diagram of a graph

Description automatically generated

A graph and diagram of a graph

Description automatically generated

* The residual histogram shows slight asymmetry which may suggest that the errors are not normal.
* The QQ normal plot shows heavier tails. This could suggest that the errors are not normal and may more closely follow the t-distribution instead.

**Shapiro-Wilk Normality Test**

Using residual data:

|  |  |
| --- | --- |
| W-value | p-value |
| 0.98703 | 0.001459 |

The Shapiro-Wilk test for normality is statistically significant with a p-value of 0.001459. This suggests that the model errors significantly deviate from the normal distribution.

1. Complete the ANOVA table and determine whether the F-statistic will result in a significant result. Interpret your result.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA Table for Overall Significance | | | | | |
| Source | df | SS | MS | F-value | Pr(>F) |
| Regression | 1 | 349.4033 | 349.4033 | 134.0171 | <2.2e-16 |
| Residual | 389 | 1014.183 | 2.593818 |  |  |
| Total | 390 | 1863.11 | 1.66 |  |  |

R-squared: 0.2562385

Adjusted R-squared: 0.2524047

The F-value of 134.0171 for the model is statistically significant at a p-value of approximately 0. This shows that the overall model appears to be significant for predicting hba1c.

1. Fit a simple linear regression model between hba1c and fbg for all subjects
2. Obtain the parameter estimations.

**Linear Model:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Residuals | | | | |
| Min | 1st Quartile | Median | 3rd Quartile | Max |
| -4.3542 | -0.6610 | -0.1731 | 0.4275 | 5.6008 |

Parameter Estimates using Least Squares Method:

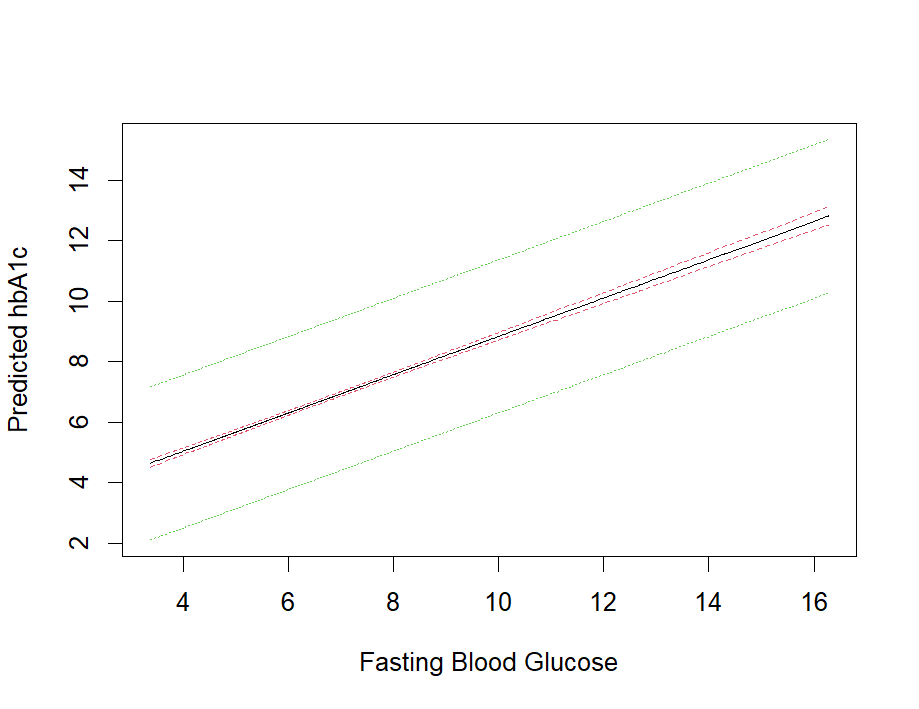
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | | | | |
|  | Estimate | Standard Error | t-value | Pr(>|t|) |
| Intercept | 2.49764 | 0.11143 | 22.41 | <2e-16 |
| fbg | 0.63449 | 0.01547 | 41.02 | <2e-16 |

Summary:

* Residual standard error: 1.289 on 1121 degrees of freedom
* Multiple R-squared: 0.6001
* Adjusted R-squared: 0.5998
* F-statistic: 1682 on 1 and 1121 DF
* p-value: < 2.2e-16

**Confidence Intervals of Parameter Estimates:**

|  |  |  |
| --- | --- | --- |
|  | 2.5% | 97.5 |
| Intercept | 2.2789992 | 2.7162775 |
| fbg | 0.6041396 | 0.6648427 |

****

**Estimated Model:**

Or

1. Check on the lack-of-fit. Again, we need to “create” pure errors before we could conduct the test. Use similar process described in A.2 above, and use intervals <4; (4, 4.5); (4.5, 5); …; (9, 9.5); (9.5, 10); >10.

**Plot to examine LOF**

A graph of a red dot

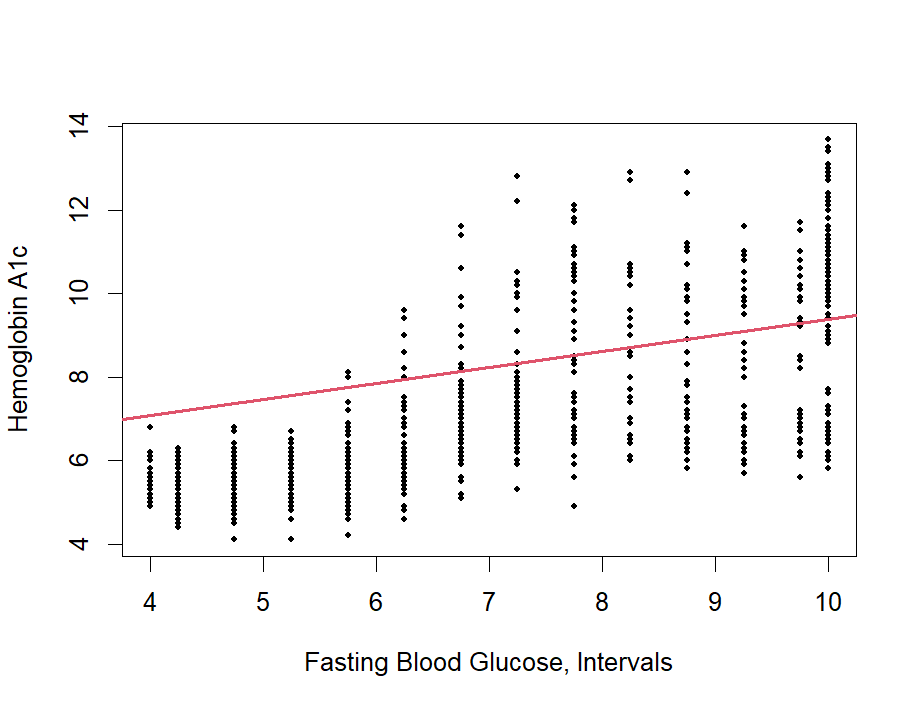
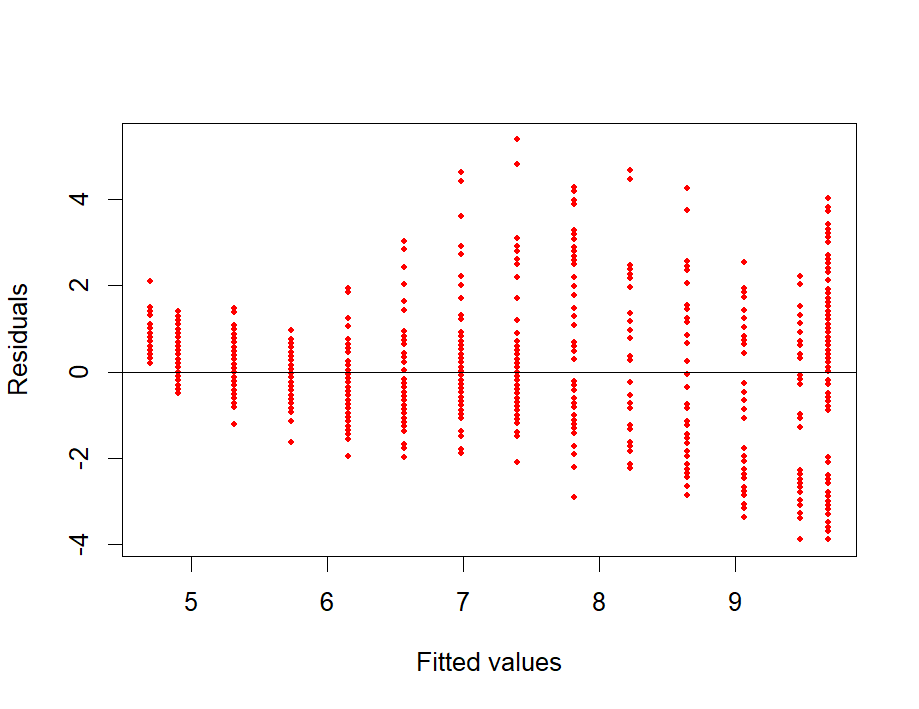
Description automatically generated with medium confidence

* The model plot follows linear trend relatively, but slightly curved positively.
* The residual plot shows a funnel shape and visible negative curvature.
* This model could be missing predictors. May also need to be weighted or undergo a Y transformation. The linear model may not be suitable.

**Evaluate Pure Errors:**

Separate predictor values into intervals to use in place of the predictor variable.

Plots to represent the binned data:

****

Conduct an ANOVA of two models:

Model 1:

Model 2:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA Table | | | | | | |
| Model | Res df | RSS | df | SS | F | Pr(>F) |
| 1 | 1121 | 1867.2 |  |  |  |  |
| 2 | 1109 | 1660.2 | 12 | 207.03 | 11.525 | <2.2e-16 |

The ANOVA of the regular model and the model with the predictor as a factor is not statistically significant with a p-value of approximately 0. There appears to be reason to doubt the adequacy of the model.

1. Do you think this model similar to the model you obtain in A)? (No formal test needed here. We will study how to conduct a formal statistic test later.)

No, this model appears to differ from the model in part A. The plots and calculated values are different and signify some contradictory analysis. Part B’s plot shows more lack of fit with increased curvature in the model plot and the residual plot. Plot B also showed a significant lack of fit when using an ANOVA model to examine pure errors.

1. Fit a multiple linear regression model between hba1c and fbg + tg.
2. Obtain the parameter estimations and complete the ANOVA table.

**Linear Model:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Residuals | | | | |
| Min | 1st Quartile | Median | 3rd Quartile | Max |
| -4.3483 | -0.6595 | -0.1752 | 0.4311 | 5.6142 |

Parameter Estimates using Least Squares Method:

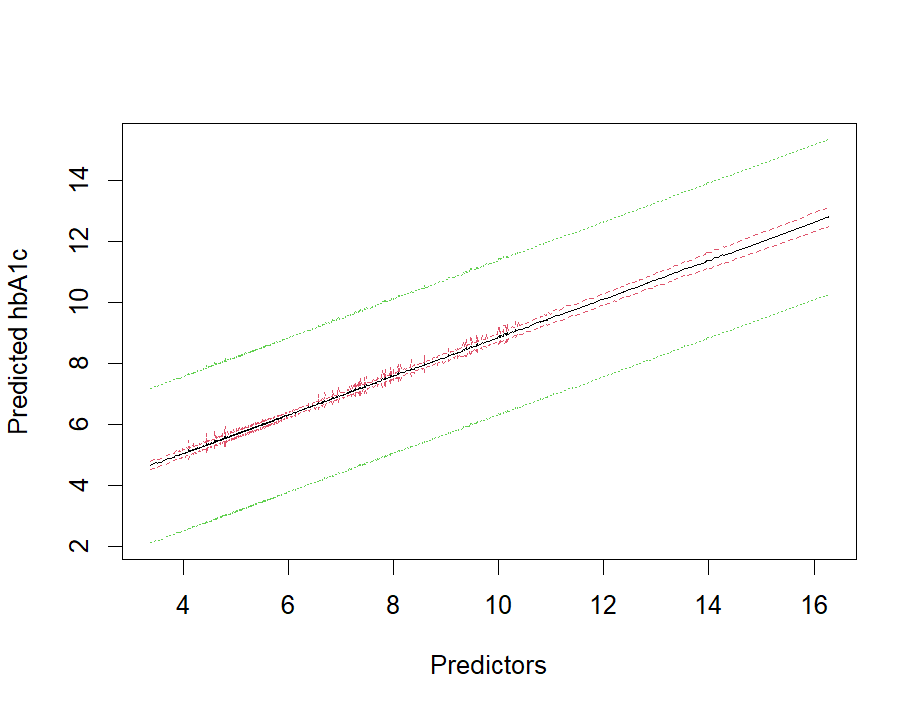
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | | | | |
|  | Estimate | Standard Error | t-value | Pr(>|t|) |
| Intercept | 2.47787 | 0.12337 | 20.085 | <2e-16 |
| fbg | 0.63302 | 0.01597 | 39.650 | <2e-16 |
| tg | 0.01685 | 0.04506 | 0.374 | 0.708 |

Summary:

* Residual standard error: 1.29 on 1120 degrees of freedom
* Multiple R-squared: 0.6002
* Adjusted R-squared: 0.5995
* F-statistic: 840.6 on 2 and 1120 DF
* p-value: < 2.2e-16

**Confidence Intervals of Parameter Estimates:**

|  |  |  |
| --- | --- | --- |
|  | 2.5% | 97.5 |
| Intercept | 2.23580216 | 2.7199341 |
| fbg | 0.60169716 | 0.6643479 |
| tg | -0.07155407 | 0.1052596 |



**Estimated Model:**

Or

**ANOVA for predictors:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA Table | | | | | |
|  | df | SS | MS | F-value | Pr(>F) |
| fbg | 1 | 2796.47 | 2796.47 | 1681.0876 | <2.2e-16 |
| tg | 1 | 0.23 | 0.23 | 0.1399 | 0.7085 |
| Residuals | 1120 | 1863.11 | 1.66 |  |  |

The F-value of 134.02 for fbg is significant at a p-value of approximately 0. This shows that fbg appears to be a statistically significant predictor for hbA1c. The F-vale of 0.1399 for tg is not statistically significant at a p-value of 0.7085. This signifies that tg is not a significant predictor for hba1c.

1. Conduct a test for the whole model and interpret your result.

**ANOVA for overall model:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA Table for Overall Significance | | | | | |
| Source | df | SS | MS | F-value | Pr(>F) |
| Regression | 1 | 2796.706 | 2796.706 | 1682.729 | <2.2e-16 |
| Residual | 1121 | 1863.109 | 1.662007 |  |  |
| Total | 1122 | 1363.586 |  |  |  |

R-squared: 0.6001753

Adjusted R-squared: 0.5994613

The F-value of 134.02 for the regression model is statistically significant at a p-value of approximately 0. This shows that the model significantly predicts hba1c.

1. Manually calculate the corresponding and adjusted based on the quantities in the ANOVA table. You need to write out the formula and include at least one or two intermediate steps showing the values of the quantities.