In a diabetes study, several traits and information were collected, including (see data in the attached file “diabetes\_study.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

The investigators are interested in the effects of different factors on hba1c. You, as a biostatistician, is helping in the data analysis. Based on prior knowledge, the focus will be put on fbg, tg, fins, and bmi. For the following issues, please briefly summarize and report what you have done and what you have found out. Make sure that your report is not just a simple aggregate of software outputs!

1. Using the data in their original scales, please find an appropriate regression model. The regression assumptions should be evaluated and considered by the fitted model if necessary.
2. Conduct appropriate data transformation and find a corresponding fitted model.
3. Assess which of the two models established in a and b is more appropriate for the data. Provide a brief justification for your conclusion.
4. Data description

In the complete data set of 1,845 individuals, HBA1C and other 4 factors of focus do not follow the normal distribution (Figure 1). The Shapiro-Wilk normality test p-value for HBA1C and other 4 factors are all < 2.2e-16.



**Figure 1. Histogram of FBG, BMI, TG and HBA1C**

1. I use scatter plot with a linear regression fitted line to show the relationship between HBA1C and the 4 factors of focus (Figure 2). From Figure 3, FBG, TG and FINS are positively correlated with HBA1C, while BMI is possibly negatively correlated with HBA1C. The effect magnitude of FBG is likely greater than other 3 factors.

Table 1 summarizes the univariate analysis for the relations between the 4 factors of focus and HBA1C. Except BMI is not significant at level of 0.05, FBG, TG and FINS are significant correlated with HBA1C. FBG explains ~46.4% variance of the HBA1C, which is greater than other 3 factors.

The forward model selection based on AIC is used to find the best model. The 4 factors of focus and other covariates are all considered. Their interactions are not included in the analysis. The final model consists of FBG, BMI, HDL, TCHO and Age. The overall model is statistically significant (F-test p-value < 2.2e-16). It explains ~47.6% (adjusted R squared) variance of the HBA1C. The summary of the model fit is shown in Table 2. Based on the selected model, FBG and BMI are significantly associated with HBA1C. Their effect sizes are 0.508 and -0.041 respectively. TG and FINS are not significantly associated with HBA1C.



**Figure 2. Scatter plot of FBG, BMI, TG and FINS vs. HBA1C with linear regression line**

**Table 1. Summary for the univariate analysis**

|  |  |  |
| --- | --- | --- |
|  | p-value | Adjusted R-squared |
| FBG | <2.2e-16 | 0.4646 |
| TG | 2.87e-13 | 0.0285 |
| BMI | 0.0917 | 0.0015 |
| FINS | 1.84e-08 | 0.0170 |

**Table 2. Summary for the multiple linear regression coefficients**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | p-value |
| (Intercept) | 3.837 | 0.330 | 11.61 | <2.00E-16 |
| fbg | 0.508 | 0.014 | 37.301 | <2.00E-16 |
| hdl | -0.296 | 0.071 | -4.193 | 2.88E-05 |
| bmi | -0.041 | 0.010 | -3.926 | 8.96E-05 |
| tcho | 0.076 | 0.032 | 2.39 | 0.0169 |
| Age | 0.006 | 0.003 | 2.29 | 0.0221 |



**Figure 3. Diagnostic plot of the forward selected model: a) residuals vs. fitted value; b) normal quantile plot of the standardized residuals; c) standardized residuals vs. fitted value; d) Cook’s distance of the observations; e) Component residual plot for FBG; f) Component residual plot for BMI**

A set of plots is used to check the model assumption. In Figure 3a, the variance of the residuals is increased with the increase of the fitted value. The assumption of the homogeneity variance is not satisfied. Figure 3c provides another evidence for the violation of homogeneity variance, in which the square root of the standardized residual is correlated with the fitted values. Figure 4e and 3f are the component residual plot for FBG and BMI to check their linear relationship with HBA1C. The fitted line closely lies along the expected lines (dash line). The linear relationship is not violated. Figure 3b is the normal Q-Q plot of the standardized residuals. Under the assumption of normally distributed residuals, the points in the Figure 3b should lie along the dash line. The deviation of the fitted points from the expected line indicates the assumption of the normal distributed residuals is not satisfied.

1. Using the Box-Cox transformations for linear models, HBA1C is transformed by the Box-Cox transformation with the power of -1.75, which is found by the profile log-likelihoods (using R function boxCox()). After the transformation, the forward model selection by AIC produces a new regression model with all variables included. The overall model is statistically significant with F-test p-value < 2.2e-16. It explains ~49.4% (adjusted R squares) variance of the HBA1C.

The regression coefficients are summarized in Table 4. After adjusting the other covariates, FBG, TG, HDL, AGE, SEX, and LDL are significantly associated with HBA1C at significance level of 0.05.

**Table 4. Summary for the multiple linear regression coefficients for transformed data**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Estimate | Std. Error | p-value |
| (Intercept) | 0.53091 | 0.00137 | <2e-16 |
| fbg | 0.00205 | 0.00006 | 9.02E-206 |
| tg | 0.00057 | 0.00017 | 6.67E-04 |
| hdl | -0.00092 | 0.00031 | 2.69E-03 |
| Age | 0.00004 | 0.00001 | 1.28E-04 |
| sexmale | 0.00073 | 0.00029 | 1.25E-02 |
| ldl | 0.00063 | 0.00022 | 3.83E-03 |
| tcho | -0.00037 | 0.00019 | 5.28E-02 |
| bmi | -0.00008 | 0.00004 | 6.78E-02 |

1. Model fitted in b is more appropriate than that in a.

The adjusted R squared for the model fitted in c is ~49.4%. It is greater than the adjusted R square for the model fitted in b, which is ~47.6%. The model in c explains more variance than model in b does.

Moreover, the diagnostic analysis in a reveals the normality and homogeneity assumption are violated. After the data transformation, the model in c results residuals much closer to be normally distributed (Figure 4b). The change of the residuals variance with the fitted value is less compared with model in a (Figure 4a and 4c). The model in c also reduces the Cook’s distance of some influential observations, such as the 33rd, 232nd observations and so on. Consequently, the model in c is more appropriate than model in b in terms of variance explained and satisfying model assumption.



**Figure 4. Diagnostic plot of the data transformed model: a) residuals vs. fitted value; b) normal quantile plot of the standardized residuals; c) standardized residuals vs. fitted value; d) Cook’s distance of the observations**