**BIOS 663: Diabetes Investigation Final Report**

Group 9: Joyce Choe, Joyce Liu, Meng Pan, Anthony Wang

**Introduction**

Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar) due to insufficient insulin, which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves (CDC, 2023). Early detection of diabetes progression is essential for secondary prevention. In the present study, we aim to determine the predictors that are most important for diabetes progression and associated with another.

**Data**

We obtained data from a diabetes study introduced in the 2004 article *Least Angle Regression* (Efron et al., 2004).The data set consists of 10 predictor variables at baseline and one response variable for 442 diabetes patients. There are no missing entries in the dataset. We provide the summary statistics of the variables in our dataset as follows:

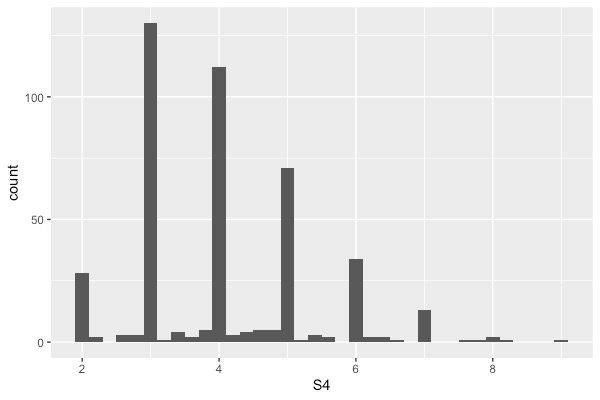
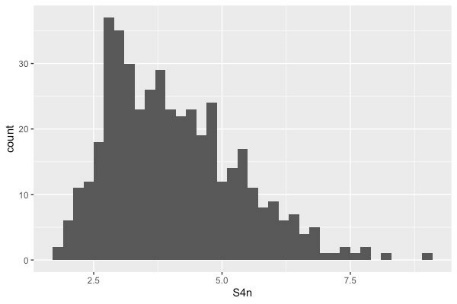
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | | **Description** | **Mean (SD)** | **Range** |
| **AGE** | | Age in years | 48.5 (13.11) | (19, 79) |
| **SEX** | | Categorical, binary, de-identified, value = {1, 2} | 1: 235 (53%)  2: 207 (47%) | |
| **BMI** | | Body Mass Index | 26.38 (4.42) | (18, 42.2) |
| **BP** | | Average blood pressure (mmHg) | 94.65 (13.83) | (62, 133) |
| Blood Serum  Measurements  S1-S6 | **S1**. TC | Total serum cholesterol (mg/dL) | 189.14 (34.61) | (97, 301) |
| **S2**. LDL | Low-density lipoproteins (LDL) (mg/dL) | 115.44 (30.41) | (41.6, 242.4) |
| **S3**. HDL | High-density lipoproteins (HDL) (mg/dL) | 49.79 (12.93) | (22, 99) |
| **S4**. TCH | Total cholesterol/HDL ratio | 4.05 (1.26) | (2, 9.1) |
| **S5**. LTG | Log of serum triglycerides (log mg/dL) | 4.64 (0.52) | (3.3, 6.1) |
| **S6**. GLU | Blood glucose level (mg/dL) | 91.26 (11.50) | (58, 124) |
| **Y** (response variable) | | Quantitative measure of disease progression one year after baseline | 152.13 (77.10) | (25, 346) |

One caveat in our predictor variables is a potential measurement error associated with variable S4. In the dataset, most observations have integer values for S4, but the rest have fraction values for S4, as seen in the histogram of S4 on the left. We suspect this is likely a measurement error caused by the differing preference for rounding method for each health practitioners while recording the S4 value. To confirm our suspicion, we constructed a new S4 variable, namely S4n, based on the description of S4—ratio of total cholesterol over HDL—which corresponds to S1 divided S3 in our dataset. The resulting values of S4n are all fractions, which confirms our suspicion. As seen in the histogram of S4n, the distribution of S4n is smoother than S4 but roughly maintains the same shape. This is because the observations with integer values for S4 have spread out between integers for S4n. We decided to proceed with modeling using S4n in place of S4.

Additionally, we compared Pearson’s correlation coefficients between all potential predictors before fitting the model. In particular, S1 and S2 had a very high positive correlation (r = 0.897) and S3 and S4 had a high negative correlation (r = - 0.738). Therefore, we made sure to check for multicollinearity as we fit the models.

S4 values

S4n values



**Methodology**

Before fitting models, we mean-centered continuous predictor variables by group. This helped us interpret our model estimates for prediction. Also, instead of x=0, the intercept is a more realistic and meaningful value when at the predictor mean (Newsom, 2024). We chose our final model based on the following six steps:

1. Multiple linear regression: full model

For the full model, we used quantitative measure of diabetes progression one year after baseline (Y) as the response variable and all predictor variables (age, BMI, BP, S1-S6, and sex) as independent variables and fitted a linear model (adj. R2 = 0.5065, RMSE = 54.16).

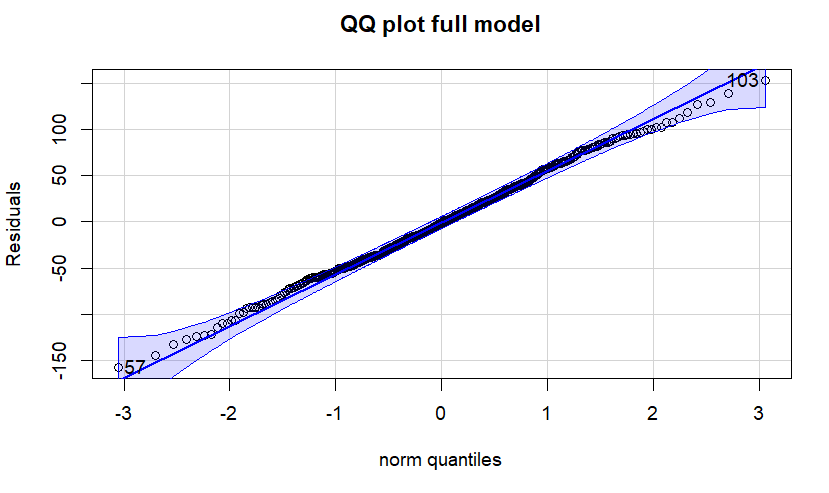
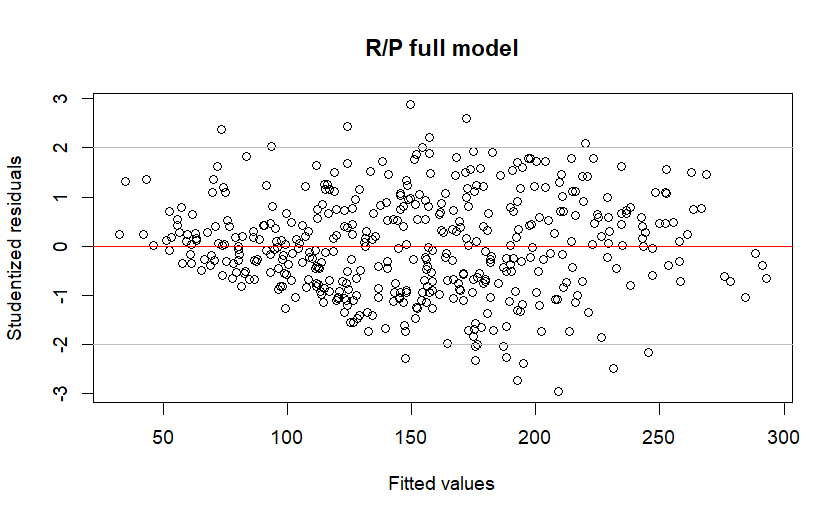
1. Diagnosis: multicollinearity and slight heterogeneity of error variance

The full model had the following variance inflation factors (VIFs) for the predictors:

A number on a white background

Description automatically generated

We noticed S1 and S2 had especially high VIFs (> 30) and S3, S4, and S5 exhibited VIFs greater than 10. This indicated a high level of multicollinearity. Hence, S1 and S2 variables were excluded from further modeling, and we would check VIFs for S3, S4, and S5 after fitting newer models.



The full model residual plot shows minor violations of constant error variance because the errors are not evenly distributed about fitted values (50 to 150). However, the residual plot shows linearity and the QQ plot shows Gaussian errors.

1. Transformation

Since we wanted to improve error variance by minimizing sum of squared errors (SSE), we applied a Box-Cox transformation on Y and found the ideal transformation for Y was √Y (λ = 0.5). Additionally, the power transform test confirmed that we should reject the null hypotheses, λ = 0 or λ = 1, and transform our response variable. The sqrt transformation is known to stabilize the error variance as the variance is proportional to the mean. However, interpreting coefficients may be difficult after transformation.

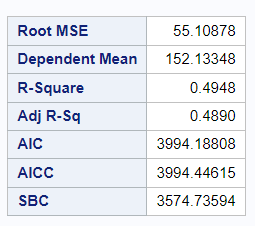
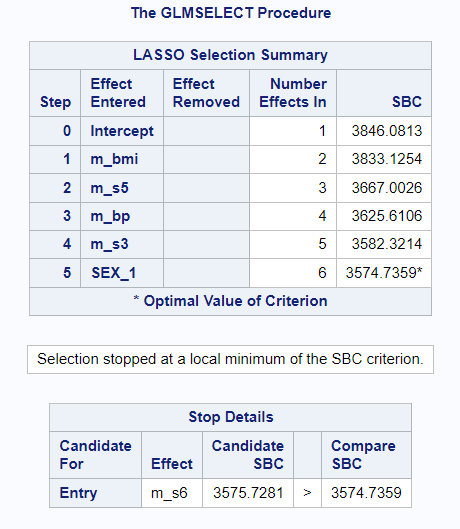
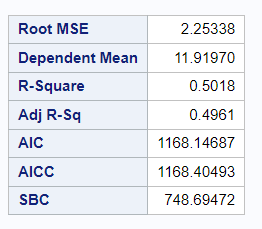
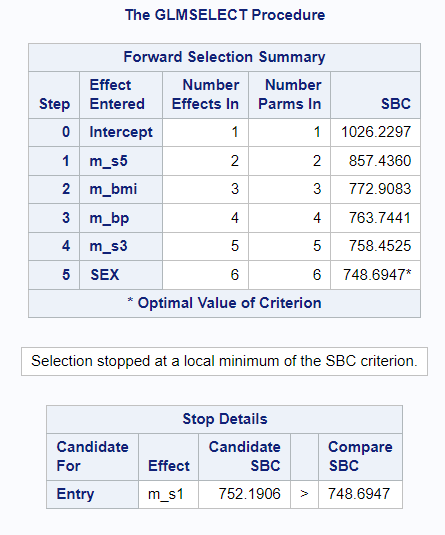
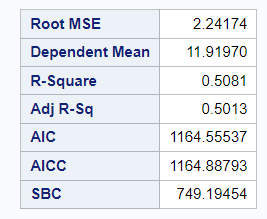
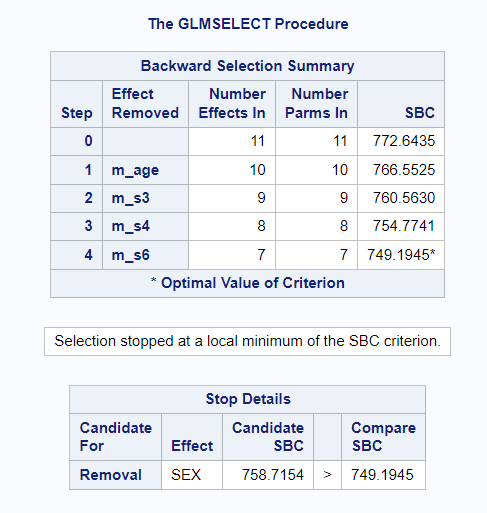
A graph of a function

Description automatically generatedA black text on a white background

Description automatically generated

1. Model selection

Our goal was to narrow down the most useful predictors for diabetes progression. As a result, we decided to select a model with few predictors. We used Schwarz Criterion (SBC) or Bayesian Information Criterion (BIC) as our model criteria measure because this statistic is determined by the number of predictors in a model and the model’s SSE. SBC increases with larger models and increases as SSE increases, so we looked for the lowest SBC when choosing a model. We ran backwards selection, forward selection, and LASSO. The output for each model is shown below:



LASSO selection in SAS did not accept √Y since LASSO in SAS standardizes the values for us, so Y was added as the response variable. This explains a higher SBC in LASSO than in backward or forward selection, even though LASSO is recommended for multicollinear data because it attempts to minimize the MSE. Interestingly, forward and LASSO selected the same important predictors (BMI, sex, BP, S3, and S5). We chose forward selection for our final model because it had the lowest SBC. In addition, forward selection begins with the smallest possible model (intercept-only model) and builds on that using added-in-order tests to test a predictor’s significance above and beyond preceding predictors already in the model. Using forward selection, we discovered a unique set of predictors that were important to contributing to changes in the response variable.

1. Interaction

We tested all potential interactions with five predictors in our model, but all of them had relatively low significance (Appendix Table 1). Since interaction effects were not as important as main effects, we excluded interaction terms to avoid data overfitting in the final model.

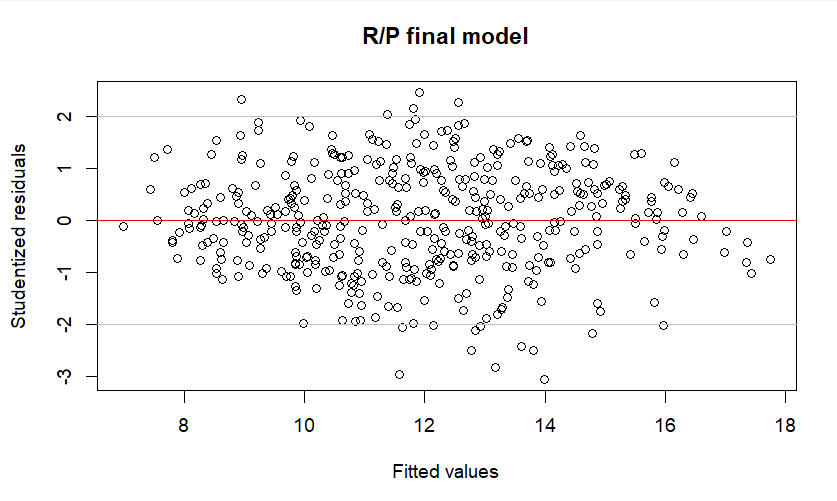
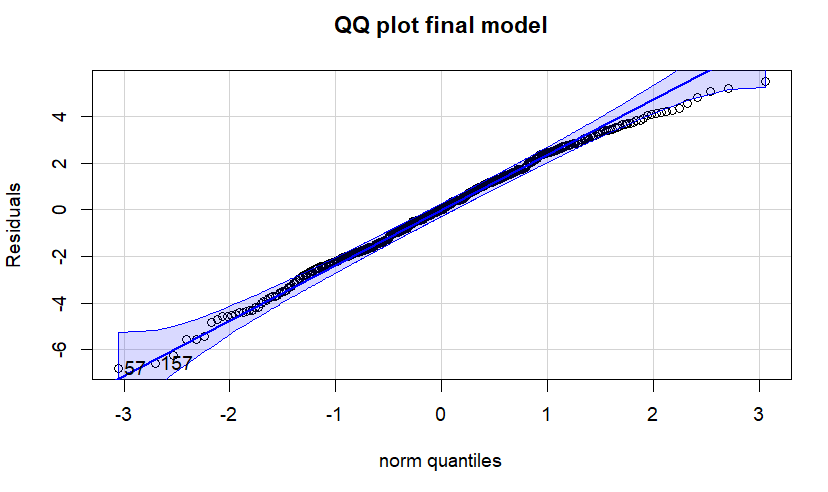
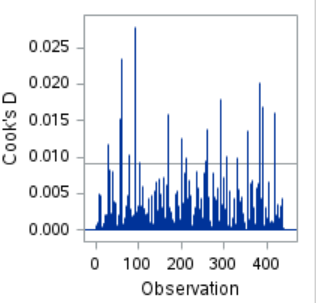
1. Diagnostic check of final model

The VIFs for each predictor in the final model are shown below:

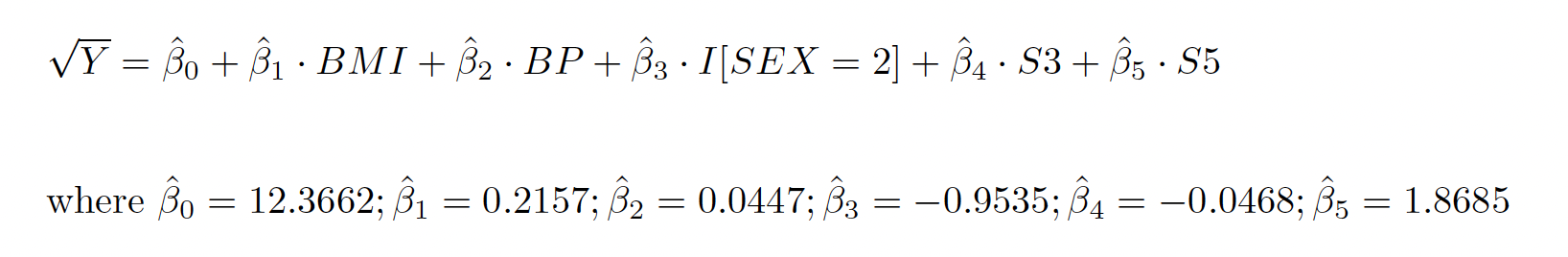
A number and text on a white background

Description automatically generated

Each predictor has a VIF much lower than 10, which indicates that multicollinearity is no longer an issue in the final model. The following residual plot for the final model also indicates homogeneity was met. In addition, the residual plot shows linearity and the QQ plot shows Gaussian errors. A side-by-side comparison of residual plots shows that the fanning pattern is less visible in the final model than in the full model. We also observed one less outlier in the final model than in the full model. (We define outliers in this report as residuals ± 2 standard deviations from the mean). For outliers in the final model, we tested their Cook’s Distance to check their degree of influence on the model fit by measuring the standardized shift in and when that observation is deleted from the model. No outlier was excessively influential since none had a Cook’s distance near or above 1.



**Results**

We determined that BMI, BP (blood pressure), sex, S3 (high-density lipoproteins), and S5 (triglyceride) were the most important predictors for diabetes progression. The final model met all regression assumptions, so the parameter estimates were defined as below:

The model reduced to four continuous predictors and one categorical predictor. Fitting estimates gave an intercept that represents the mean diabetes progression value when all predictors are at their mean values if Y had not been transformed. However, we transformed Y to √Y in our model, and square root transformation does not allow linear or proportional relationships between variables. Instead, we interpreted estimates by positive or negative sign.

A screenshot of a computer

Description automatically generatedThe R analysis output:

Each parameter estimate resulted in a p-value of less than 0.001, which is significant at an alpha level of 0.005. This alpha level was calculated with the Bonferroni correction of 10 pairwise tests. Additionally, the adjusted R2 was 0.4961, so the model explained 49.61% of the variation in the data.

In general, we predict diabetes progression to increase on average for patients with a high BMI, high blood pressure, de-identified sex of 1, low HDL cholesterol, and high triglyceride levels one year after baseline. From this model, the most significant predictors of diabetes progression are BMI and triglyceride (t-value >7.3). Although there was high collinearity in S3 and S5 in the full model, we were surprised to see both variables without collinear issues in the final model. Reduced collinearity may be due to removal of confounder variables in the full model and variables with poor fit. Also, it makes sense that health measures like diet, weight, and blood pressure may be related to diabetes, but we did not expect sex to be related to diabetes as well.

**Discussion and Conclusion**

The results suggested that BMI, blood pressure, serum triglycerides are risk factors for diabetes progression, while female sex and high-density lipoproteins are protective variables of diabetes progression.

Our study has several limitations. Firstly, some of the predictor variables are not thoroughly documented in our data source. The sex variable was deidentified, so we assumed 1 represented male and 2 represented female in the analysis. Similarly, the response variable Y was only described as a measure of diabetes progression and we assumed higher values signified greater disease progression (i.e., worse health). If either of the assumption of variables was not met, our interpretation of findings would have to change accordingly. Secondly, we fitted the model based on a relatively small sample size, some of the predictors would be accidentally included or excluded using the forward selection method based on the p-values. Thirdly, there are potential unmeasured variables outside of our data that can be significant predictors for diabetes disease progression, such as race, smoking status, dietary factors, and family history of diabetes. Lastly, our study result may have limited external validity, which means the final model we proposed may not be fit for other populations of diabetes patients that display distinct characteristics from our study population.

To address these limitations, more data needs to be collected from diabetes patients with diverse range of demographics and biomarkers, so we can obtain a potentially more accurate statistical model and properly assess adaptability of the use of our model. Additionally, interviews with clinicians and medical practitioners can provide us with domain knowledge to inform our model selection process, which can be helpful improving the interpretability of our disease progression model in secondary diabetes prevention.

**Reference**

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani "Least angle regression," The Annals of Statistics, Ann. Statist. 32(2), 407-499, (April 2004)

Centers for Disease Control and Prevention. (2023, September 5). What is diabetes?. Centers for Disease Control and Prevention. <https://www.cdc.gov/diabetes/basics/diabetes.html>

Newsom, J. T. (2024). Centering in multilevel regression. https://web.pdx.edu/~newsomj/mlrclass/ho\_centering.pdf

**Appendix**

Table 1: Results from Interaction tests

