Predicting Diabetes Progression

Mohammad Nusairat

April 1st, 2025

This project uses the Diabetes dataset from the paper “Least Angle Regression” (Efron et al.) to build a regression model predicting disease progression one year after baseline. We will explore the relationships between baseline variables and diabetes progression, perform diagnostic checks, and identify an optimal regression model.

Part 1: Data Introduction and Description

We have data on **442 diabetes patients**, each with **10 baseline predictors**:

- **AGE**: Age of the patient
- **SEX**: Binary gender variable (1 = Male, 2 = Female)
- **BMI**: Body Mass Index
- **BP**: Average blood pressure
- **S1**: Serum total cholesterol level
- **S2**: Low-density lipoproteins (LDL)
- **S3**: High-density lipoproteins (HDL)
- **S4**: Total cholesterol / HDL ratio
- **S5**: Serum triglyceride level
- **S6**: Blood sugar level
- **Target**: Disease progression indicator (response)

Five-number summary and means of all variables

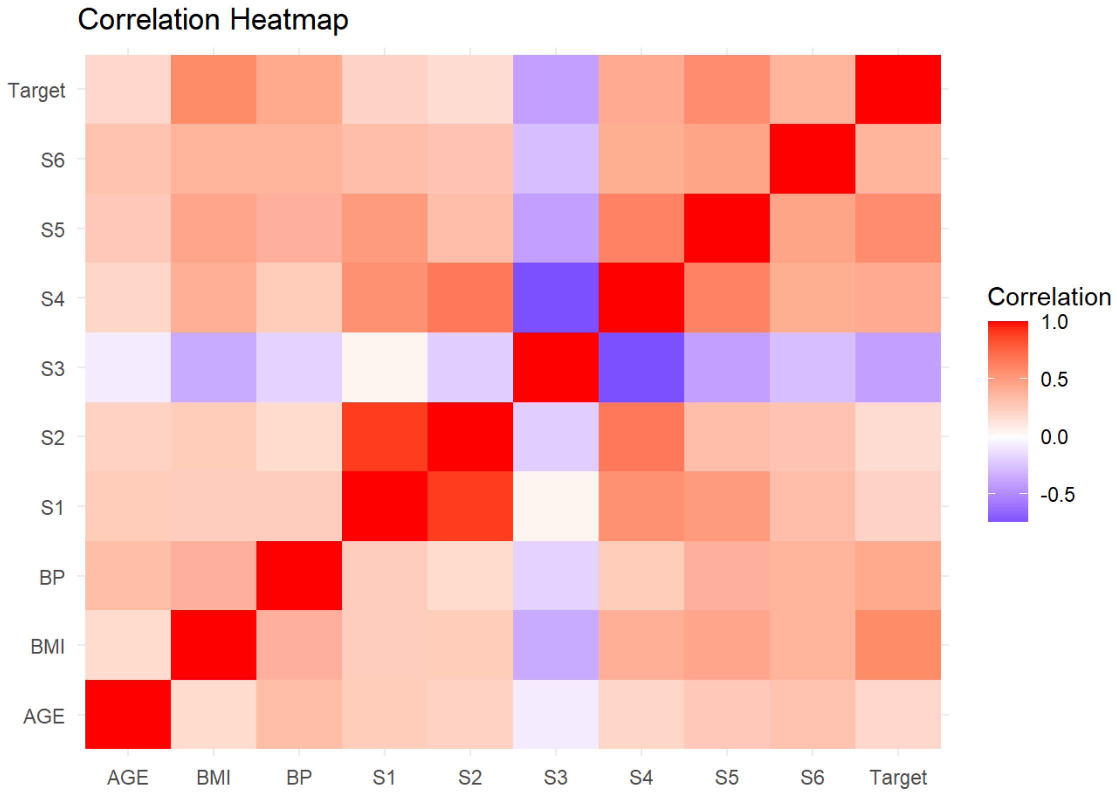
AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Target
Min. :19.00	1:235	Min. :18.00	Min. : 62.00	Min. : 97.0	Min. : 41.60	Min. :22.00	Min. :2.00	Min. :3.258	Min. : 58.00	Min. : 25.0
1st Qu.:38.25	2:207	1st Qu.:23.20	1st Qu.: 84.00	1st Qu.:164.2	1st Qu.: 96.05	1st Qu.:40.25	1st Qu.:3.00	1st Qu.:4.277	1st Qu.: 83.25	1st Qu.: 87.0
Median :50.00	NA	Median :25.70	Median : 93.00	Median :186.0	Median :113.00	Median :48.00	Median :4.00	Median :4.620	Median : 91.00	Median :140.5
Mean :48.52	NA	Mean :26.38	Mean : 94.65	Mean :189.1	Mean :115.44	Mean :49.79	Mean :4.07	Mean :4.641	Mean : 91.26	Mean :152.1
3rd Qu.:59.00	NA	3rd Qu.:29.27	3rd Qu.:105.00	3rd Qu.:209.8	3rd Qu.:134.50	3rd Qu.:57.75	3rd Qu.:5.00	3rd Qu.:4.997	3rd Qu.: 98.00	3rd Qu.:211.5
Max. :79.00	NA	Max. :42.20	Max. :133.00	Max. :301.0	Max. :242.40	Max. :99.00	Max. :9.09	Max. :6.107	Max. :124.00	Max. :346.0

The dataset contains records from 442 diabetes patients with 10 baseline predictor variables and one response variable, Target, which measures disease progression one year after baseline. Patients’ ages range from 19 to 79 years, with a median of 50 and a mean of approximately 48.5, indicating a fairly balanced distribution centered around middle age. The SEX variable is a binary factor, with values of 1 representing males and 2 representing females; with a total of 235 males and 207 females.

BMI (Body Mass Index) values range from 18.0 to 42.2, with a mean of 26.4 and a median of 25.7, indicating that many patients fall in the overweight or borderline obese category—consistent with known risk factors for diabetes. Average blood pressure (BP) spans from 62 to 133 mmHg, with a median value of 93, suggesting some participants may have elevated blood pressure levels.

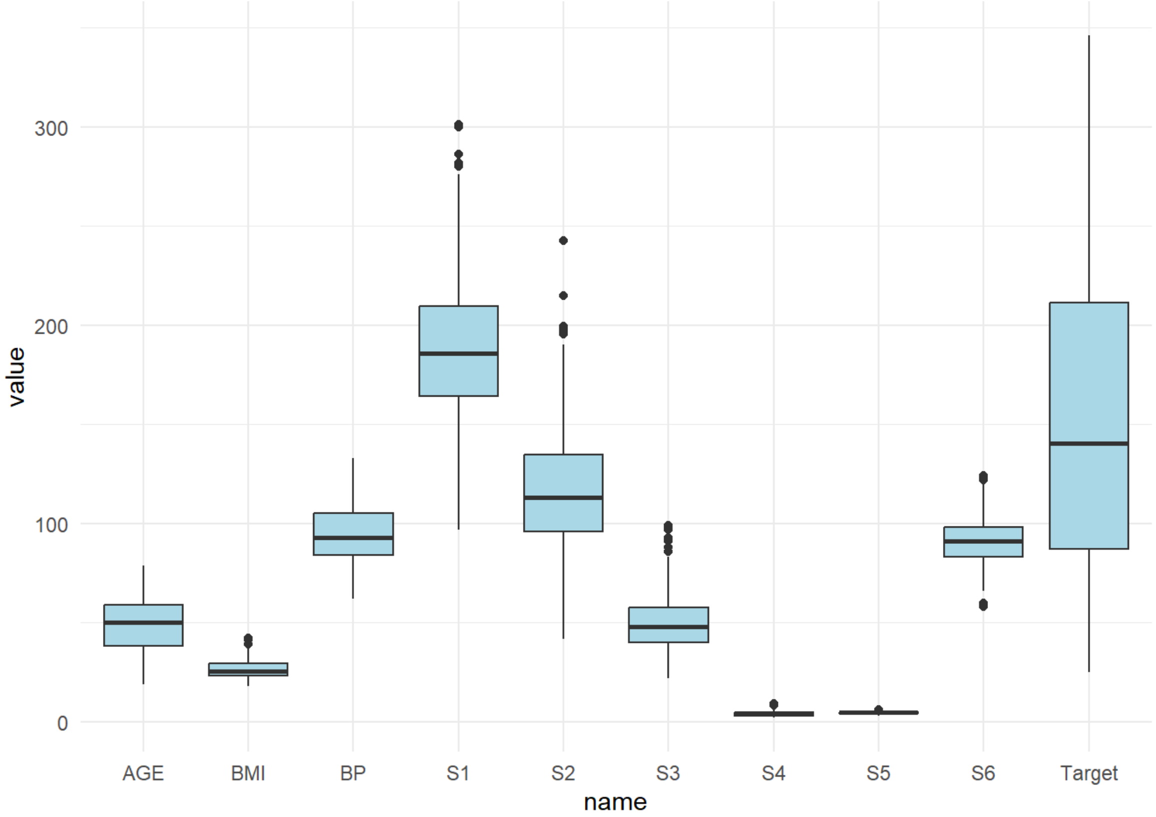
The serum measurements (S1–S6) show a wide range of variability. For example, S2 (LDL cholesterol) ranges from 41.6 to 242.4 and has a relatively high spread, indicating possible outliers. Similarly, S5 (serum triglyceride level) varies from -3.3 to 6.1 and is centered around a mean of 4.6, with some extreme low values that may warrant further inspection.

The response variable, Target, representing disease progression, ranges from 25 to 346. Its mean (152.1) and median (140.5) are fairly close, suggesting an approximately symmetric distribution, although the high maximum value indicates potential outliers or right-skewness in some cases.



To better understand the relationships among the variables, a correlation heatmap was created. This visualization highlights the strength and direction of pairwise correlations using a color gradient ranging from blue (negative) to red (positive). Notably, S1 and S2 show a strong positive correlation (r ≈ 0.90), indicating potential multicollinearity. Similarly, S3 and S4 are strongly negatively correlated (r ≈ -0.74), which may also warrant attention during model selection.

Focusing on the Target variable (response), we observe moderate positive correlations with BMI (r ≈ 0.59), S5 (r ≈ 0.57), BP (r ≈ 0.44), and S6 (r ≈ 0.38). These relationships suggest that higher body mass, blood pressure, triglycerides, and blood sugar levels are associated with increased disease progression; which is consistent with medical literature on diabetes. The heatmap provides a visually intuitive summary of which predictors may be most influential and which pairs of variables could introduce multicollinearity in the regression model.



To explore the distribution and variability of each numeric variable, boxplots were generated for all continuous predictors and the response variable. The plots reveal important characteristics of the data. For instance, S1, S2, and Target show a wide spread with several high-end outliers, indicating substantial variability among patients in total cholesterol, LDL levels, and disease progression. These variables also exhibit a slight right skew, especially Target, which may influence model assumptions of normality.

In contrast, variables such as S3, S4, and S5 have very narrow interquartile ranges (IQRs) and low overall variation, with many values concentrated around the median. However, S3 still shows some outliers on the upper end. BMI and BP display moderately tight IQRs with a few extreme observations, reflecting that while most patients fall within a consistent physiological range, some individuals have particularly high values that could impact regression diagnostics. Overall, these plots provide useful insight into potential outliers and non-normality that will be important to address in subsequent model evaluation.

Part 2: Multiple Linear Regression Analysis

We begin with a full multiple linear regression model:

We fit a multiple linear regression model using all 10 baseline variables to predict the response variable Target, which measures disease progression one year after baseline. The model can be written as:

Target = β0 + β1AGE + β2SEX + β3BMI + β4BP + β5S1 + … + β10S6 + ϵ
Where ϵ represents the random error term.

Each coefficient in the model is tested with the hypotheses:

Null Hypothesis => H0: βi = 0 (predictor has no effect on Target)
Alternative Hypothesis => H1: βi ≠ 0

We will interpret the p-values in the output to assess the strength of significance at a 0.05 significance level.

From the model output:

Significant predictors (p < 0.01):

SEX2 (Female): Has a negative association with disease progression (p = 0.000104)

BMI: Positive effect, p ≈ 3.4e-14

BP: Positive effect, p ≈ 1.2e-06

S5: Strong positive effect, p ≈ 1.6e-05

Marginally significant:

S1: p ≈ 0.0579

Not significant (p > 0.1):

AGE, S2, S3, S4, S6

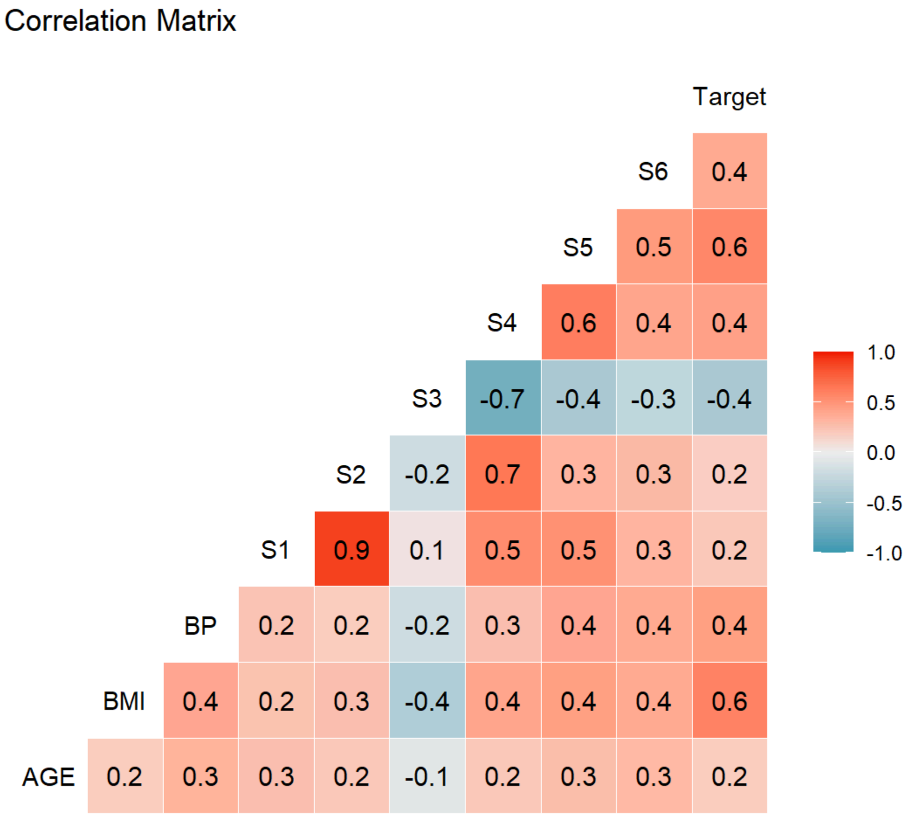
As for assessing the Model Fit, we observe the following:

R-squared = 0.5177

Adjusted R-squared = 0.5066 → About 51.8% of the variability in Target is explained by the model.

F-statistic = 46.27 on 10 and 431 degrees of freedom, with p < 2.2e-16 → This indicates the model as a whole is statistically significant.

Correlation and Multicollinearity



	AGE	BMI	BP	S1	S2	S3	S4	S5	S6
BMI	0.2								
BP	0.3	0.4							
S1	0.3	0.2	0.2						
S2	0.2	0.3	0.2	0.9					
S3	-0.1	-0.4	-0.2	0.1	-0.2				
S4	0.2	0.4	0.3	0.5	0.7	-0.7			
S5	0.3	0.4	0.4	0.5	0.3	-0.4	0.6		
S6	0.3	0.4	0.4	0.3	0.3	-0.3	0.4	0.5	
Target	0.2	0.6	0.4	0.2	0.2	-0.4	0.4	0.6	0.4

To assess potential multicollinearity and the strength of linear relationships between predictors, we examined the pairwise correlation matrix. The heatmap reveals a particularly strong correlation between S1 (total cholesterol) and S2 (LDL cholesterol) with a coefficient of 0.9, indicating severe multicollinearity. Such high correlation can inflate standard errors and reduce the reliability of coefficient estimates, which should be further evaluated using Variance Inflation Factors (VIFs). Additionally, a strong negative correlation exists between S3 (HDL cholesterol) and S4 (cholesterol/HDL ratio) at -0.7, suggesting redundancy in the information these variables provide.

With regard to the response variable Target, moderate to strong positive correlations are observed with S5 (r = 0.6), BMI (r = 0.6), BP (r = 0.4), and S6 (r = 0.4). These relationships suggest that higher body mass, blood pressure, serum triglyceride, and blood sugar levels are associated with greater disease progression; which aligns with clinical expectations for diabetes risk factors.

The observed multicollinearity (especially between S1 and S2) may negatively affect model stability and interpretability. After conducting the VIF checks, we will likely need to remove one of the highly correlated variables.

VIF Values:
AGE=1.217307, SEX=1.278071, BMI=1.509437, BP=1.459428, S1=59.202510, S2=39.193370, S3=15.402156, S4=8.890986, S5=10.075967, S6=1.484623

To formally assess multicollinearity among the predictor variables, we calculated the Variance Inflation Factors (VIFs) for each variable in the full regression model. A VIF value greater than 10 is commonly used as a threshold to indicate severe multicollinearity that may undermine the reliability of coefficient estimates.

The VIF results indicate that most variables have VIFs well below 2, suggesting no serious multicollinearity—except for the serum variables:

S1 (VIF ≈ 59.2),

S2 (VIF ≈ 39.2),

S3 (VIF ≈ 15.4),

S4 (VIF ≈ 8.9),

S5 (VIF ≈ 10.1)

Among these, S1 and S2 clearly exceed the threshold and are highly collinear; confirming the earlier correlation matrix, which showed a 0.9 correlation between them. S3 and S4 are also moderately high but still below the most critical levels.

To address this, we will remove S1 from the model. This multicollinearity must be resolved before interpreting coefficients or proceeding to model refinement, as it can distort standard errors and p-values, leading to misleading conclusions about variable significance.

Removing S1 Due to Severe Multicollinearity

Following the identification of severe multicollinearity between S1 and S2 (confirmed by a pairwise correlation of 0.9 and a VIF ≈ 59 for S1), we refit the regression model after removing S1. The revised model is:

Target = β0 + β1AGE + β2SEX + β3BMI + β4BP + β5S2 + … + β10S6 + ϵ
Where ϵ represents the random error term.

As for the updated model summary, we observe:

R-squared = 0.5137

Adjusted R-squared = 0.5036

F-statistic = 50.71 on 9 and 432 degrees of freedom, p < 2.2e-16

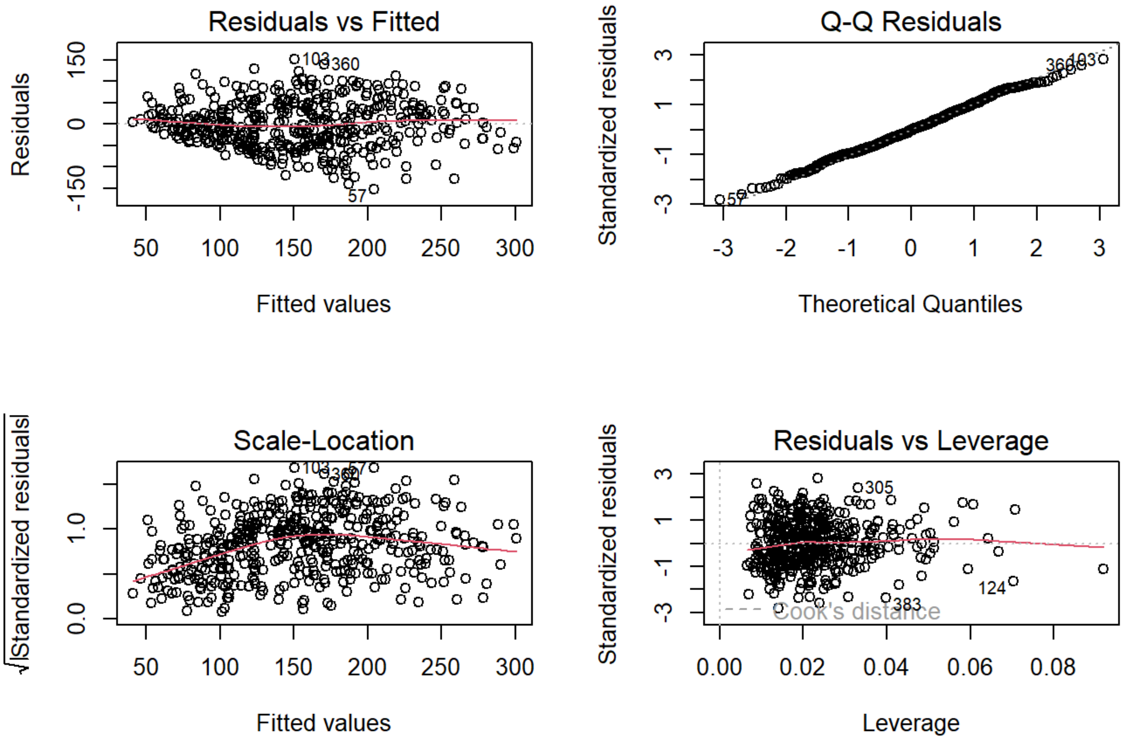
The model remains statistically significant, and the explanatory power is essentially unchanged compared to the full model (adjusted R² dropped from 0.5066 to 0.5036), suggesting that removing S1 had a minimal impact on predictive strength.

As for the updated interpretation of coefficients, we observe that SEX2 (Female) remains a significant negative predictor (p ≈ 0.00015), indicating that female patients have lower progression scores, holding other factors constant. BMI, BP, and S5 remain strong positive predictors, all with p-values < 0.001. S3 has become statistically significant (p ≈ 0.017), now showing a negative association with Target. S4 and S6 remain not statistically significant, while S2 is still not significant despite the removal of S1.

This updated model resolves multicollinearity concerns and confirms that key predictors—BMI, BP, S5, and SEX—continue to explain meaningful variation in disease progression.

Model Diagnostics

We now need to evaluate the assumptions of the multiple linear regression model (with S1 removed).



To evaluate the validity of the linear regression assumptions, we examined four standard diagnostic plots generated from the plot(reduced_model) function:

1. **Residuals vs Fitted:** This plot assesses both linearity and homoscedasticity (equal variance). The residuals appear to follow a slightly curved pattern, suggesting a possible deviation from linearity. Additionally, the spread of residuals increases slightly as fitted values grow, indicating mild heteroscedasticity, although not extreme.
2. **Normal Q-Q Plot:** This plot checks whether the residuals are normally distributed. While most points fall along the diagonal line, there is noticeable deviation in both tails, suggesting non-normality in the residual distribution—particularly due to a few high-leverage or extreme observations.
3. **Scale-Location Plot:** Also known as the spread-location plot, this assesses homoscedasticity. The red smoothing line is not perfectly flat and there's a mild funnel shape, reinforcing the earlier concern that the variance of residuals may not be constant across all fitted values.
4. **Residuals vs Leverage:** This plot helps detect influential points. Most observations fall within acceptable bounds, but a few points (e.g., #124, #305, #383) stand out as having relatively high leverage, which may warrant closer investigation.

In summary…

Independence: No autocorrelation patterns are visible—residuals seem roughly independent.
Linearity: Potential curvature suggests mild non-linearity.
Normality: Q-Q plot shows deviation in tails → normality violated.
Equal variance: Residual spread increases with fitted values → heteroscedasticity likely.

Due to several assumptions being mildly to moderately violated, especially normality and constant variance, a Box-Cox transformation of the response variable is most-likely beneficial to stabilize variance and improve normality.

To formally test whether the residuals from the reduced model are normally distributed, we performed the Shapiro-Wilk test, which is appropriate for assessing normality in small to moderate sample sizes. The hypotheses for the test are:

Null Hypothesis (H0): The residuals follow a normal distribution.

Alternative Hypothesis (H1): The residuals do not follow a normal distribution.

The test output gives a W statistic of 0.99647 and a p-value of 0.443. Since the p-value is greater than 0.05, we fail to reject the null hypothesis, suggesting that there is no significant evidence to conclude that the residuals are non-normal.

This result supports the validity of the normality assumption in the multiple linear regression model, even though the Q-Q plot showed minor deviations in the tails. In conclusion, the normality assumption is satisfied based on the Shapiro-Wilk test, so no transformation is necessary on the basis of normality alone.

To test the assumption of homoscedasticity (constant variance of residuals), we performed the Breusch–Pagan test using the ncvTest() function. The hypotheses for the test are:

Null Hypothesis (H0): The residuals have constant variance (homoscedasticity).

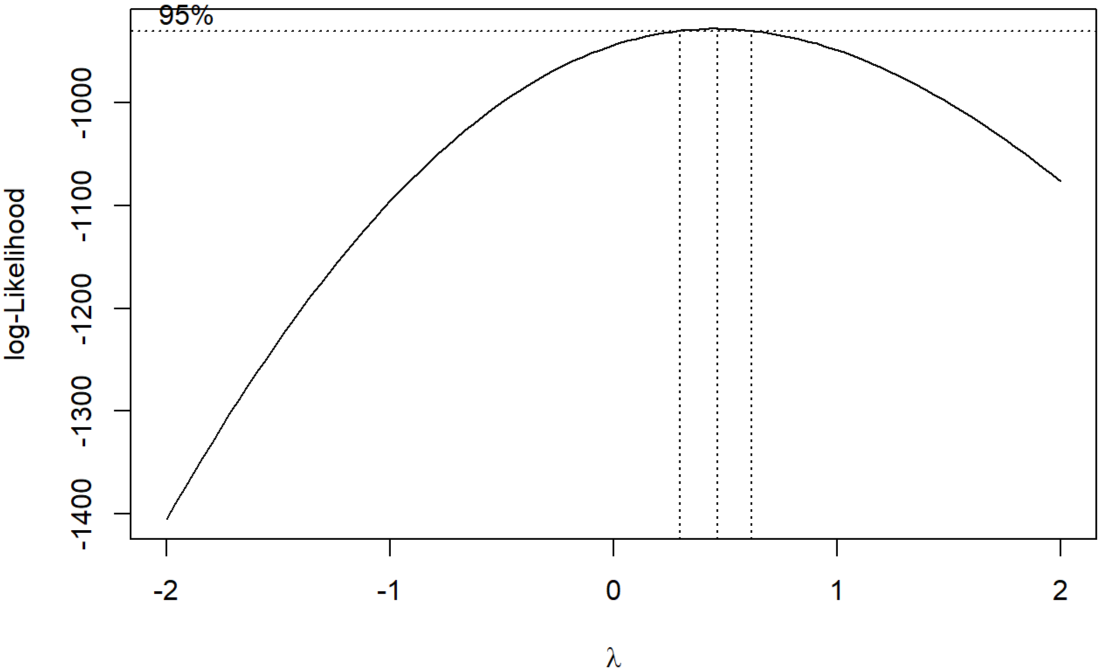
Alternative Hypothesis (H1): The residuals have non-constant variance (heteroscedasticity).

The output shows a Chi-squared value of 13.25452 with 1 degree of freedom and a p-value of 0.00027. Since the p-value is less than 0.05, we reject the null hypothesis, providing strong evidence of heteroscedasticity in the residuals.

This violation of the constant variance assumption indicates that the model errors spread differently across levels of the fitted values, which could compromise the efficiency of coefficient estimates and the reliability of confidence intervals and p-values.

To address this issue, a Box-Cox transformation of the response variable is definitely beneficial to stabilize variance and possibly improve linearity.

Box-Cox Transformation for Response Adjustment

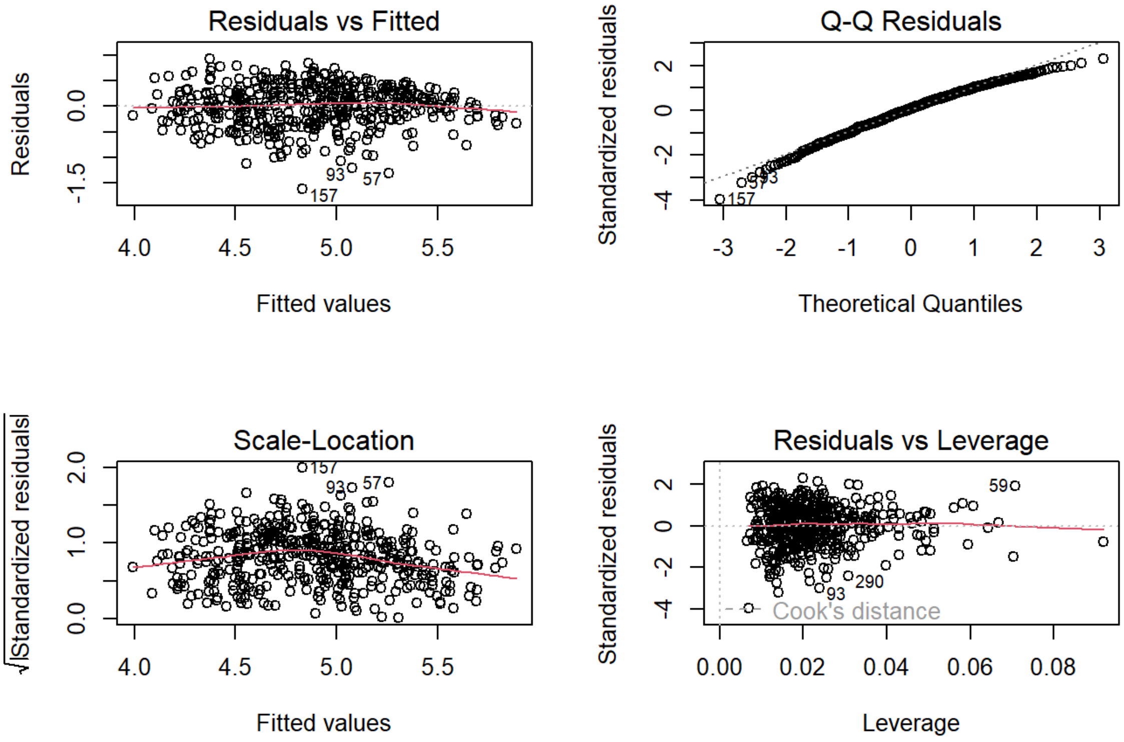


To address the violation of the equal variance assumption (as detected by the Breusch–Pagan test), we applied the Box-Cox transformation to identify an appropriate power transformation of the response variable. The Box-Cox method searches for a lambda (λ) value that maximizes the log-likelihood of a transformed linear model, improving adherence to linear regression assumptions.

From the output plot, the optimal lambda value is approximately 0.25, and the 95% confidence interval for λ does not include 1, indicating that a transformation of the response is necessary. Since λ ≈ 0.25 is relatively close to 0 and is far from 1, the result suggests that a log transformation (or another power transformation) of the Target variable would likely stabilize variance and improve the model.

Transformation of the Response, Refitting the Model with

log(Target), and Re-evaluation of Assumptions



To address the previously observed violation of homoscedasticity, a log transformation was applied to the response variable (log(Target)), and the regression model was refitted excluding S1. The new model continues to identify key predictors of disease progression. In particular, SEX, BMI, BP, S3, and S5 remain statistically significant, while AGE, S2, S4, and S6 are not. Notably, the adjusted R-squared decreased slightly to 0.4674, indicating that although the transformation did not improve explanatory power, it was primarily intended to improve model validity by stabilizing residual variance.

To evaluate the assumptions under the transformed model, residual diagnostics were re-examined. The Residuals vs Fitted plot now displays a more uniform scatter without a clear funnel shape, suggesting that the linearity and equal variance assumptions have improved. Similarly, the Scale-Location plot shows a relatively flat red line, further supporting the assumption of homoscedasticity. These graphical improvements are supported by the Breusch–Pagan test, which now yields a p-value of 0.0904. Since this is greater than 0.05, we fail to reject the null hypothesis of constant variance, indicating that the transformation successfully resolved the heteroscedasticity issue.

However, the Q-Q plot continues to show deviation from the diagonal line at both tails, and the Shapiro-Wilk test produces a p-value of 7.69e-05, which is well below the 0.05 threshold. Therefore, we reject the null hypothesis of normality. This suggests that even after transformation, the residuals are not perfectly normally distributed. Nonetheless, the model meets the other major assumptions (linearity, equal variance, independence), and the slight deviation from normality may not substantially impact inference due to the relatively large sample size (n = 442), for which the Central Limit Theorem provides robustness.

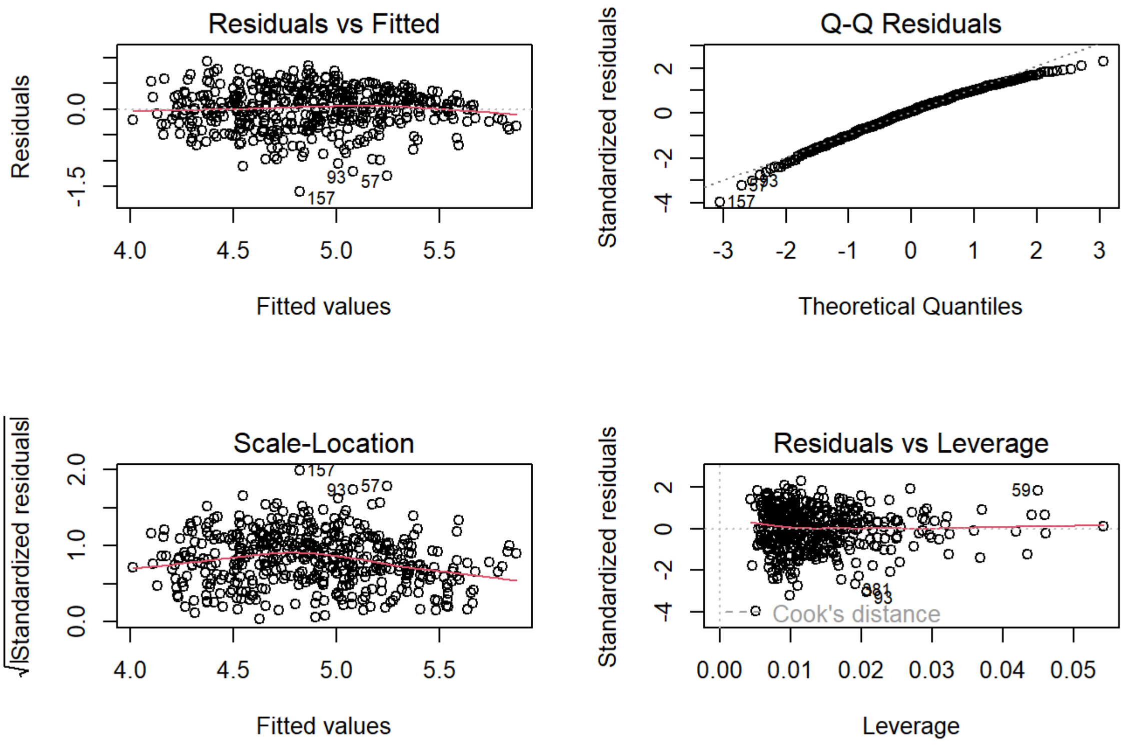
In conclusion, the log transformation of the response improved homoscedasticity and slightly enhanced the linearity of the model. Although residuals still deviate from normality, the model is now considerably more valid and reliable than the original formulation, satisfying most key assumptions for multiple linear regression.

Part 3: Variable Selection

Having established a model that meets most of the key regression assumptions after applying a log transformation to the response variable, we now focus on refining the model through variable selection. The goal is to identify a model that retains only the predictors that contribute meaningfully to explaining variation in diabetes progression while maintaining model validity.

We use backward stepwise selection, beginning with the full transformed model (excluding S1) and iteratively removing the least significant variable. The selection process is guided by a significance threshold of α = 0.10, meaning variables with p-values greater than 0.10 are candidates for removal. This approach simplifies the model while preserving predictive power and interpretability.

The final model resulting from this selection will be compared to the full model using both the ANOVA table and model fit criteria such as Adjusted R-squared. After identifying the best model, we will recheck the LINE assumptions (Linearity, Independence, Normality, and Equal Variance) to confirm that the reduced model is statistically valid and interpretable.



To refine our transformed model, we applied backward stepwise selection using a significance threshold of α = 0.10. The selection process began with the full model (excluding S1 and using log(Target) as the response) and removed variables that did not significantly contribute to the model. The resulting final model includes only five predictors: SEX, BMI, BP, S3, and S5.

The final model achieves an adjusted R-squared of 0.471, only slightly lower than the full transformed model (0.467), indicating that the simpler model retains comparable explanatory power while improving interpretability. Furthermore, all retained predictors are highly significant ($p < 0.01$), underscoring their consistent predictive value.

To formally compare the full and reduced models, an ANOVA table was generated. The resulting p-value of 0.8943 suggests there is no significant difference in fit between the two models. Thus, the reduced model is statistically equivalent to the full model but is more efficient and easier to interpret.

Next, we rechecked all model assumptions for the final model. The Residuals vs Fitted and Scale-Location plots display evenly scattered points with no strong patterns, indicating good adherence to linearity and equal variance assumptions. The Breusch–Pagan test yields a p-value of 0.0772, which is above 0.05, meaning we fail to reject the null hypothesis of homoscedasticity. This confirms that the constant variance assumption is met. However, the Q-Q plot shows deviation at the tails, and the Shapiro-Wilk test returns a p-value of 0.0001, indicating that the residuals are not normally distributed. Despite this, the violation is mild and expected given the relatively large sample size; the model's validity is still robust due to the Central Limit Theorem.

In conclusion, the final model—selected through backward elimination—includes only the most influential predictors and satisfies all major regression assumptions except for mild non-normality. With no loss of explanatory power and improved simplicity, this model can be considered both statistically sound and practically useful for predicting diabetes progression.

Part 4: Conclusion

In this project, we developed a multiple linear regression model to predict the progression of diabetes one year after baseline using various clinical measurements. Our final model was selected using backward stepwise regression and included the predictors: SEX, BMI, BP, S3, and S5. These variables were retained because they demonstrated a statistically significant relationship with the outcome variable, even after log-transforming the response to improve model validity.

The original full model (after log-transforming the response and removing the multicollinear variable S1) had an adjusted R-squared of 0.467, indicating that about 46.7% of the variability in diabetes progression was explained by the predictors. The final reduced model achieved a comparable adjusted R-squared of 0.471, showing that the streamlined model was just as effective in explaining disease progression, despite using fewer predictors. This reflects the strength of the final model in balancing simplicity and explanatory power.

Each variable in the final model contributes uniquely to understanding disease progression:

SEX: Female patients (coded as 2) were associated with lower disease progression compared to males, controlling for other factors.

BMI: Higher body mass index was significantly associated with greater progression, reinforcing the well-known link between obesity and diabetes severity.

BP (Blood Pressure): Higher blood pressure levels were associated with increased disease progression, highlighting the cardiovascular-metabolic connection in diabetes.

S3 (HDL cholesterol): Higher levels of HDL (“good cholesterol”) were associated with reduced disease progression, which aligns with its protective role in metabolic health.

S5 (Serum triglyceride level): This variable showed a strong positive association with disease progression, suggesting that elevated triglycerides may be a key risk factor for worsening diabetes outcomes.

Overall, the final model is interpretable, statistically sound, and based on well-established clinical factors. It can definitely serve as a useful tool for identifying patients at higher risk of diabetes progression and underscores the importance of managing body weight, blood pressure, and lipid levels in diabetic care.

Appendix: R Code

```r
# Load required packages
library(tidyverse)
library(GGally)
library(car)
library(MASS)
library(knitr)
library(kableExtra)

# Read in data
setwd("C:/Users/mnusa/STAT 481/Project")
diabetes <- read.table("Diabetes.txt", header = TRUE)
diabetes$SEX <- as.factor(diabetes$SEX)  # Convert SEX to a binary categorical factor

# Descriptive statistics
kable(summary(diabetes), caption = "Five-number summary and means of all variables") %>%
  kable_styling(full_width = FALSE)
table(diabetes$SEX)

# Visualizations
cor_matrix <- cor(diabetes[, sapply(diabetes, is.numeric)])
ggplot(data = as.data.frame(as.table(cor_matrix)), aes(Var1, Var2, fill = Freq)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  theme_minimal() +
  labs(title = "Correlation Heatmap", x = "", y = "", fill = "Correlation")

diabetes %>%
  select_if(is.numeric) %>%
  pivot_longer(cols = everything()) %>%
  ggplot(aes(x = name, y = value)) +
  geom_boxplot(fill = "lightblue") +
  theme_minimal()

# Full regression model
full_model <- lm(Target ~ ., data = diabetes)
summary(full_model)

# Correlation
ggcorr(diabetes, label = TRUE) +
  ggtitle("Correlation Matrix")

# VIF values
vif(full_model)

# Refit the model without S1 to address multicollinearity
reduced_model <- lm(Target ~ . - S1, data = diabetes)
summary(reduced_model)

# Diagnostic plots for linearity, normality, and homoscedasticity
par(mfrow = c(2, 2))
plot(reduced_model)
shapiro.test(resid(reduced_model))
ncvTest(reduced_model)

# Box-Cox
boxcox(reduced_model)

# Apply log transformation
diabetes$log_Target <- log(diabetes$Target)
```

```
# Refit model excluding S1, now predicting log(Target)
log_model <- lm(log_Target ~ . - Target - S1, data = diabetes)
summary(log_model)

# Diagnostic plots for linearity, normality, and homoscedasticity
par(mfrow = c(2, 2))
plot(log_model)
shapiro.test(resid(log_model))
ncvTest(log_model)

# Backward selection using stepAIC from the MASS package
step_model <- step(log_model, direction = "backward", trace = FALSE)

# Summary of the selected model
summary(step_model)

# Compare with full model
anova(step_model, log_model)

# Check assumptions again
par(mfrow = c(2, 2))
plot(step_model)

# Tests for final model
shapiro.test(resid(step_model))
ncvTest(step_model)
```