Inferring the Relationship Between Blood Glucose Levels and Clinical Health Indicators

Callixta Cahyaningrum, Fabrizzio Bernie, Syifa Izzah, Kezia Wijaya

---

1. Introduction

1.1 Data

Our project uses the Pima Indian Diabetes Dataset from Kaggle. The dataset contains health information for 768 female patients of Pima Indian heritage, aged 21 years and older. It also consists of seven independent variables and one dependent variable. The key variables analyzed are:

1. BMI (Body Mass Index): A measure of body fat based on weight and height.
2. Pregnancies: Number of times the patient has been pregnant.
3. Glucose: Blood glucose level (mg/dL).
4. Blood Pressure: Diastolic blood pressure (mm Hg).
5. Skin Thickness: Measurement of skinfold thickness (mm).
6. Insulin: Blood insulin level (μU/mL).
7. Diabetes Pedigree Function: A genetic risk score for diabetes.
8. Age: Patient's age (years)

We acknowledge the limitation this dataset has since it comprises only a certain ethnic group and is further limited by having only one gender.

1.2 Motivation and Research Question

Type 2 diabetes is a chronic health condition that affects over 400 million people worldwide (World Health Organization, 2023). Its projections stipulate that there will be a rise in prevalence in the coming decades. Monitoring blood glucose levels can be used for early detection and prevention of diabetes since elevated levels of it indicate weakened insulin regulation. Which in turn causes abnormal glucose regulation that is a precursor to type 2 diabetes.

We are interested in how other phsyiological factors in the dataset affect blood glucose levels, hence why we aim to explore their relationship. Our focus is on understanding this relationship using linear regression, variable selection techniques, and log transformations if needed. Our end goal is to interpret and quantify these relationships in a meaningful way. This leads us to our research question:

What is the relationship between blood glucose levels and physiological variables in the Pima Indian female population?

---

2. Analysis

2.1 Data Pre-processing and Cleaning

We imported the dataset directly from the Kaggle-hosted URL and used the read.csv() to directly read the dataset and store it into an object named diabetes. After examining the dataset, we found multiple 0 entries in Insulin, Glucose, BloodPressure, and BMI, which is not physiologically possible. For example, a BMI or BloodPressure of 0 most likely reflects missing or unrecorded data instead of the actual value being 0.

To address this issue, we removed all observations with a 0 value in any of these four variables. This is done to ensure the dataset is not affected by missing or invalid entries and ensure the reliability and interpretability of our analysis.

2.2 Exploratory Data Analysis

Before selecting variables for modelling and fitting the model, we examined the summary statistics of the dataset. This initial step helped us better the relationships between covariates and blood glucose level. Specifically, we computed the linear correlation coefficient between glucose level and each of the potential covariates.

Table 1.1. Correlation between potential covariates and the response variable

Covariates	**Correlation with glucose**
Pregnancies	0.1982910
BloodPressure	0.2100266
SkinThickness	0.1988558
Insulin	**0.5812230**
BMI	0.2095159
DiabetesPedigreeFunction	0.1401802
Age	0.3436415

This analysis revealed that most variables have weak positive linear correlation with glucose levels. From all the variables examined, Insulin showed the strongest positive linear correlation with glucose level (r = 0.581) while the others showed moderate or weak associations. These linear correlations suggest that no single variable is a dominant predictor and that a multivariate approach might be preferred.

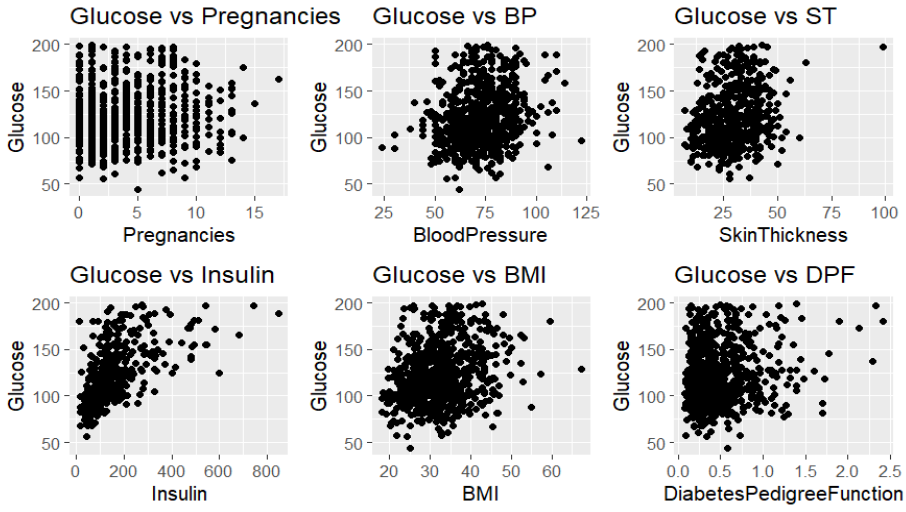


Figure 2.1. Scatterplot of Glucose With Each Covariate

To better understand how each potential covariates correlate to glucose level, we visualized all the scatterplots of glucose against all other covariates (**Figure 2.1**). As we can see, most covariates tend to have a weak positive trend except for insulin, which shows a clearer upward trend.

Initial exploration revealed high right-skewness in Insulin and SkinThickness, justifying a log transformation. Pairwise correlation analysis suggested moderate positive correlations between Glucose and variables like Insulin, Age, and BMI.

These findings are consistent with our earlier correlation analysis from the summary statistics. Insulin seems to be the most informative predictor, but it seems the overall pattern points toward the need for multivariate modelling to capture the combined effects present in the data.

2.2 Model Selection and Multicollinearity

We began by fitting a full linear model using all predictors (with log transformations where appropriate). To assess multicollinearity, we used the vif() function. All predictors had VIF values well below the common threshold of 5, indicating no serious multicollinearity concerns and justifying their inclusion in candidate models. We then applied exhaustive subset selection using the regsubsets() function to evaluate all possible combinations of predictors. Models were assessed using two criteria: Akaike Information Criterion (AIC), which balances model fit with complexity, and Adjusted R^2, which accounts for the number of predictors used. Lower AIC and higher Adjusted R^2 values indicate better model performance.

Features	**Pregnancies**	**BloodPressure**	**BMI**	**Age**	**DiabetesPedigreeFunction**	**log_Insulin**	**log_SkinThickness**
VIF	1.898333	1.220503	1.971500	2.074562	1.047623	1.141762	1.802811

Table 2.2.1 VIF Values for Each Predictor

2.3 Selected Model

Table 2.3.1 AIC and Adjusted R^2 Values for Each Model Size

Model Size	**AIC**	**Adjusted R^2**
4	3591.487	0.4245870
3	3592.195	0.4220912
5	3592.504	0.4245419
6	3594.446	0.4231328
2	3594.821	0.4167352
7	3596.428	0.4216560
1	3617.741	0.3800455

The model with the lowest AIC (3591.49) and highest Adjusted R^2 (0.4246) included four predictors:

- BloodPressure
- Age
- DiabetesPedigreeFunction
- log_insulin

However, upon further examination using backward selection, we found that DiabetesPedigreeFunction was not statistically significant at a 5% level (p-value: 0.10224). Since

the AIC and the Adjusted R^2 values after removing DiabetesPedigreeFunction are not very far from the best model obtained before (ΔAIC = 0.708 & Δadj.$R^2 \approx 0.002$). Given that removing this variable led to only a marginal increase in AIC and a slight drop in Adjusted R^2, we opted for a more parsimonious 3-variable model: BloodPressure, Age, and log(Insulin).

To assess robustness, we evaluated Cook's Distance, which identifies observations with a disproportionate influence on the regression coefficients. We applied the common rule-of-thumb threshold:

$$\text{Cook's distance} > \frac{2k}{n},$$

where k is the number of parameters in your model, including the intercept, and n is the number of observations (rows) used to fit the model. After removing high-influence observations, we refit the 3-variable mode and compare them.

Table 2.3.2 Comparison Between the Optimal Model Before and After (Cleaned) Cook's Distance

	AIC	**Adjusted R^2**
Model 3	3592.195	0.4220912
Model 3 Cleaned	3478.591	0.475858

These improvements suggest that a few data points had a strong impact on model performance and inference. While our final 3-variable model remains the primary focus, this robustness check confirms that our model is sensitive to certain observations and benefits from influence diagnostics. Therefore, we decided to choose the cleaned model as our final model, and hence the summary statistic of our final model chosen is:

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-40.10524	9.64148	-4.160	3.94×10^{-5}
BloodPressure	0.21582	0.09513	2.269	0.0238
Age	0.56879	0.12086	4.706	3.54×10^{-6}
log_insulin	26.86977	1.73453	15.491	$< 2 \times 10^{-16}$

R^2	**Adjusted R^2**
0.48	0.4759

Table 2.3.3 Summary Statistics of The Final Model

So, the fitted model would be:

$$\widehat{Glucose} = -40.105 + 0.216 \cdot (BloodPressure) + 0.569 \cdot (Age) + 26.870 \cdot (log\_insulin)$$

The interpretation of the intercept and each predictor is as follows,

- **Intercept**: This is the expected glucose level when all predictors (BloodPressure, Age, and log_Insulin) are 0. **Note**: Since log(Insulin) = 0 implies Insulin = 1 μU/mL (after transformation), and Age and BloodPressure = 0 are not realistic in a medical setting, this intercept has limited interpretability.
- **BloodPressure**: Holding Age and log(Insulin) constant, for each 1 mm Hg increase in diastolic blood pressure, the glucose level increases by approximately 0.22 mg/dL.
- **Age**: Holding BloodPressure and log(Insulin) constant, each additional year in age is associated with an average increase of 0.57 mg/dL in glucose level.
- **log_insulin**: Holding Age and BloodPressure constant, a 1-unit increase in the natural log of Insulin corresponds to an average increase of about 26.87 mg/dL in glucose.

All of the variables were significant at a 5% significance level, and about 48% of the variance in glucose levels is explained by BloodPressure, Age, and log(Insulin) in the fitted model. Then, since the adjusted R^2 value is close to R^2, it suggests that all included predictors contribute meaningfully, and the model is not overfitting.

2.4 Model Diagnostics

To evaluate the validity of the linear regression model, the model diagnostic check was performed to assess the key assumptions of linearity, normality of residuals, homoscedasticity, and leverage. For the model that had been selected by the exhaustive subset selection, we plotted the final model residuals and noticed that there were notable observed outliers from the model, which can be seen in **Appendix 5.3**. Therefore, we consider removing it from the model, and then after we removed the outliers by the Cook's distance, we plotted the final model residuals and got the result as follows,

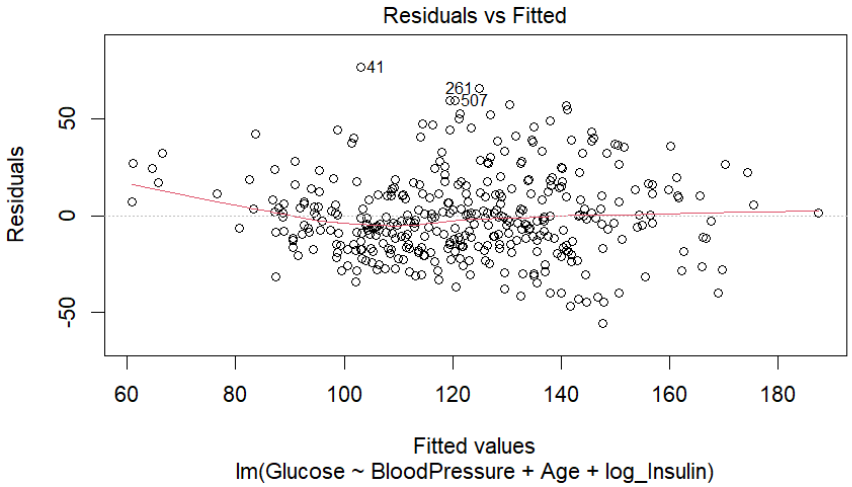


Figure 2.4.1. Residuals vs Fitted Plot of The Final Model

It can be seen from **Figure 2.4.1** that the red line is relatively flat, suggesting that the assumption of linearity is generally satisfied. Although there are a few points that show small deviations at the middle fitted values, there is no strong evidence of curvature, indicating that the model does not have severe multicollinearity.

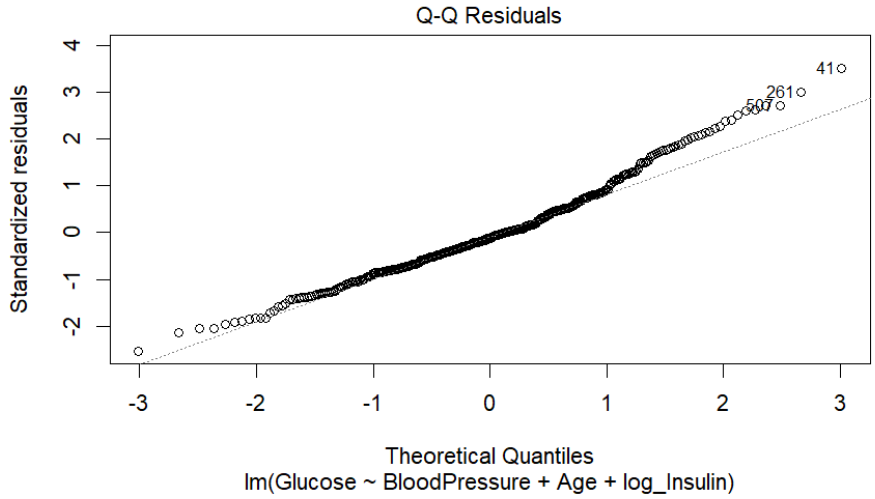


Figure 2.4.2 Normal Q-Q Plot of The Final Model

The normality plot from **Figure 2.4.2** showed that most residuals are close to the 45-degree reference line, suggesting that the residuals are approximately normally distributed. There are only a few observations at the higher theoretical quantiles that deviate from the line in the upper tail, suggesting mild right skewness, but still not indicating extreme deviations.

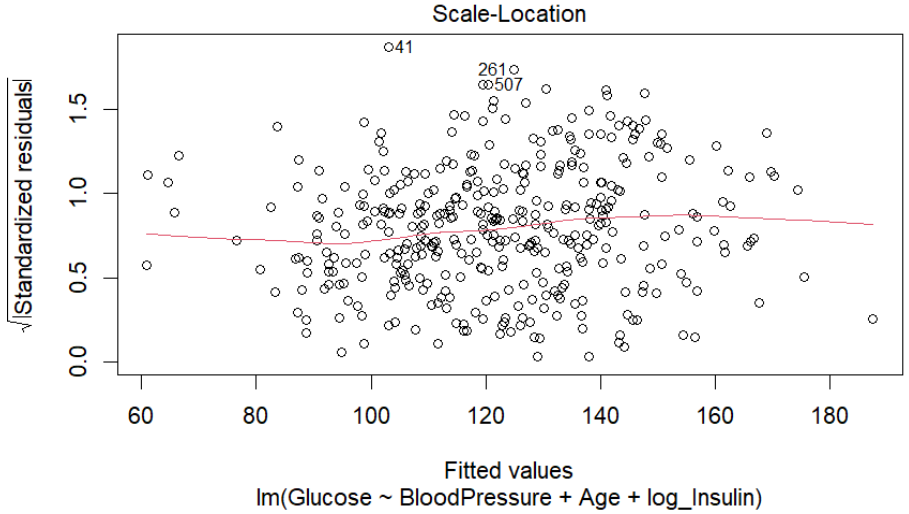


Figure 2.4.3. Scale Location Plot of The Final Model

As we can see from **Figure 2.4.3**, the red trend line appears relatively horizontal, and the spread of residuals remains fairly consistent across the range of fitted values, indicating that there is no heteroskedasticity and the assumptions of the homoskedasticity (constant variance) are not violated (although there is some variability there is no clear funnel shape observed).

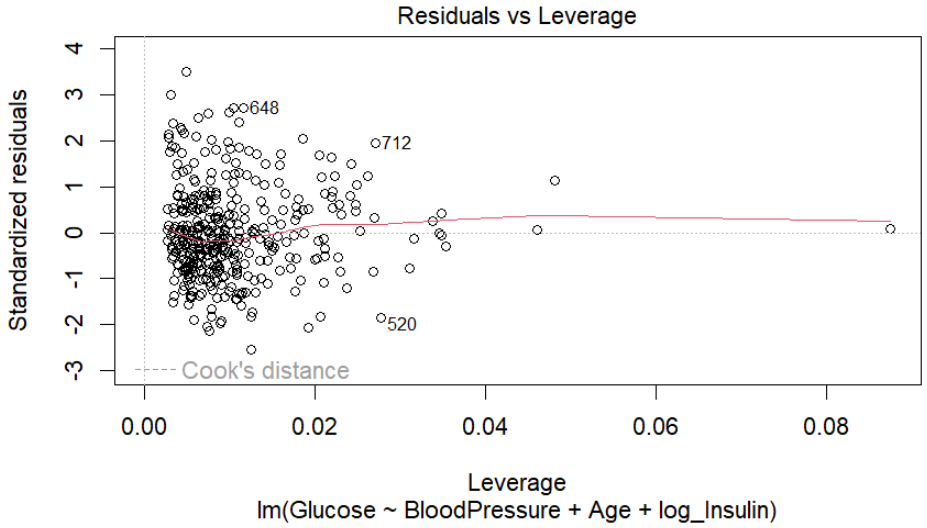


Figure 2.4.4. Residuals vs Leverage Plot of The Final Model

The plot from **Figure 2.4.4** showed that the majority of the observations clustered around low leverage and low residual values, stable and consistent predictions across most of the dataset. Although a few points exhibit slightly elevated leverage or residuals, none of them exceed Cook's distance threshold, which suggests that these observations are not overly influential. The absence of high-influence points after outlier removal supports that the model is now more robust and its estimates are more reliable.

---

3. Discussion & Conclusion

This study explored how various physiological variables relate to blood glucose concentration in adult women of Pima Indian heritage from the data collected by the National Institute of Diabetes and Digestive and Kidney Diseases. Using a multiple linear regression model, we aimed to identify which health indicators are significantly associated with glucose levels.

Our final model identifies three physiological variables significantly associated with blood glucose levels:

- **log(Insulin):** The strongest predictor in our model, insulin levels showed a strong positive association with blood glucose. The coefficient for log(Insulin) is 25.245, with a p-value of 0.000, suggesting that higher log(Insulin) is associated with higher blood glucose levels. This finding aligns with the biological role of insulin in glucose regulation, where higher insulin levels may indicate insulin resistance, a core feature of Type 2 Diabetes.
- **Age:** Age has a positive correlation with blood glucose level, with its coefficient being 0.535 and p-value of 0.000. This suggests that older individuals tend to have higher glucose levels, which is consistent with known age-related declines in insulin sensitivity and increased diabetes risk with age.
- **BloodPressure:** Blood pressure showed a positive association with blood glucose level, with a coefficient of 0.214 and a p-value of 0.032, suggesting that hypertension and high blood glucose level may occur, potentially due to the shared metabolic pathways.

Although DiabetesPedigreeFunction initially appeared to be a promising predictor, it was excluded from the final model. While its inclusion reduced the AIC value slightly, its p-value was 0.10224, indicating that it was not statistically significant at the 5% level. Additionally, it did not

meaningfully improve the Adjusted R^2 value. Based on the principle of parsimony, we opted for a simpler model with fewer covariates, as including this variable made the model more complicated without really improving how well it explained the data.

Together, these findings give insight into the physiological mechanisms underlying glucose regulation. Moreover, they also assist in early risk profiling of diabetes, particularly in populations with similar demographics to the study group. While the analysis focused on inference rather than prediction, the associations found were both statistically significant and medically meaningful to reinforce existing knowledge about metabolic health.

While the model offered meaningful insights, there are several limitations that should be acknowledged.

Limitations:

- **Sample:** The dataset is obtained from and limited to women of Pima Indian descent, which limits the ability to generalize our findings to other populations or to men, as results may not apply the same across different demographic or gender groups.
- **Missing data:** Several physiological variables contained zero values, which we treated as missing data and removed. This approach reduced the sample size which could possibly introduce bias if the missingness is not completely random.
- **Cross-sectional design:** As the data was collected at just one point in time, we are not able to identify what came first: the higher glucose levels or the changes in other variables like insulin or blood pressure. While positive associations were identified, the cross-sectional nature of the data prevents us from drawing causal conclusions.

Lastly, there is an implication of our analysis. The findings from our analysis can be used to support the medical community in early identification and risk assessment for individuals at higher risk of high blood glucose levels and potentially Type 2 Diabetes. By identifying physiological factors such as insulin levels, age, and blood pressure as significantly associated with glucose concentration, healthcare providers can focus on these indicators during routine screenings, even before diabetes is formally diagnosed. For instance, patients with higher insulin or blood pressure may benefit from earlier lifestyle interventions or closer monitoring, even if their glucose levels have not yet reached a critical threshold. Additionally, the model reinforces the importance of considering age as part of metabolic health. While our model is not predictive in nature, the associations it highlights can still inform preventative strategies and individualized care plans in clinical settings.

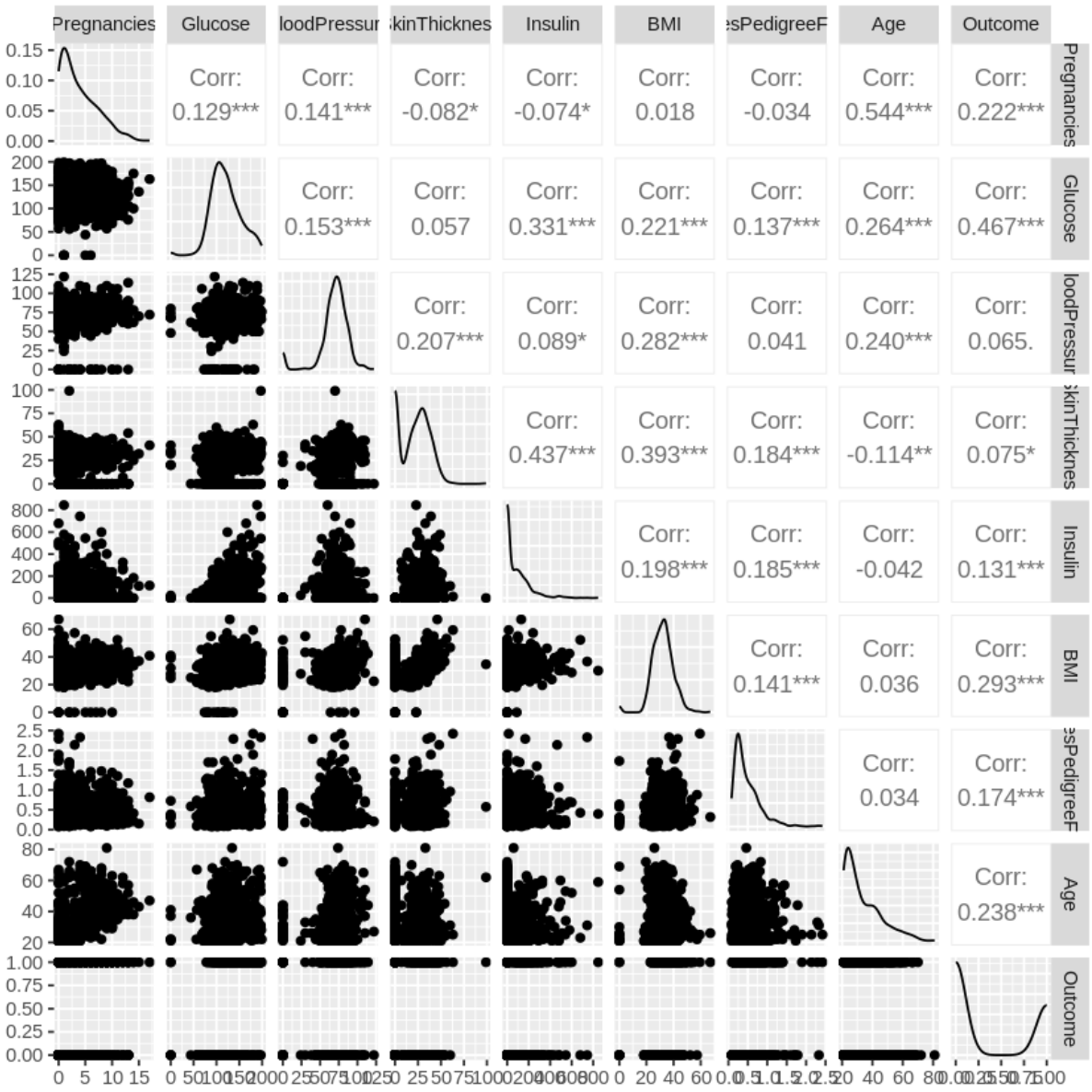
4. References

1. World Health Organization. (2023). Diabetes. Retrieved from: https://www.who.int/news-room/fact-sheets/detail/diabetes
2. https://storage.googleapis.com/kagglesdsdata/datasets/2619659/4476380/diabetes.csv?X-Goog-Algorithm=GOOG4-RSA-SHA256&X-Goog-Credential=gcp-kaggle-com%40kaggle-161607.iam.gserviceaccount.com%2F20250405%2Fauto%2Fstorage%2Fgoog4_request&X-Goog-Date=20250405T173357Z&X-Goog-Expires=259200&X-Goog-SignedHeaders=host&X-Goog-Signature=658079b58f57c534d681f9bc814b17978b44d8669d4ceebd1bad9b284ef6cc5fe06e83d551b80ead623990ab2957b2c83da752eb324e5cf48d940ce27da8ea61729e0adb69a88fca76aaa229a0336052bd864a8271d420f6dcd62f0db6186c7bce1621cf8fd791288d244ed500f150d39649541a0f91dc2cea43efb207d5e16d49595c4e07f9752c0b12e2033e7e6fc1841c2d65fd97e97a5c9c7e6c2296128c4521681de4e1287f7cf90f1aa13de62e730c4b13428e52c98e666b9031ad91d526c915562a479c96e49da565b6ed03cbcd5d97c0e211e0cef4b17b2daf77deb6d1b07d6a09a0a2b0b132bd2be168d7da066cdb4477a5eaf3cd990677cdd2a485 [link to the dataset]

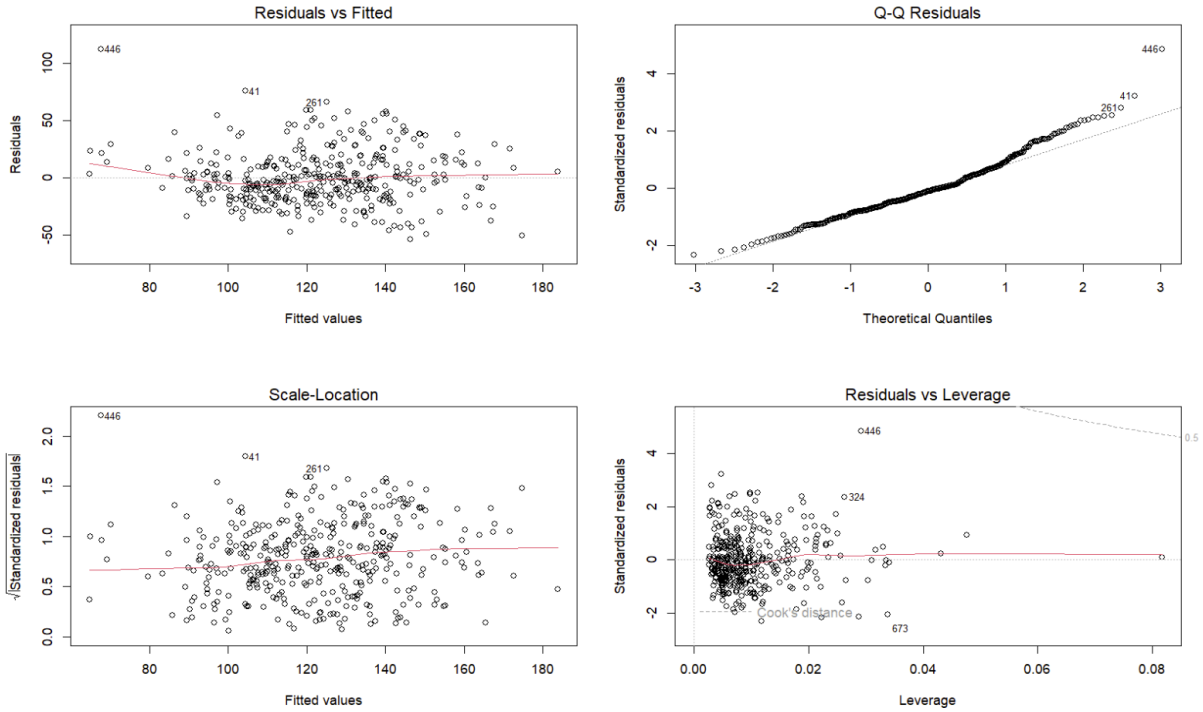
5. Appendix

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Pregnancies	0.00	1.00	3.00	3.85	6.00	17.00
Glucose	44.00	99.00	117.00	121.70	141.00	199.00
BloodPressure	24.00	64.00	72.00	72.41	80.00	122.00
SkinThickness	0.00	0.00	23.00	20.54	32.00	99.00
Insulin	14.00	76.25	125.00	155.55	190.00	846.00
BMI	18.20	27.50	32.30	32.46	36.60	67.10
DiabetesPedigreeFunction	0.078	0.2437	0.3725	0.4719	0.6262	2.42
Age	21.00	24.00	29.00	33.24	41.00	81.00

Appendix 5.1. Summary Statistics for Each Variable



Appendix 5.2. Pairwise Correlation and Scatterplots for Full Set of Covariates



Appendix 5.3. Residual Plot of The Model Before Removing Outliers