Team Project Milestone 2: Kruse Truman, Kuo Yun-Hsien, Ukagwu Arnold, Gajula Keerthan

**Changes From Milestone 1:** Changed the explanatory and response variable, Graphs and Questions.

**Variables:** Serum Creatinine( Level of serum creatinine in the blood (mg/dL), Serum Sodium ( Level of serum sodium in the blood (mEq/L), Anemia: decrease of red blood cells( 0= no, 1=yes), Age( Numerical value)

**Physiological Relationship:**

**Serum Creatinine and Anemia:**

Serum creatinine is a key marker for assessing kidney function. Elevated levels may indicate impaired renal function due to the kidneys' role in creatinine filtration. Anemia, characterized by low red blood cells or hemoglobin, can affect renal function by reducing oxygen-carrying capacity. In response, the body may increase blood flow, impacting creatinine filtration and influencing serum levels.

**Serum Sodium and Anemia:** Serum sodium levels are tightly regulated, influenced by hydration and hormones. Anemia can indirectly impact these levels by affecting blood volume and fluid dynamics. Severe anemia may increase heart rate and cardiac output, potentially influencing sodium balance.

**Interaction Between Serum Creatinine, Serum Sodium, and Anemia:** The relationship among serum creatinine, serum sodium, and anemia is complex. Anemia-induced changes in blood viscosity may impact creatinine filtration by the kidneys. Moreover, anemia-related alterations in fluid balance can influence sodium concentrations

**Questions:**

**1. What is the initial relationship between serum sodium and serum creatinine in the context of anemia?**

Correlation Between Serum Sodium and Serum Creatinine:

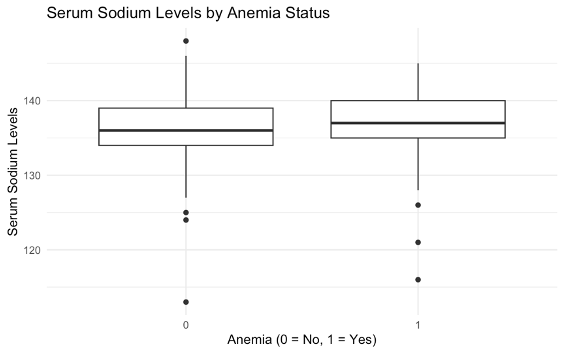
A negative correlation between serum sodium and serum creatinine levels indicates that as serum sodium levels increase, serum creatinine levels tend to decrease.

This correlation was slightly stronger in patients with anemia (−0.23) compared to those without anemia (−0.15).

Influence of Anemia on the Relationship:

The standard deviations of serum sodium were similar for patients with and without anemia, indicating that the presence of anemia did not significantly affect the overall variability in serum sodium levels. This is further supported by the large overlap in the confidence intervals for the correlation coefficients, with bound of [-0.39, -0.06254] for anemia positive, and [-0.3, -0.0043] for anemia negative.

**2. Do individuals with anemia exhibit higher variability in serum sodium levels, and does this variability relate to fluctuations in serum creatinine?**



The boxplot displays the distribution of serum sodium levels for patients with and without anaemia.

• The mean serum sodium levels are slightly different for patients with anaemia (136.84) compared to those without anaemia (136.46).

• The t-test results (t-statistic = 0.72, p-value = 0.47) indicate that this difference is not statistically significant. This suggests that the presence of anaemia does not predict variations in serum sodium levels.

• Patients with anaemia, the correlation coefficient between serum sodium and serum creatinine is −0.23

• Patients without anaemia, the correlation coefficient is −0.15

As serum sodium levels increase, serum creatinine levels tend to decrease, and vice versa. This correlation is stronger in patients with anemia. However, both values suggest a moderate negative correlation.

**3. Are there age-related differences in the associations between serum creatinine, serum sodium, and anemia?**

It appears that, after splitting the data into 10 year gaps, there is no significant difference in the correlation of serum creatinine and serum anemia based on anemia status. In fact, within a given anemia status, some 95% confidence intervals for the correlation contain zero. Specifically for the groups, anemia positive age 50-60, age 70-80, age 80-90, age 90-max and anemia negative age 60-70, age 80-70, age 90-80. Overall, given how many confidence intervals contain 0 for given ages, it brings some doubt on the overall correlation.

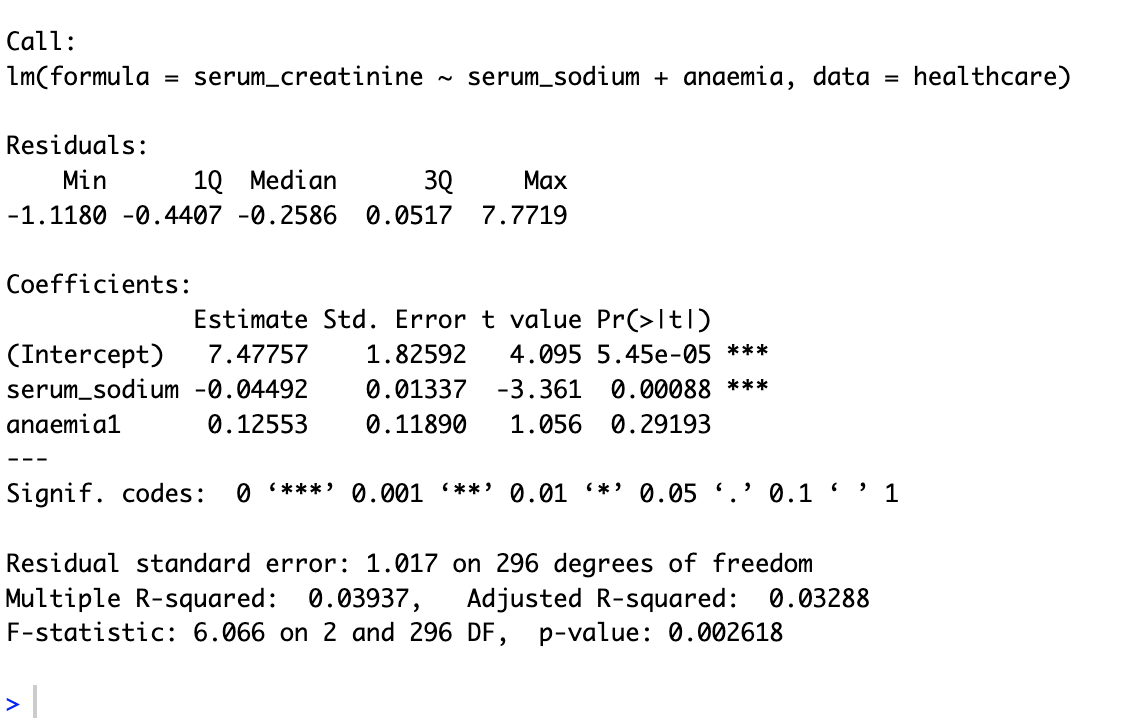
Model of Assumptions:

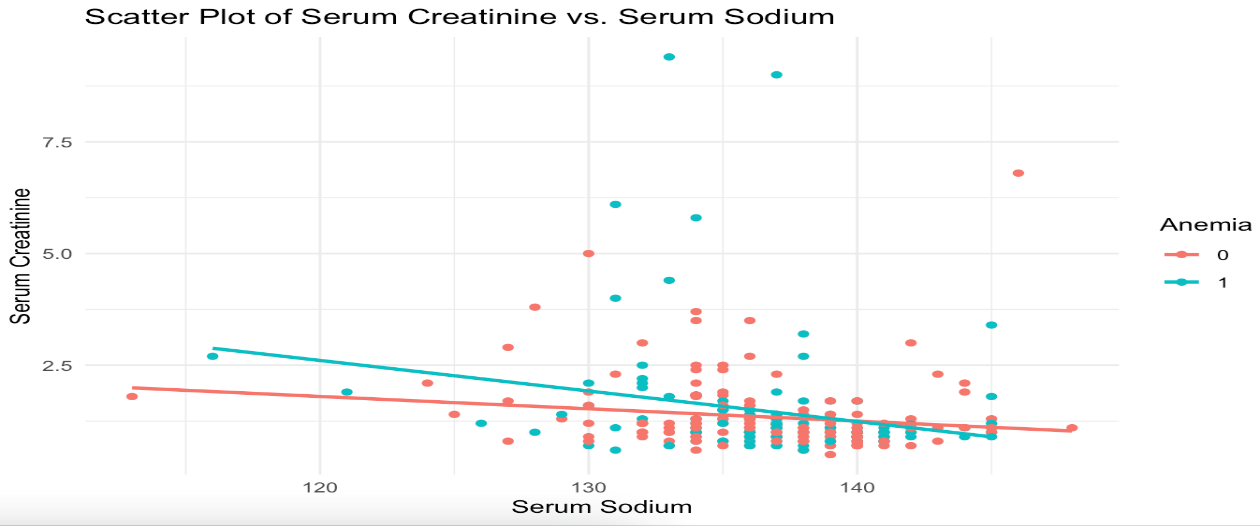
**Constant Variance:** The examination of residuals indicates the presence of constant variance. Both positive and negative residuals exhibit similar spreads across the range of predicted values, satisfying the assumption of homoscedasticity.

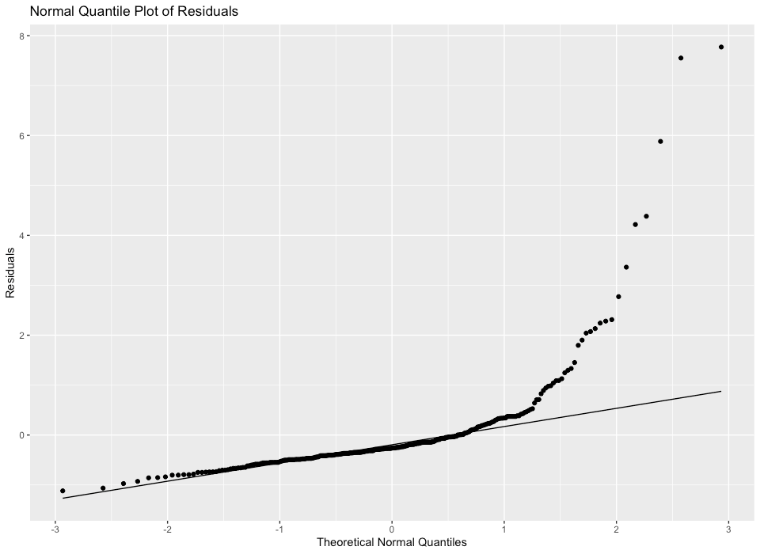
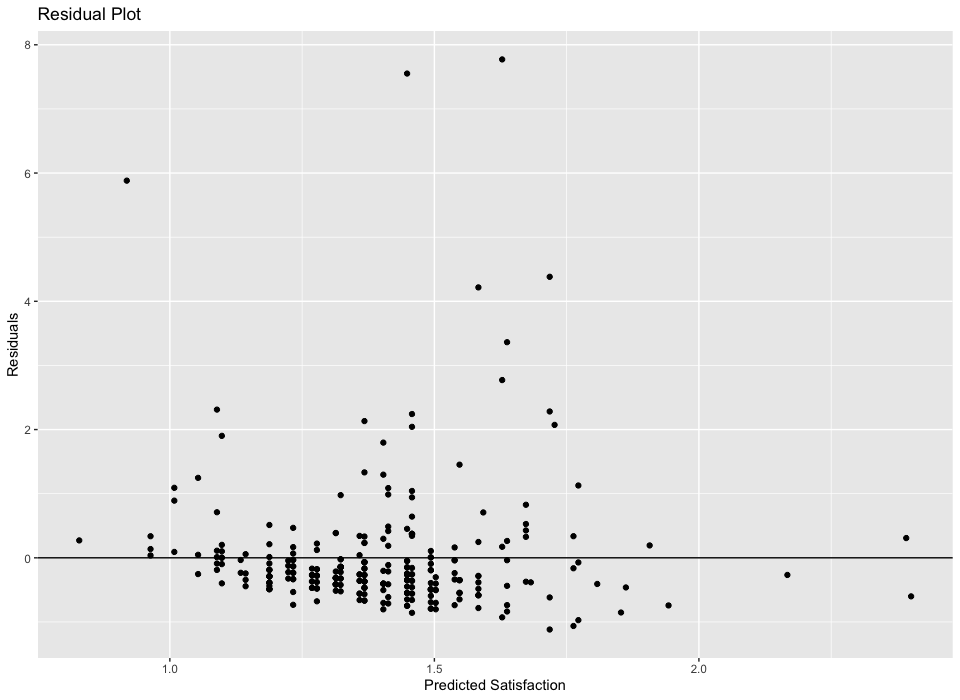
**Independence:** The residuals from the model do not display any discernible pattern or trend, suggesting that each observation is independent of the others.

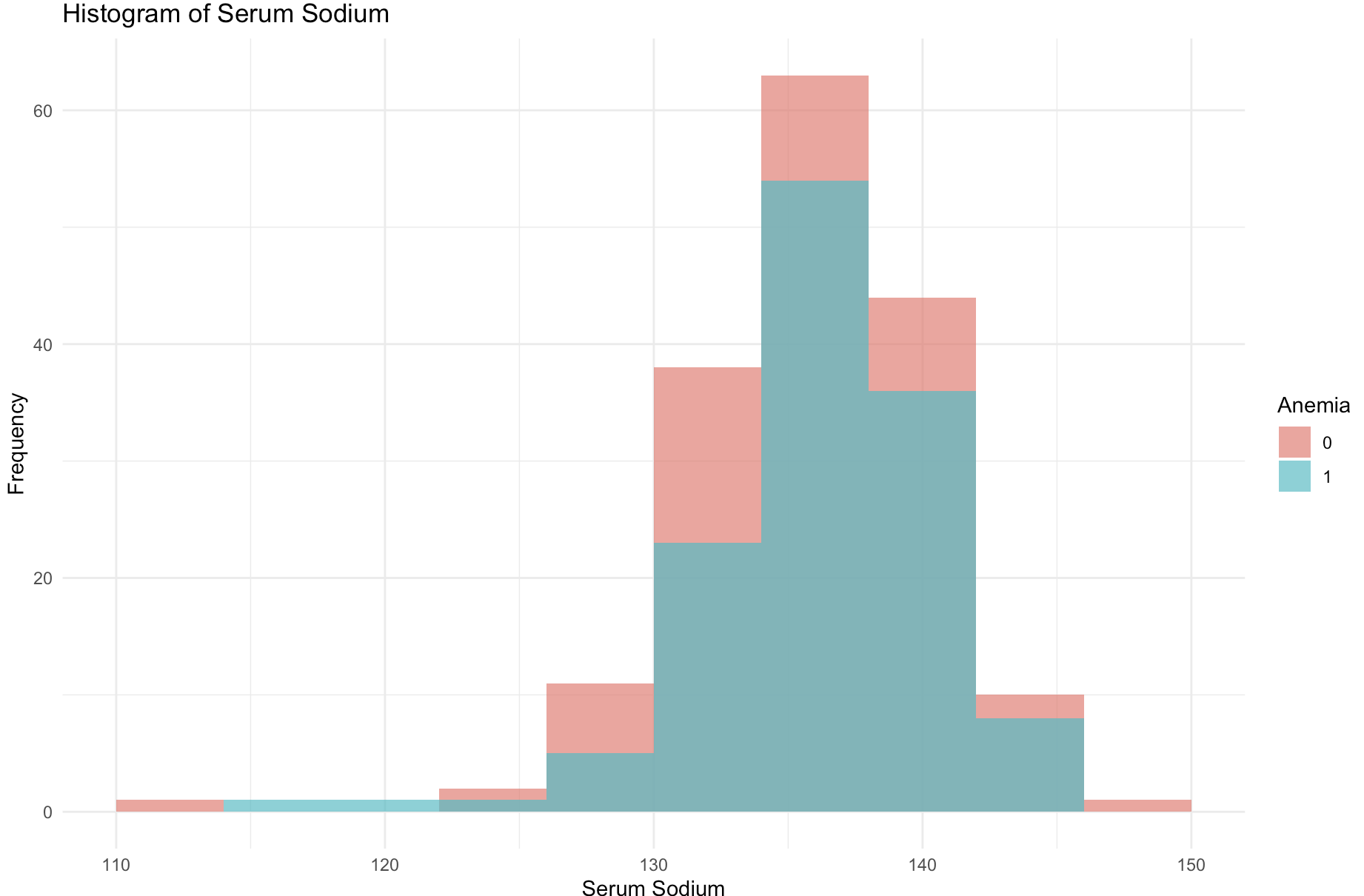
**Normal QQ Plot:** The normal quantile-quantile (QQ) plot reveals a curved shape in the residuals, indicating potential non-normality. To address this, further transformations may be necessary to achieve a more symmetrical distribution of residuals.

Performed T test:

* Null Hypothesis (H0): There is no significant relationship between serum sodium and serum creatinine in individuals with anemia
* Alternative Hypothesis (H1): There is a significant relationship between serum sodium and serum creatinine in individuals with anemia
* The coefficient for serum sodium is -0.043215.
* The t-value associated with serum sodium is -3.260.
* The p-value for serum sodium is 0.001246, which is less than the typical significance level of 0.05.
* Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis. Therefore, there is evidence to suggest that there is a significant relationship between serum sodium and serum creatinine in individuals with anemia.

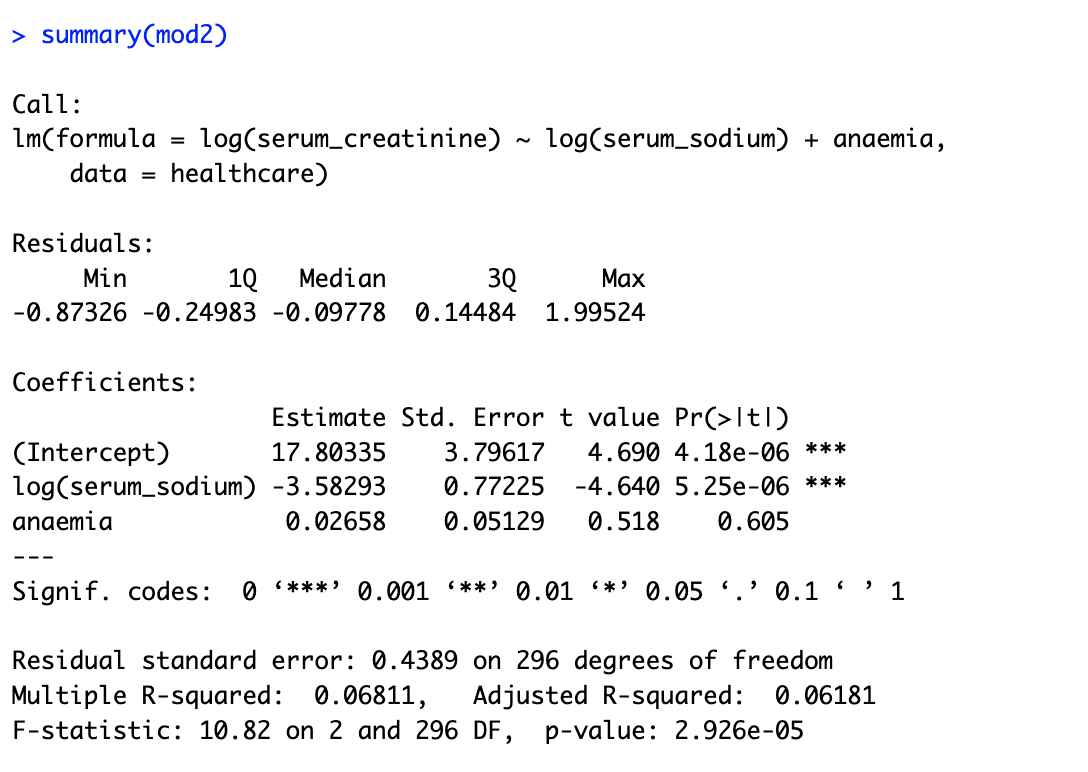


The scatterplot indicates that as serum sodium levels increase, there is a tendency for serum creatinine levels to decrease. This negative association holds true for both individuals with and without anemia, suggesting a consistent pattern across different health conditions.



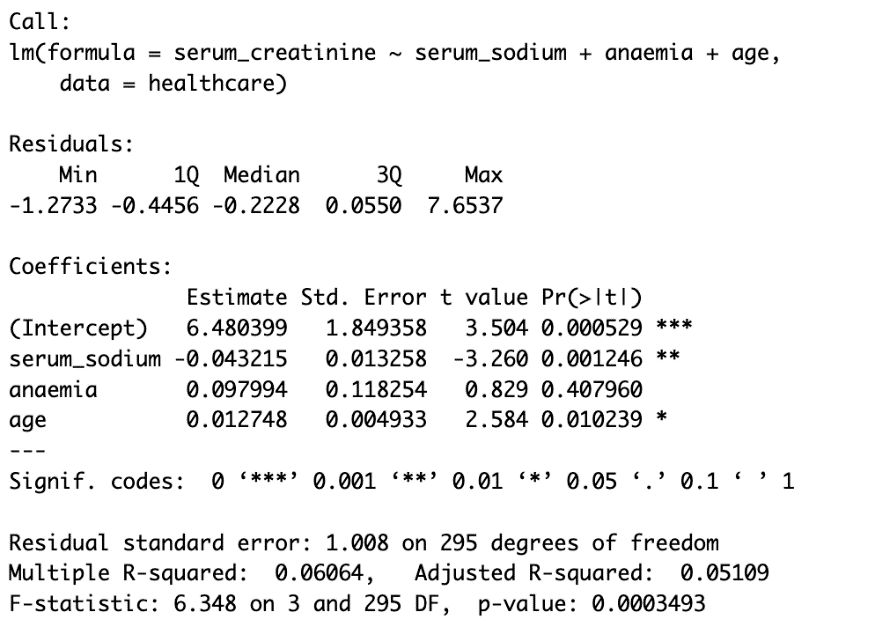
Based on model of assumptions we decided to use a log-log model due to non-linearity and non constant variance presented in the scatterplot or residual plots. Lastly there is a right skew in the histogram of the explanatory variable and response variables.

**Log-Log model Transformation:**

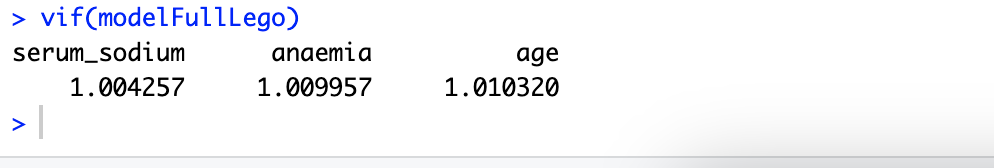


In the log-log model, the p-value is reduced compared to Model 1. Additionally, there is an increase in the multiple R-squared, indicating a greater explained variance of 68%

**Full Model:**



**Test for Multicolinearity:**



The full model

The (VIFs) for all variables in the model are below 10, indicating the absence of multicollinearity.

Reference

* Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020)
* Heart failure clinical records. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5Z89R.