Statistical Modeling of Blood Pressure

A Regression Analysis report

Prepared by:
Ezz Eldin Ahmed
Mohamed Amir
Abdulrahman Mostafa Kamel

Under Supervision of:
Dr. Wafaa Ibrahim
T.A. Moamen Ahmed

Statistics Department
Arabic Section
Fall 2024

Introduction

This study focuses on understanding the effect of various variables on blood pressure, a key indicator of cardiovascular health. Blood pressure is commonly measured in terms of systolic (the pressure in arteries when the heart beats) and diastolic (the pressure in arteries between heartbeats) values, expressed in millimeters of mercury (mmHg). The dataset under investigation includes measurements from a population sample, it includes the following variables: **Age**, **BSA** (Body Surface Area), **Weight**, **DUR** (Drug Utilization Review), **Pulse** and **Stress**.

Objective of the Study

The primary objective of this study is to analyze the relationship between blood pressure and various factors to identify potential determinants and risk indicators of hypertension. Specifically, the study aims to:

1. Determine whether the variables influence blood pressure.
2. Interpretation of the marginal effect of each variable on blood pressure.

By addressing these objectives, this study seeks to contribute to a better understanding of blood pressure patterns and the factors influencing them, thereby supporting efforts to improve cardiovascular health.

In order to study this impact, a sample of 20 individuals with high blood pressure was chosen and It had been measured the following:

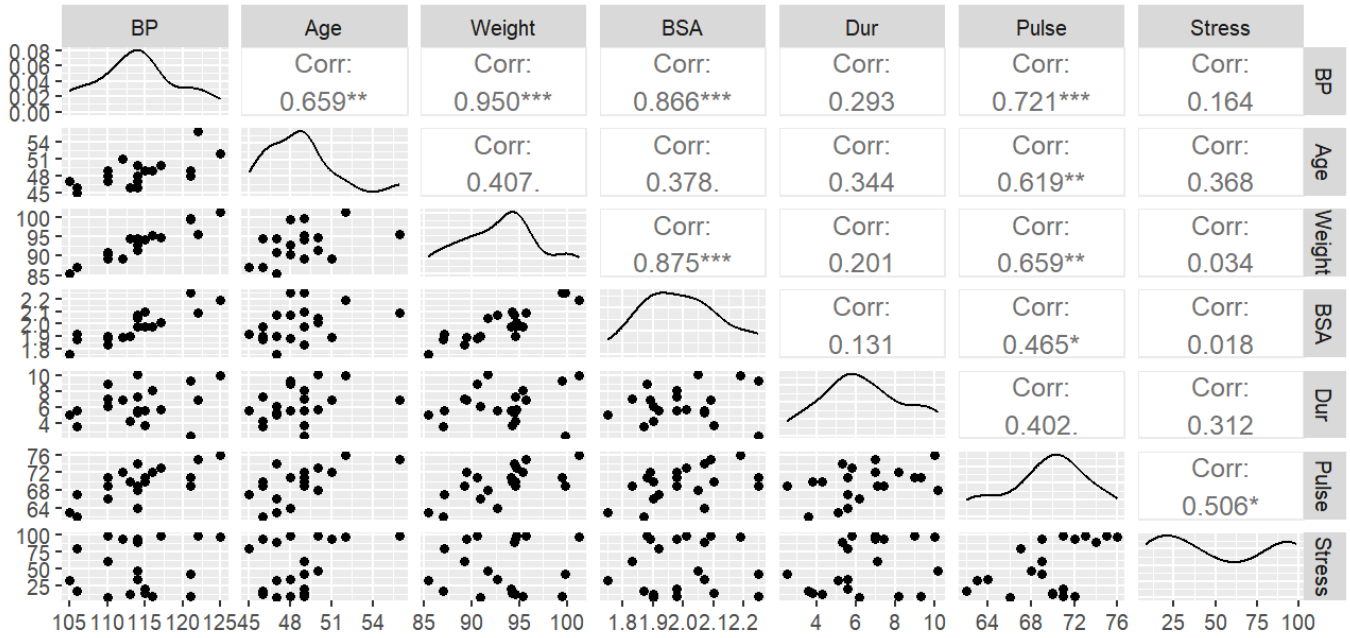
Response variable(s):

Blood Pressure: Measurement of the pressure or force of blood inside the arteries, measured in **millimeters of mercury (**mmHg).

Explanatory variable(s):

1. **BSA (Body Surface Area)**: The area of the external surface of the body, expressed in square meters (m^2); used to calculate metabolic, electrolyte, nutritional requirements, drug dosage, and expected pulmonary function measurements.

2. **DUR (Drug Utilization Review)**: Ongoing, systematic quality-improvement activity constructed to ensure the effective and appropriate use of medicines. It can also be considered a formulary system management technique.
3. **Weight**: Weight of the respondents measured in kilograms.
4. **Age**: Age of the respondents measured in years.
5. **Pulse**: The regular movement of blood through the body when the heart is beating, measured in beats per minute.
6. **Stress**: State of worry or mental tension caused by a difficult situation. Stress is a natural human response that prompts us to address challenges and threats in our lives. Everyone experiences stress to some degree.



Correlation Matrix visualization

Relations between explanatory variables and response variable:

1. A strong positive linear relationship between **Age** and **Blood Pressure** ($r = 0.659$)
2. A very strong positive linear relationship between **Weight** and **Blood Pressure** ($r = 0.95$)
3. A very strong positive linear relationship between **BSA** and **Blood Pressure** ($r = 0.866$)
4. A weak to moderate positive linear relationship between **DUR** and **Blood Pressure** ($r = 0.293$)
5. A strong positive linear relationship between **Pulse** and **Blood Pressure** ($r = 0.721$)

6. A very weak positive linear relationship between **Stress** and **Blood Pressure** ($r = 0.164$)

Relations between explanatory variables:

1. A moderate positive linear relationship between **Weight** and **Age** ($r = 0.407$).
2. A weak to moderate positive linear relationship between **BSA** and **Age**($r = 0.378$).
3. A weak to moderate linear relationship between **DUR** and **Age** ($r = 0.344$).
4. A strong linear relationship between **Pulse** and **Age** ($r = 0.619$).
5. A moderate to weak linear relationship between **Stress** and **Age** ($r = 0.368$).
6. A very strong positive linear relationship between **BSA** and **Weight** ($r = 0.875$),
7. A weak to moderate positive linear relationship between **DUR** and **Weight** ($r = 0.201$).
8. A strong positive linear relationship between **Pulse** and **Weight** ($r = 0.659$).
9. A very weak positive linear relationship between **Stress** and **Weight** ($r = 0.034$).
10. A very weak positive linear relationship between **DUR** and **BSA** ($r = 0.131$).
11. A moderate positive linear relationship between **Pulse** and **BSA** ($r = 0.465$).
12. A very weak positive linear relationship between **Stress** and **BSA** ($r = 0.018$).
13. A moderate positive linear relationship between **Pulse** and **DUR** ($r = 0.402$).
14. A weak to moderate positive linear relationship between **Stress** and **DUR** ($r = 0.312$).
15. A moderate positive linear relationship between **Stress** and **Pulse** ($r = 0.506$).

Nearly all relationships between the explanatory variables had r less than 0.7. However, in the relationship between **BSA** and **Weight,** r is greater than 0.7. Meaning they're highly correlated hence there are concerns for multicollinearity.

Full multiple-linear regression model:

Mathematical notation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon_i$$

Substituting the y and x's with their variable names:

$$Blood\ Pressure = \beta_0 + \beta_1 Age + \beta_2 Weight + \beta_3 BSA + \beta_4 Dur + \beta_5 Pulse + \beta_6 Stress + \epsilon_i$$

Deterministic part assumptions

Assumption 1. The parameters are linear in the deterministic part of the model.

- The assumption is valid since all regression coefficients in the suggested model are linear.

Assumption 2. The values of the explanatory variables are recorded without error.

- The data values appear to be valid, with no evident outliers. However, certainty cannot be ensured without knowledge of the precision of the measurement tools used during data collection.

Assumption 3. The explanatory variables are fixed in repeated samples.

- We assume the validity of this assumption.

Assumption 4. Reasonable variation in the values of the explanatory variables.

- This assumption is valid as the data values are different to each other.

Assumption 5. The sample size must be greater than the number of parameters to be estimated.

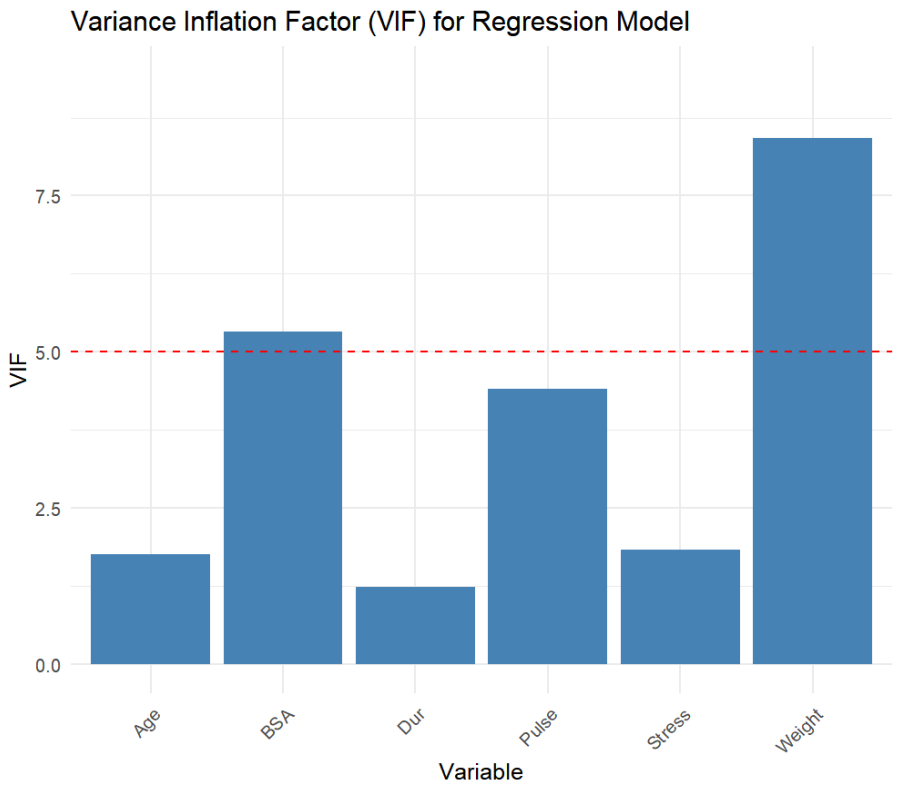
- Sample size (n) = 20 and parameters (p) = 6, meaning that$degrees\ of\ freedom = n - p = 20 - 6 = 14$.
  The assumption is valid since $n > p$.

Assumption 6. No multicollinearity between the explanatory variables in multiple regression models.

- Check multicollinearity using **Variance Inflation Factor (VIF)**.

R output:

```
--- Multicollinearity Check: Variance Inflation Factor ---
> vif_values <- vif(model_full)
> print(vif_values)
     Age   Weight      BSA      Dur    Pulse   Stress
1.762807 8.417035 5.328751 1.237309 4.413575 1.834845
```



The **VIF** values and bar chart show **BSA** and **Weight** having values **greater** than 5, which indicates potential multicollinearity.

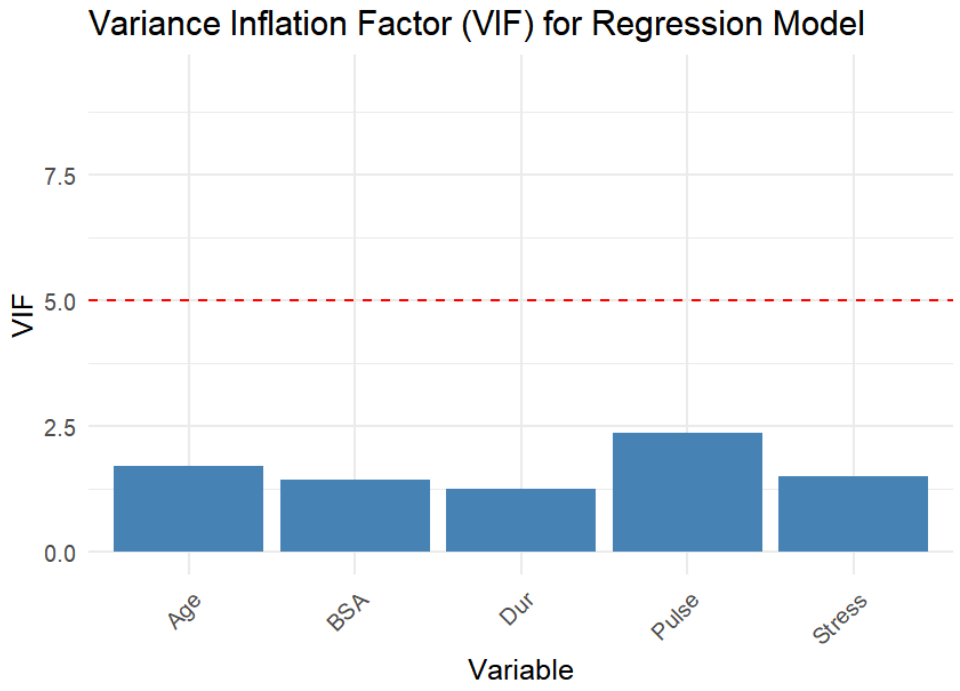
BSA formula: $BSA\ (m^2) = \sqrt{\frac{Weight\ (kg) * Height\ (cm)}{3600}}$

The formula shows **BSA** as a function of **Weight**, confirming that they're directly related. Thus, it is better to drop **Weight** and keep **BSA**. Likely, **BSA** is a better overall measure for studying blood pressure as it accounts for body size in a standardized way.

- Re-Checking multicollinearity using **Variance Inflation Factor (VIF)**.

R output:

```
--- Multicollinearity Check: Variance Inflation Factor ---
> vif_values <- vif(model_reduced)
> print(vif_values)
     Age      BSA      Dur    Pulse   Stress
1.703115 1.428349 1.237151 2.360939 1.502936
```



Dropping Weight has significantly reduced **BSA**'s **VIF** value. Multicollinearity assumption is now satisfied.

Model fitting

```
Call:
lm(formula = BP ~ Age + BSA + Dur + Pulse + Stress, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3687 -0.9135  0.1546  0.9053  2.8020

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.21215    9.42927   0.659   0.5207
Age          0.56297    0.20572   2.737   0.0161 *
BSA         24.55378    3.45160   7.114 5.22e-06 ***
Dur          0.07682    0.20437   0.376   0.7126
Pulse        0.45644    0.15925   2.866   0.0124 *
Stress      -0.01673    0.01303  -1.284   0.2199
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.718 on 14 degrees of freedom
Multiple R-squared:  0.9262,    Adjusted R-squared:  0.8998
F-statistic: 35.14 on 5 and 14 DF,  p-value: 1.921e-07
```

From R output:

- The residuals range from **-2.3687** to **2.8020**. The range suggests that there are some observations where the model predictions deviate significantly from the actual values.
- The explanatory variables **Age**, **BSA** and **Pulse** are statistically significant (**P-Value < 0.05**), i.e. their contribution to the model is higher than the other explanatory variables, while **Dur** and **Stress** are statistically insignificant (**P-Value > 0.05**) and can be removed from the model without losing much of its predictive power.
- The **Adjusted** R^2 value is **0.8998**, meaning that **89.98%** of the variation in **Blood Pressure** is explained by **Age**, **BSA**, **Dur**, **Pulse** and **Stress**.
- The **F-statistic** value is **35.14**. Since the value is large enough and P-Value is smaller than 0.05, then the entire model is statistically significant and is a good fit.

A.N.O.V.A

Analysis of variance is a hypothesis test used in regression analysis to study the effect of the regression coefficients.

$$H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$
$$H_1: At\ least\ one\ coefficient \neq 0$$

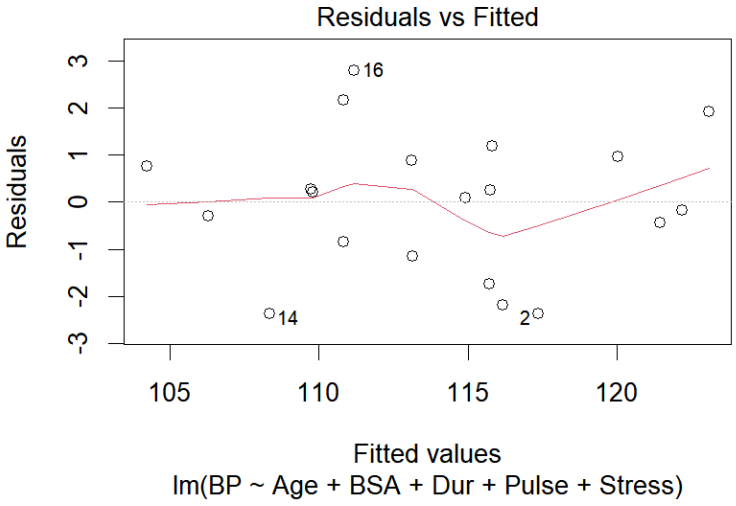
```
Analysis of Variance Table

Response: BP
          Df  Sum Sq Mean Sq F value    Pr(>F)
Age        1 243.266 243.266 82.4079 3.053e-07 ***
BSA        1 248.375 248.375 84.1385 2.692e-07 ***
Dur        1   2.761   2.761  0.9352   0.34992
Pulse      1  19.402  19.402  6.5726   0.02251 *
Stress     1   4.869   4.869  1.6493   0.21989
Residuals 14  41.328   2.952
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The **A.N.O.V.A** table confirms the statistical significance of the **Age**, **BSA** and **Pulse** variables, being the most important variables to the model. As well as the statistical significance of the model and it being a good fit.

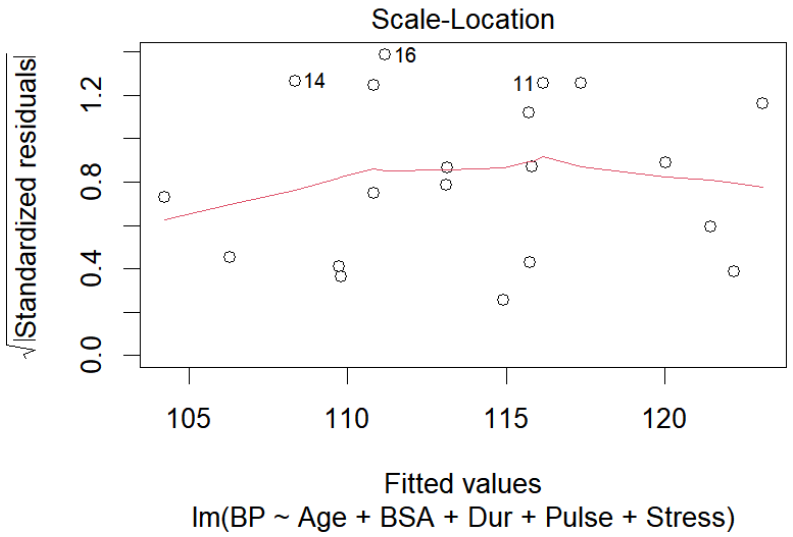
Decision: **Reject** the null hypothesis, the model is significant.

Random part assumptions

- Check that $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2$ for $i = 1,2,\ldots,20$



(1)



(2)

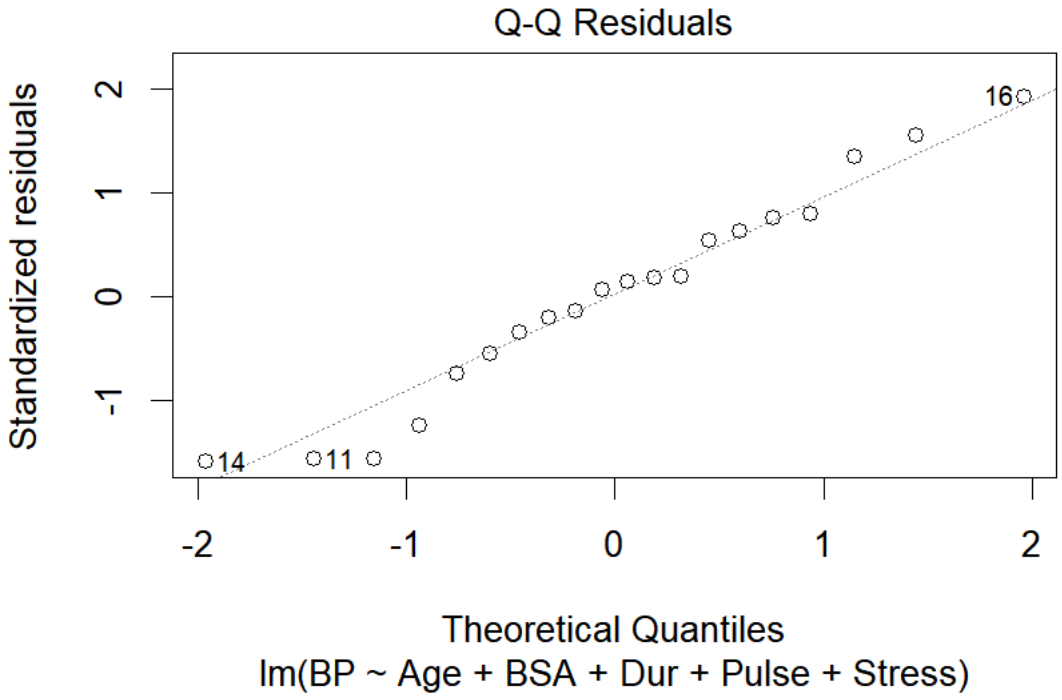
From plot (1): The data points seem to be randomly scattered around zero thus making $E(\epsilon_i) = 0$, variance seems to be constant across the residuals.
Meaning that the deterministic part of the model captures the non-random structure in the data and the errors scale of variability are constant at all values of the covariate.

From plot (2): The plot is used to have a better look at the homoscedasticity assumption. Here, it shows a relatively flat horizontal line. Meaning that the homoscedasticity is likely satisfied, and the variance is constant, confirming plot (1)'s assumption.

- Check the assumption of independent errors, i.e. $Cov(\epsilon_i, \epsilon_j) = E(\epsilon_i \epsilon_j) = 0$, for $i \neq j$.

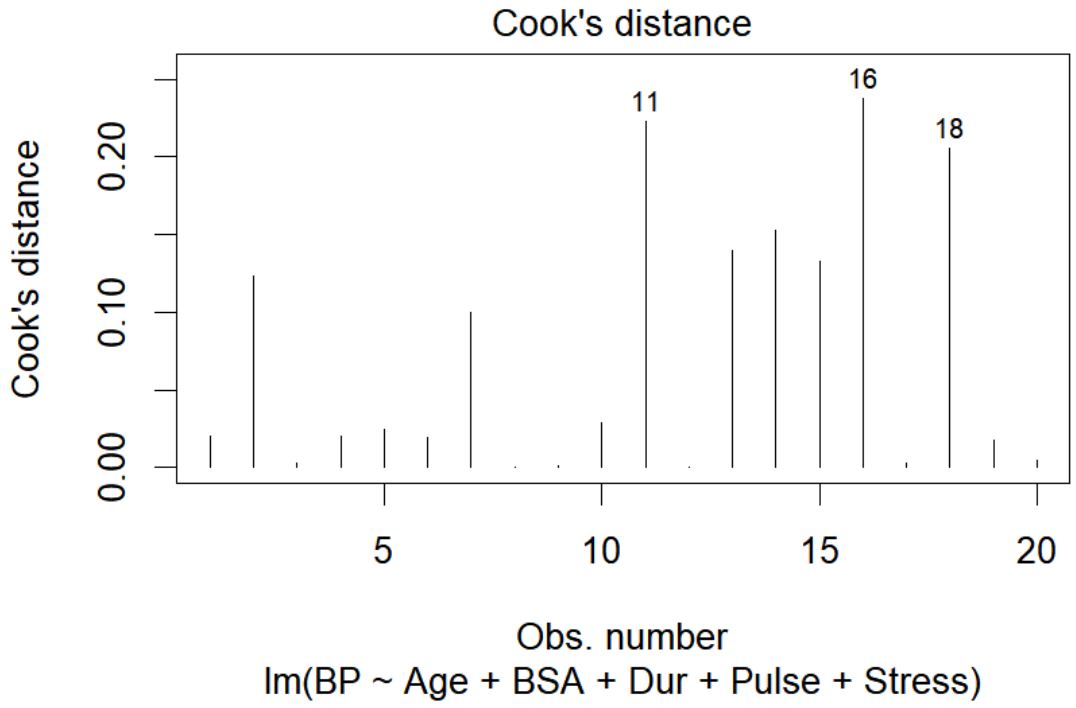
In this assumption, there is no intuitive natural order that we know about in the explanatory variable. So, we can assume independence between the errors.

- Check that $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1,2,\ldots,20$, i.e. The normality of the error term.



From the plot: There are deviations from the equity line, meaning that the normality assumption is most likely invalidated.

- Checking Cook’s distance for influential points or outliers.



From the plot: The points 11, 16 and 18 are potential outliers that need to be addressed.

Assumptions Conclusions

After checking the assumptions, it appears that a data transformation is needed to alleviate the normality of the error term and the outliers.

The most common transformation is the **Log transformation**.

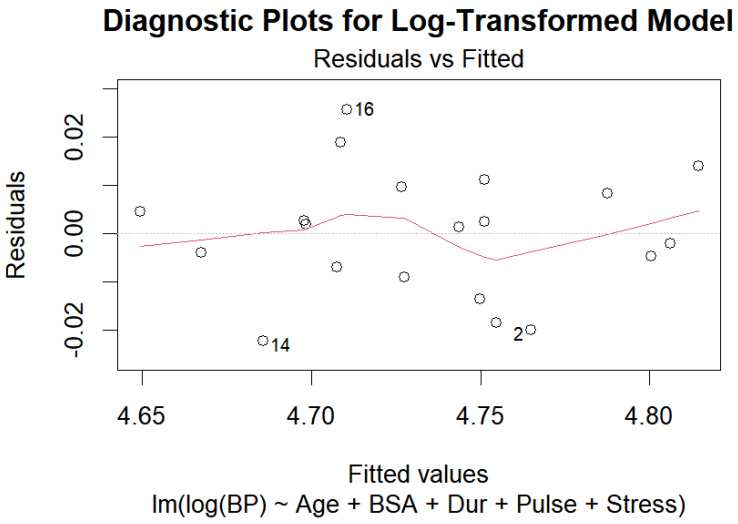
Model and Assumptions after Log-Transformation

```
Call:
lm(formula = log(BP) ~ Age + BSA + Dur + Pulse + Stress, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.022232 -0.007464  0.001687  0.008602  0.025774

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.7897406  0.0820264  46.201  < 2e-16 ***
Age          0.0047987  0.0017896   2.682  0.01789 *
BSA          0.2132823  0.0300259   7.103 5.31e-06 ***
Dur          0.0006456  0.0017778   0.363  0.72190
Pulse        0.0041698  0.0013854   3.010  0.00937 **
Stress      -0.0001563  0.0001133  -1.379  0.18955
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01495 on 14 degrees of freedom
Multiple R-squared:  0.927,     Adjusted R-squared:  0.901
F-statistic: 35.58 on 5 and 14 DF,  p-value: 1.775e-07
```

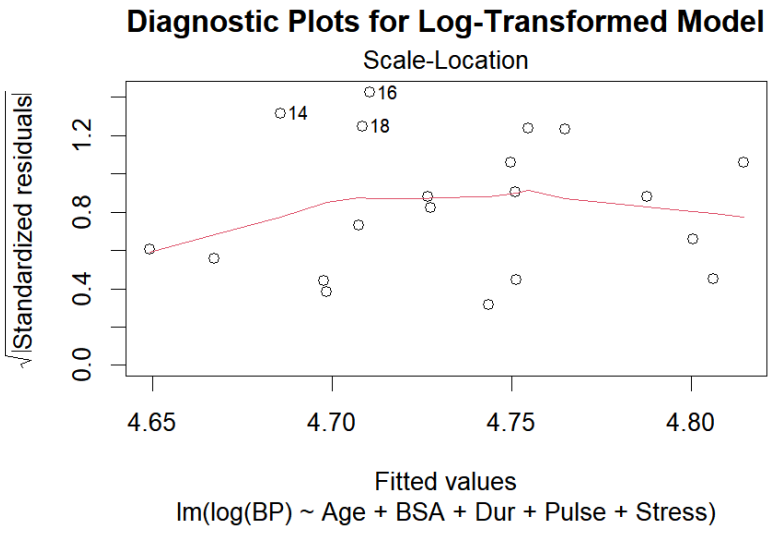
From R output:

- The residuals range has changed to: from **-0.022232** to **0.025774**, and this range is relatively small. This suggests that the model's predictions are quite accurate, with the largest error being only about **0.0257** units.
- The **multiple** R^2 value is **0.901**, which is slightly better than the normal model’s value **(0.8998).** However, the value is still neighboring **1**, thus making **90.1%** of the variation in **Blood Pressure** is explained by **Age**, **BSA**, **Dur**, **Pulse** and **Stress**.
- The statistical significance of the variables has not changed from the full model.
- The **F-statistic** value is **35.5**. Since the value is still large enough with a small P-Value, then the entire model is statistically significant and is a good fit.

- Check that $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2$ for $i = 1,2,\ldots,20$



(1)

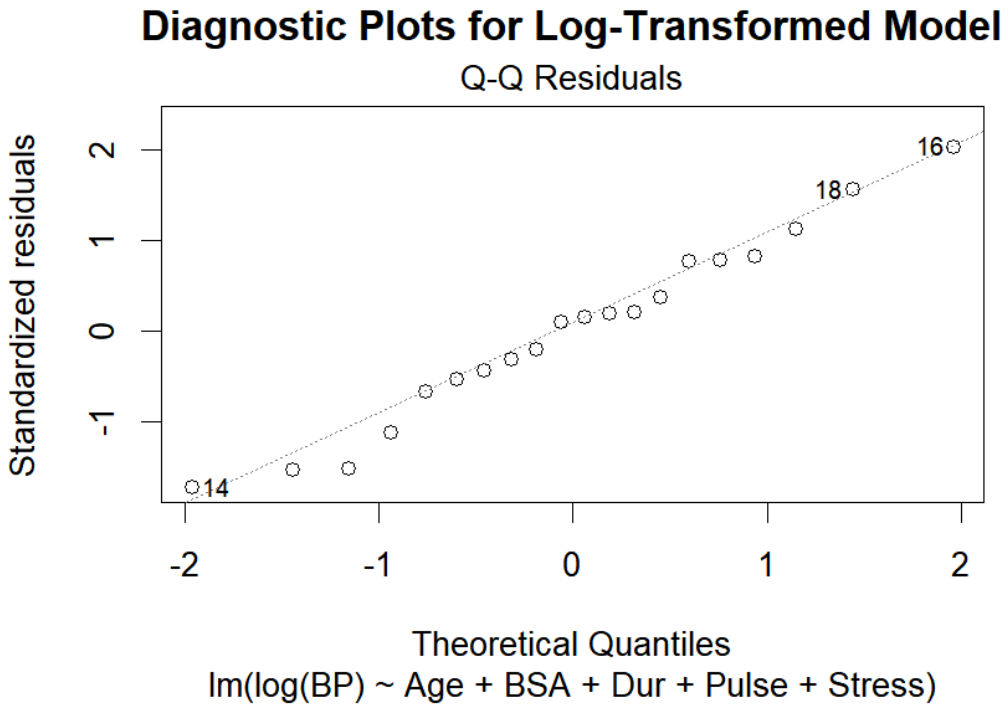


(2)

From plot (1): The data points seem to be still randomly scattered around zero and $E(\epsilon_i) = 0$ and variance of the error term is constant across the residuals.
Meaning that the deterministic part of the model still captures the non-random structure in the data and the errors scale of variability are constant at all values of the covariate, even after the log-transformation.

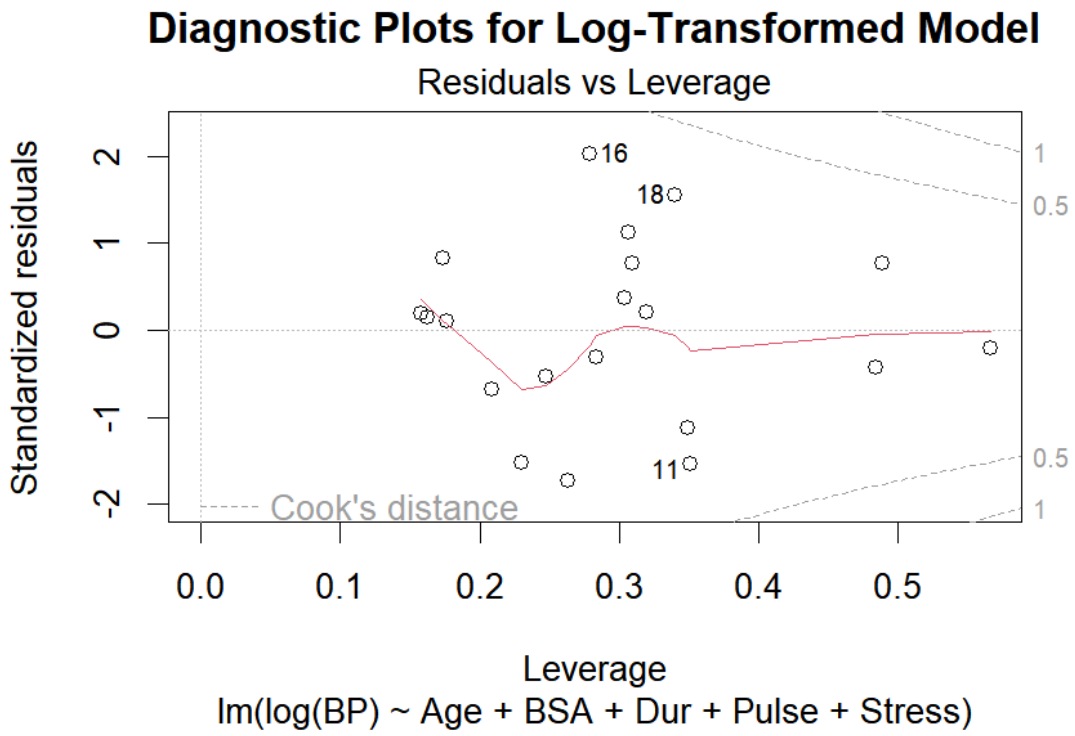
From plot (2): The horizontal line is still relatively flat; thus, the homoscedasticity assumption is likely satisfied.

- Check that $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1,2,\ldots,20$, i.e. The normality of the error term.



From the plot: The normality seems to have improved due to the log-transformation, since the deviations from the equity line have lessened.

- Check for outliers or influential points.



From the plot: There are no influential points, since the points do not fall on the dashed lines (which represent Cook's distance).

Conclusion

After conducting the regression analysis, the best fitted model is:

$$Log(\widehat{Blood\ Pressure}) = \hat{\beta}_0 + \hat{\beta}_1 Age + \hat{\beta}_2 BSA + \hat{\beta}_3 Dur + \hat{\beta}_4 Pulse + \hat{\beta}_5 Stress$$

Substituting $\hat{\beta}_i$ for their estimated values (for $i = 1,2,\ldots,20$):

$$Log(\widehat{Blood\ Pressure}) = 3.7897406 + (0.0047987)Age + (0.2132823)BSA + (0.0006456)Dur + (0.0041698)Pulse - (0.0001563)Stress$$

Model Interpretation

Intercept (3.7897406):

- This is the estimated log of blood pressure when all the predictors (Age, BSA, Dur, Pulse, and Stress) are equal to zero. Although interpreting the intercept alone might not always be meaningful in a real-world context (since age, pulse, and stress are unlikely to be zero), it's necessary for the model calculation.

Age (0.0047987):

- For each 1-year increase in age, the average log-transformed blood pressure is expected to increase by 0.47987%, holding all other variables constant.
- In practical terms, this suggests that older individuals tend to have slightly higher log-transformed blood pressure values, but the effect is relatively small.

BSA (0.2132823):

- For each 1-unit increase in Body Surface Area (BSA), the average log-transformed blood pressure is expected to increase by 21.32823%, holding all other variables constant.
- This suggests a positive association between BSA and blood pressure, meaning larger individuals (in terms of body surface area) tend to have higher blood pressure values.

Dur (0.0006456):

- For each 1-unit increase in Dur (perhaps duration of some activity or condition), the average log-transformed blood pressure increases by 0.06456%, holding other variables constant.

- This suggests a very slight positive relationship between duration and blood pressure, though the effect is quite small.

Pulse (0.0041698):

- For each 1-unit increase in pulse rate, the average log-transformed blood pressure increases by 0.41698%, holding other variables constant.
- This indicates a positive relationship, meaning individuals with higher pulse rates tend to have slightly higher log-transformed blood pressure.

Stress (-0.0001563):

- For each 1-unit increase in stress, the average log-transformed blood pressure decreases by 0.01563%, holding other variables constant.
- Interestingly, this suggests that higher stress is associated with slightly lower log-transformed blood pressure, although this relationship is very weak.

List of references

1. professional, C. C. medical. (2024, October 4). What is blood pressure?. Cleveland Clinic. https://shorturl.at/1DBCJ .
2. Carver, N., Anderson, A. D., & Jamal, Z. (2023, April 23). Drug utilization review: Treatment & management: Point of care. StatPearls. https://shorturl.at/0GWrF.
3. Pulse | meaning - Cambridge learner's dictionary. (n.d.). https://shorturl.at/qbjVi .
4. World Health Organization. (2023, February 21). Stress. World Health Organization. https://shorturl.at/hZK7T

Appendix

The code used in the analysis:

```
# ---------------- Load Required Libraries ----------------
# Install required packages if they are not already installed
required_packages <- c("readr", "ggplot2", "car", "MASS", "lmtest", "GGally")
missing_packages <- required_packages[!(required_packages %in%
installed.packages()[, "Package"])]
if (length(missing_packages)) install.packages(missing_packages)

# Load libraries
library(readr)   # For reading the data
library(ggplot2) # For data visualization
library(car)     # For diagnostic checks
library(MASS)    # For stepwise variable selection
library(lmtest)  # For statistical tests (e.g., Breusch-Pagan test)
library(GGally)  # For correlation pair plots

# ---------------- Load and Inspect Data ----------------
# Define the file path for the dataset
file_path <- "C:/Users/MSI/Downloads/bloodpress.txt"

# Load the data from a tab-delimited file
data <- read_delim(file_path, delim = "\t")

# Display the first few rows of the data
cat("\n--- First Few Rows of the Data ---\n")
print(head(data))

# Visualize correlation matrix
ggpairs(data)

# ---------------- Diagnostic Checks ----------------
# Fit the full model
model_full <- lm(BP ~ Age + Weight + BSA + Dur + Pulse + Stress, data = data)
```

```r
# Multicollinearity Check: Variance Inflation Factor (VIF)
cat("\n--- Multicollinearity Check: Variance Inflation Factor ---\n")
vif_values <- vif(model_full)
print(vif_values)

cor(data$Weight, data$BSA)

# Convert VIF values to a data frame for plotting
vif_df <- data.frame(Variable = names(vif_values), VIF = vif_values)

# Set a threshold to indicate high VIF
high_vif_threshold <- 5

# Create a ggplot bar plot to visualize VIF values
ggplot(vif_df, aes(x = Variable, y = VIF)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_hline(yintercept = high_vif_threshold, linetype = "dashed", color = "red")
+
  scale_y_continuous(limits = c(0, max(vif_df$VIF) + 1)) +
  labs(title = "Variance Inflation Factor (VIF) for Regression Model",
       y = "VIF",
       x = "Variable") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# ---------------- Reduced Model ----------------
# Fit a reduced model
model_reduced <- lm(BP ~ Age + BSA + Dur + Pulse + Stress, data = data)

# Display the summary of the reduced model
print(summary(model_log))
summary(model_reduced)

# ---------------- Reduced Model Diagnostic Plots ----------------
# 1. Linearity Check: Residuals vs Fitted values
plot(model_reduced, which = 1)
```

```r
# 2. Normal Q-Q Plot
plot(model_reduced, which = 2)  # Normal Q-Q

# 3. Scale-Location Plot
plot(model_reduced, which = 3)  # Scale-Location

# 4. Cook's Distance Plot
plot(model_reduced, which = 4)  # Cook's Distance

# 5. Multicollinearity Check: Variance Inflation Factor (VIF)
cat("\n--- Multicollinearity Check: Variance Inflation Factor ---\n")
vif_values <- vif(model_reduced)
print(vif_values)

# Convert VIF values to a data frame for plotting
vif_df2 <- data.frame(Variable = names(vif_values), VIF = vif_values)

# Create a ggplot bar plot to visualize VIF values
ggplot(vif_df2, aes(x = Variable, y = VIF)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_hline(yintercept = high_vif_threshold, linetype = "dashed", color = "red")
+
  scale_y_continuous(limits = c(0, max(vif_df$VIF) + 1)) +
  labs(title = "Variance Inflation Factor (VIF) for Regression Model",
       y = "VIF",
       x = "Variable") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# ---------------- Hypotheses Testing ----------------
# 1. Null hypothesis: All coefficients of predictors are zero (no effect).
# 2. Alternative hypothesis: At least one predictor has a non-zero effect.

# Perform an F-test for the overall significance of the model
anova_full <- anova(model_reduced)
cat("ANOVA Table:\n")
print(anova_full)
```

```r
cat(sprintf("F-statistic: %.2f, p-value: %.4f\n", anova_full$`F value`[1], anova_full$`Pr(>F)`[1]))
if (anova_full$`Pr(>F)`[1] < 0.05) {
  cat("Result: Reject the null hypothesis. The model is significant.\n")
} else {
  cat("Result: Failed to reject the null hypothesis. The model is not significant.\n")
}
# ---------------- Log-Transformed Reduced Model ----------------
cat("\n--- Log-Transformed Model ---\n")
# Fit a log-transformed model
model_log <- lm(log(BP) ~ Age + BSA + Dur + Pulse + Stress, data = data)

# Display the summary of the log-transformed model
print(summary(model_log))

# Diagnostic plots for the log-transformed model
plot(model_log, main = "Diagnostic Plots for Log-Transformed Model")
if (bp_test$p.value < 0.05) {
  cat("Homoscedasticity assumption violated. Log-transformed model may better satisfy assumptions.\n")
}

# ---------------- Final Model Diagnostic Plots ----------------
cat("\n--- Diagnostic Plots for Final Model ---\n")

# 1. Residuals vs Fitted Plot
plot(model_log, which = 1)  # Residuals vs Fitted

# 2. Normal Q-Q Plot
plot(model_log, which = 2)  # Normal Q-Q

# 3. Scale-Location Plot
plot(model_log, which = 3)  # Scale-Location

# 4. Cook's Distance Plot
plot(model_log, which = 4)  # Cook's Distance
```