# 24MDT0184_DA3

March 13, 2025

# 1 Name: Tufan Kundu

# 2 Registration No.: 24MDT0184

# 3 Regression Analysis and Predictive Models Lab

# 4 PMDS504P

# 5 Digital Assessment 3:Residual Analysis

## 5.1 Problem Statement

You are given a dataset containing various health-related variables for 20 individuals. Your task is to analyze the relationship between Diastolic Blood Pressure (BP) and other predictor variables using simple and multiple linear regression techniques.

### 5.1.1 Importing the necessary libraries

[2]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
```

### 5.1.2 Loading the dataset

[4]:
```python
df = pd.read_excel("bloodpress.xlsx")
df.head()
```

[4]:

| | Pt | BP | Age | Weight | BSA | Duration | Pulse | Stress |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 105 | 47 | 85.4 | 1.75 | 5.1 | 63 | 33 |
| 1 | 2 | 115 | 49 | 94.2 | 2.10 | 3.8 | 70 | 14 |
| 2 | 3 | 116 | 49 | 95.3 | 1.98 | 8.2 | 72 | 10 |
| 3 | 4 | 117 | 50 | 94.7 | 2.01 | 5.8 | 73 | 99 |
| 4 | 5 | 112 | 51 | 89.4 | 1.89 | 7.0 | 72 | 95 |

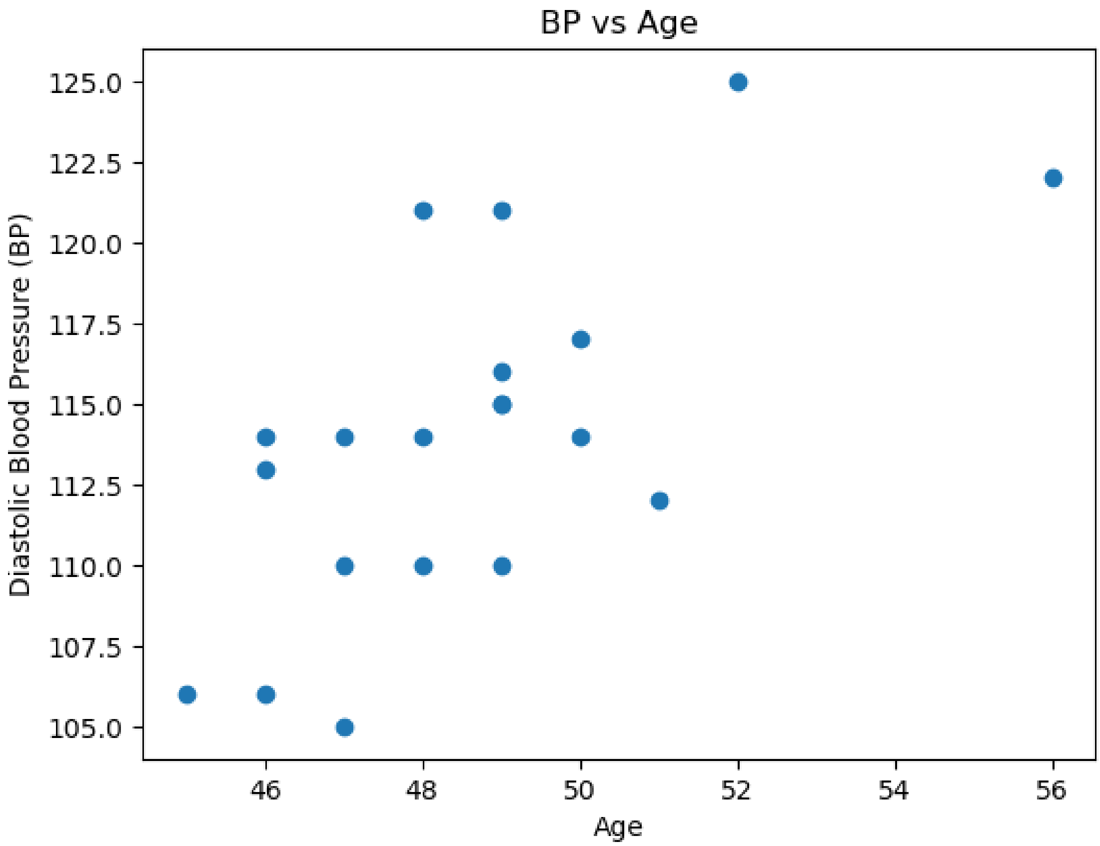
[5]:
```python
# Extract the variables
bp = df['BP'] # Response variable
age = df['Age']
```

```
weight = df['Weight']
duration = df['Duration']
```

### 5.2 Simple Linear Regression (Bp ~ Age)

[6]:
```
x_age = sm.add_constant(age)
model_age = sm.OLS(bp,x_age).fit()
print("Regression summary: BP vs Age\n")
print(model_age.summary())
plt.scatter(age,bp)
plt.xlabel("Age")
plt.ylabel("Diastolic Blood Pressure (BP)")
plt.title("BP vs Age")
plt.show()
resid_age = model_age.resid
```

```
Regression summary: BP vs Age

                            OLS Regression Results
==============================================================================
Dep. Variable:                     BP   R-squared:                       0.434
Model:                            OLS   Adj. R-squared:                  0.403
Method:                 Least Squares   F-statistic:                     13.82
Date:                Thu, 13 Mar 2025   Prob (F-statistic):            0.00157
Time:                        22:11:56   Log-Likelihood:                -56.002
No. Observations:                  20   AIC:                             116.0
Df Residuals:                      18   BIC:                             118.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         44.4545     18.728      2.374      0.029       5.109      83.800
Age            1.4310      0.385      3.718      0.002       0.622       2.240
==============================================================================
Omnibus:                        0.767   Durbin-Watson:                   1.965
Prob(Omnibus):                  0.682   Jarque-Bera (JB):                0.766
Skew:                           0.277   Prob(JB):                        0.682
Kurtosis:                       2.217   Cond. No.                         972.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

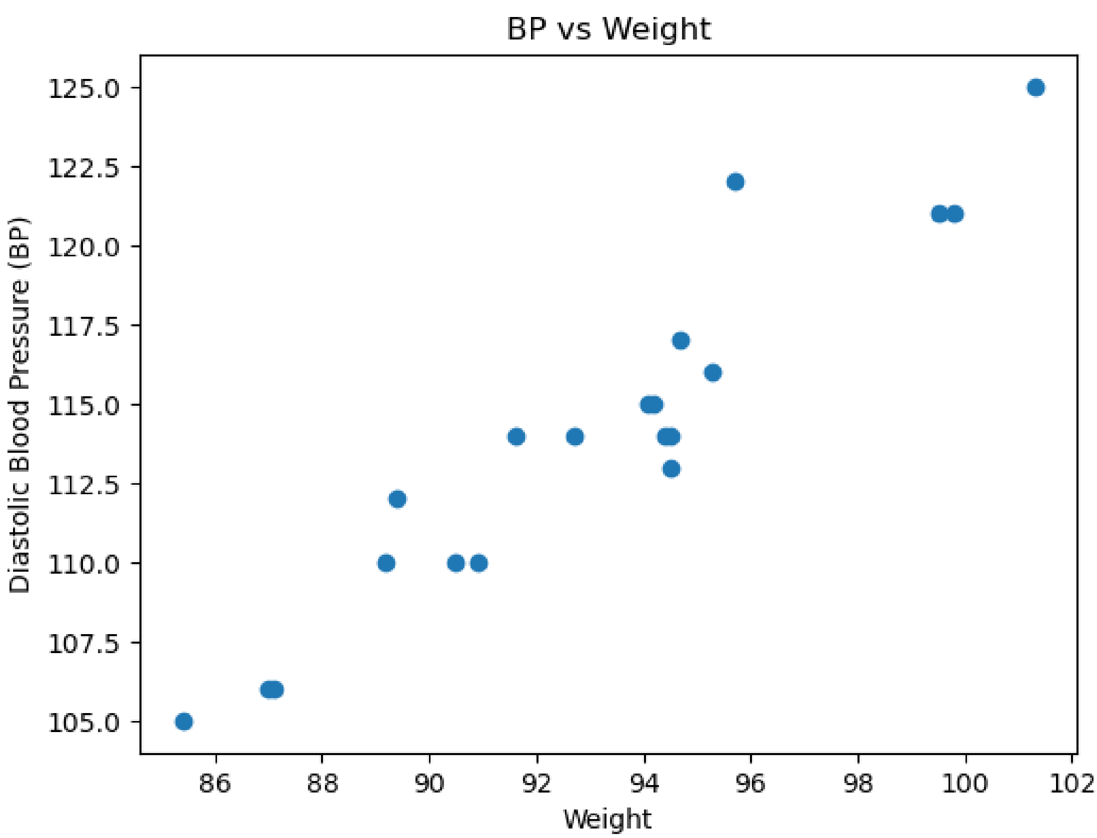



### 5.3 Simple Linear Regression (BP ~ Weight)

[10]:
```
x_weight = sm.add_constant(weight)
model_weight = sm.OLS(bp,x_weight).fit()
print("Regression Summary: BP vs Weight\n")
print(model_weight.summary())
plt.scatter(weight, bp)
plt.xlabel("Weight")
plt.ylabel("Diastolic Blood Pressure (BP)")
plt.title("BP vs Weight")
plt.show()
resid_weight = model_weight.resid
```

```
Regression Summary: BP vs Weight

                            OLS Regression Results
==============================================================================
Dep. Variable:                     BP   R-squared:                       0.903
Model:                            OLS   Adj. R-squared:                  0.897
Method:                 Least Squares   F-statistic:                     166.9
```

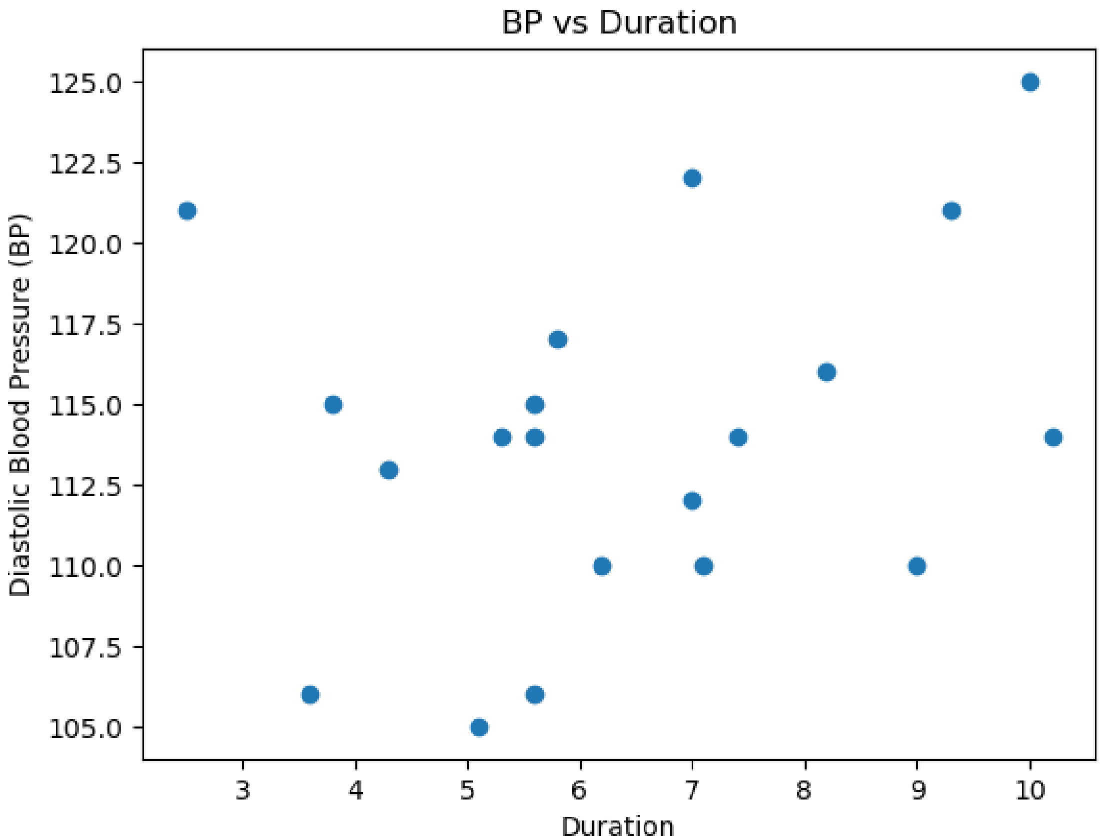
```
Date:                Thu, 13 Mar 2025   Prob (F-statistic):           1.53e-10
Time:                        22:16:43   Log-Likelihood:                -38.409
No. Observations:                  20   AIC:                             80.82
Df Residuals:                      18   BIC:                             82.81
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.2053      8.663      0.255      0.802     -15.996      20.406
Weight         1.2009      0.093     12.917      0.000       1.006       1.396
==============================================================================
Omnibus:                        9.231   Durbin-Watson:                   1.641
Prob(Omnibus):                  0.010   Jarque-Bera (JB):                6.566
Skew:                           1.157   Prob(JB):                       0.0375
Kurtosis:                       4.590   Cond. No.                     2.07e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 2.07e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```



### 5.4 Simple Linear Regression (BP ~ Duration)

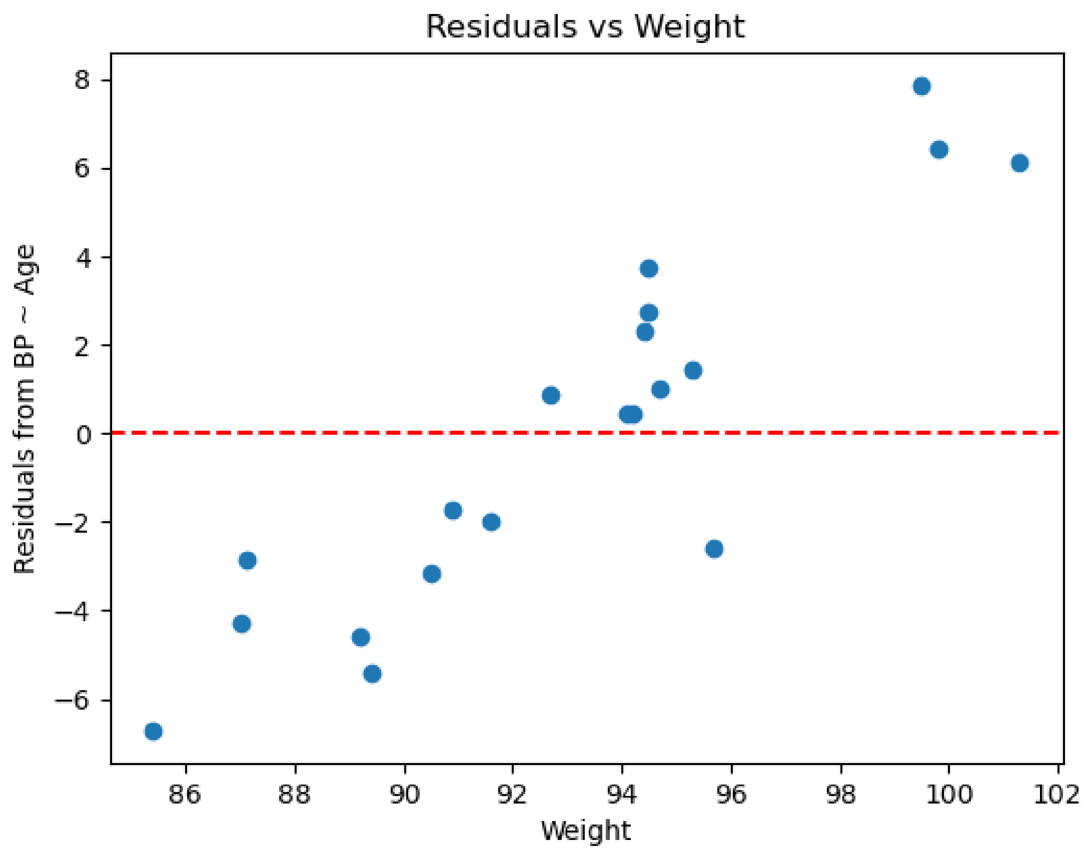
[12]:
```
x_duration = sm.add_constant(duration)
model_duration = sm.OLS(bp,x_duration).fit()
print("Regression Summary: BP vs Duration\n")
print(model_duration.summary())
plt.scatter(duration, bp)
plt.xlabel("Duration")
plt.ylabel("Diastolic Blood Pressure (BP)")
plt.title("BP vs Duration")
plt.show()
resid_duration = model_duration.resid
```

```
Regression Summary: BP vs Duration

                            OLS Regression Results
==============================================================================
Dep. Variable:                     BP   R-squared:                       0.086
Model:                            OLS   Adj. R-squared:                  0.035
Method:                 Least Squares   F-statistic:                     1.688
Date:                Thu, 13 Mar 2025   Prob (F-statistic):              0.210
Time:                        22:30:13   Log-Likelihood:                -60.804
No. Observations:                  20   AIC:                             125.6
Df Residuals:                      18   BIC:                             127.6
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        109.2350      3.856     28.327      0.000     101.133     117.337
Duration       0.7411      0.570      1.299      0.210      -0.457       1.939
==============================================================================
Omnibus:                        0.757   Durbin-Watson:                   2.199
Prob(Omnibus):                  0.685   Jarque-Bera (JB):                0.752
Skew:                           0.258   Prob(JB):                        0.687
Kurtosis:                       2.202   Cond. No.                         22.3
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```




### 5.5 Residual vs Weight plot (from BP ~ Age model)

[15]:
```
plt.scatter(weight,resid_age)
plt.axhline(y = 0, color = 'r', linestyle = '--')
plt.xlabel('Weight')
plt.ylabel('Residuals from BP ~ Age')
plt.title('Residuals vs Weight')
plt.show()
```




### 5.6 Multiple Linear Regression (BP ~ Age + weight)

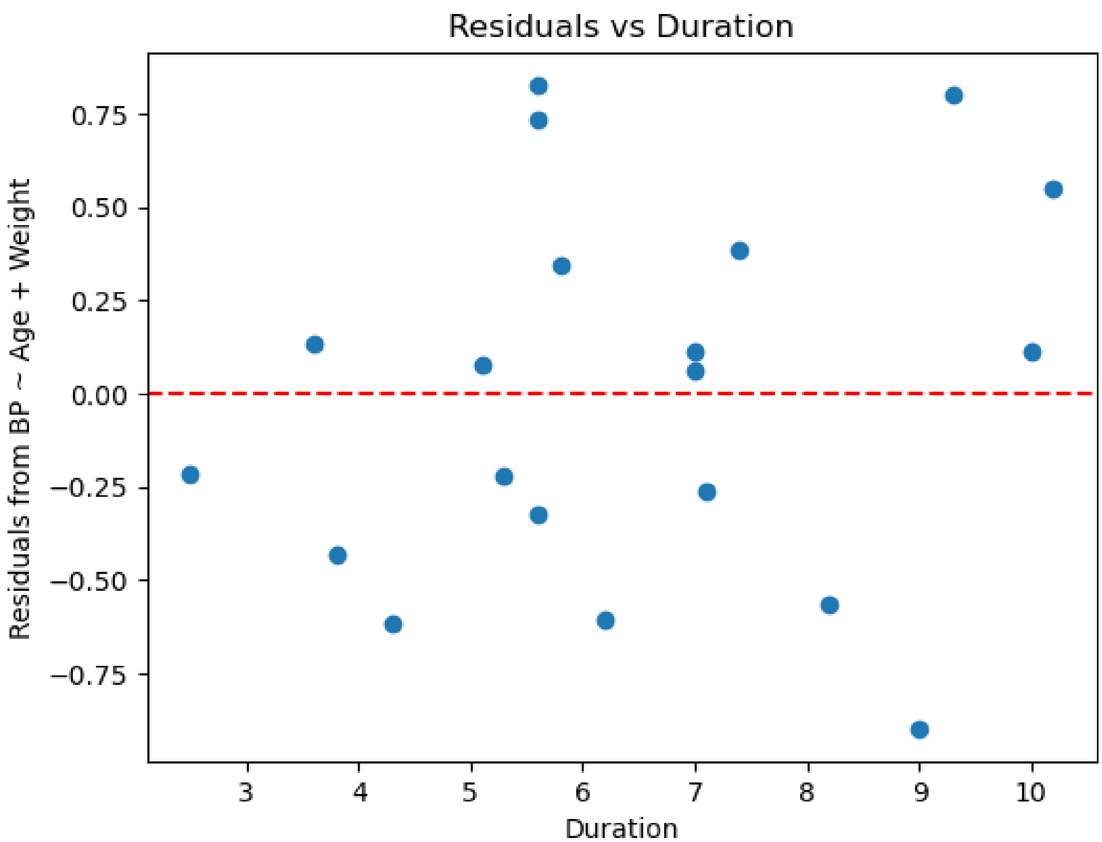
[17]:
```
x_multi = sm.add_constant(pd.DataFrame({'Age':age,'Weight':weight}))
model_multi = sm.OLS(bp,x_multi).fit()
print("Regression summary: BP vs Age + Weight\n")
print(model_multi.summary())
```

```
Regression summary: BP vs Age + Weight

                            OLS Regression Results
==============================================================================
Dep. Variable:                     BP   R-squared:                       0.991
Model:                            OLS   Adj. R-squared:                  0.990
Method:                 Least Squares   F-statistic:                     978.2
Date:                Thu, 13 Mar 2025   Prob (F-statistic):           2.81e-18
Time:                        22:35:38   Log-Likelihood:                -14.157
No. Observations:                  20   AIC:                             34.31
Df Residuals:                      17   BIC:                             37.30
Df Model:                           2
Covariance Type:            nonrobust
```

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -16.5794      3.007     -5.513      0.000     -22.925     -10.234
Age            0.7083      0.054     13.235      0.000       0.595       0.821
Weight         1.0330      0.031     33.154      0.000       0.967       1.099
==============================================================================
Omnibus:                        0.989   Durbin-Watson:                   1.688
Prob(Omnibus):                  0.610   Jarque-Bera (JB):                0.768
Skew:                           0.101   Prob(JB):                        0.681
Kurtosis:                       2.061   Cond. No.                     2.65e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 2.65e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

### 5.7 Residual vs Duration plot (from BP ~ Age + weight)

[19]:
```python
resid_multi = model_multi.resid
plt.scatter(duration, resid_multi)
plt.axhline(y = 0, color = 'r', linestyle = '--')
plt.xlabel("Duration")
plt.ylabel("Residuals from BP ~ Age + Weight")
plt.title("Residuals vs Duration")
plt.show()
```




# 6 Interpretation & Observations

## 6.1 1. Significance of Predictor Variables

- **BP ~ Age:** Significant (p = 0.002). BP increases by 1.43 per year.
- **BP ~ Weight:** Highly significant (p = 0.000). BP increases by 1.20 per kg.
- **BP ~ Duration:** Not significant (p = 0.210). Weak relationship.
- **BP ~ Age + Weight:** Both are significant (p = 0.000). Best model.

## 6.2 2. Goodness-of-Fit ($R^2$ Value)

- **BP ~ Age:** $R^2 = 0.434$ (Moderate)
- **BP ~ Weight:** $R^2 = 0.903$ (Strong)
- **BP ~ Duration:** $R^2 = 0.086$ (Weak)
- **BP ~ Age + Weight:** $R^2 = 0.991$ (Best)

## 6.3 3. Residual Behavior and Model Improvements

- **BP ~ Age model:** Residuals plotted against Weight show a pattern, suggesting that including Weight as a predictor can improve the model.

- **BP ~ Age + Weight model:** Residuals plotted against Duration do not show a pattern, confirming that the model fits well.
- **This model can be improved by:**
  1. **Checking for multicollinearity,** as the high condition number suggests possible correlation issues.
  2. **Adding more predictors** such as BSA, Pulse, or Stress, if they contribute valuable information.
  3. **Exploring polynomial or interaction terms** to capture non-linear relationships if present.
  4. **Increasing dataset size,** which can enhance model generalization and robustness.

## 6.4 Conclusion

BP is mainly influenced by Weight and Age. The best model is BP ~ Age + Weight ($R^2$ = 99.1%). Duration does not contribute significantly. This model can be refined further by checking assumptions, addressing multicollinearity, and considering additional predictors.