Foundation of Machine Learning [CS 555]

Project Report [Group 7]

BOSTON UNIVERSITY

Proposed By:

Pratyush Patel

Shreya Asoba

Varrsha Ramanna Kumar

Table of Contents

Table of Figures

1. Research Scenario:

The research aims to explore the relationship between breast cancer characteristics in patients, such as Resistin, Adiponectin, and Age and their Body Mass Index. By utilizing linear and multi regression models, the study seeks to identify significant predictors and assess the overall fit of the models. This analysis is conducted on a sampled dataset of 116 rows.

2. Research Questions:

- Is there a significant linear relationship between Resistin and Body Mass Index in individual with breast cancer?
- Is there a significant relationship between the Adiponectin levels and Body Mass Index of the patients?
- Is there a significant relationship between the Resistin, Adiponectin and Age with Body Mass Index of the patients?
- Identifying trends and outliers.

3. Data Set Description and Summary:

Dataset Variables:

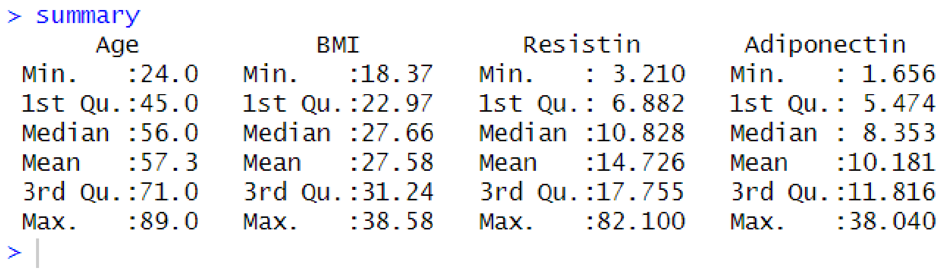
- **Age:** The age of the individuals in the dataset.
- **BMI:** Body Mass Index, a measure of body fat based on height and weight.
- **Resistin:** A biomarker associated with insulin resistance.
- **Adiponectin:** A protein hormone involved in regulating glucose levels.

Data Cleaning:

- Removed any unnecessary columns for this analysis.
- Checked for missing values.

Summary Statistics:

For each variable (Age, BMI, Resistin, Adiponectin), the table provides the following summary statistics:



```
> summary
      Age             BMI             Resistin         Adiponectin
 Min.   :24.0   Min.   :18.37   Min.   : 3.210   Min.   : 1.656
 1st Qu.:45.0   1st Qu.:22.97   1st Qu.: 6.882   1st Qu.: 5.474
 Median :56.0   Median :27.66   Median :10.828   Median : 8.353
 Mean   :57.3   Mean   :27.58   Mean   :14.726   Mean   :10.181
 3rd Qu.:71.0   3rd Qu.:31.24   3rd Qu.:17.755   3rd Qu.:11.816
 Max.   :89.0   Max.   :38.58   Max.   :82.100   Max.   :38.040
> 
```

Fig.1 Summary of Dataset

4. Statistical Methods:

- Simple Linear Regression Model
- Multiple Linear Regression Model

5. Result and Evaluations:

1) Linear Regression Model - Understanding the association between Resistin levels and BMI of the patients.

BMI = B_0 +B_1 X_{hat}

Linear regression results for the impact of resistinvels on the BMI individually.
Residual plots for the calculated results based on resistin levels.

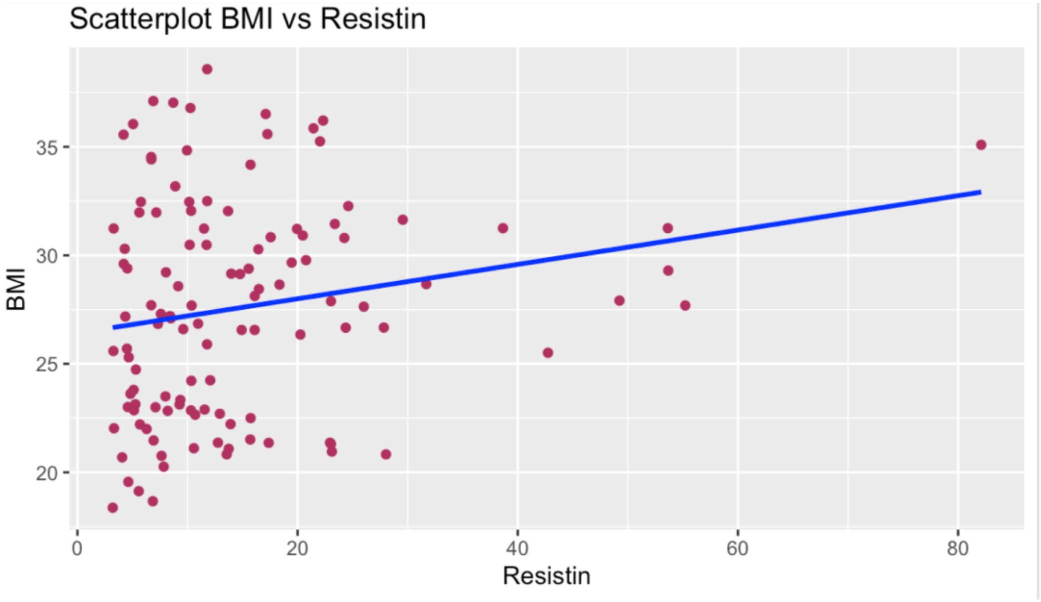
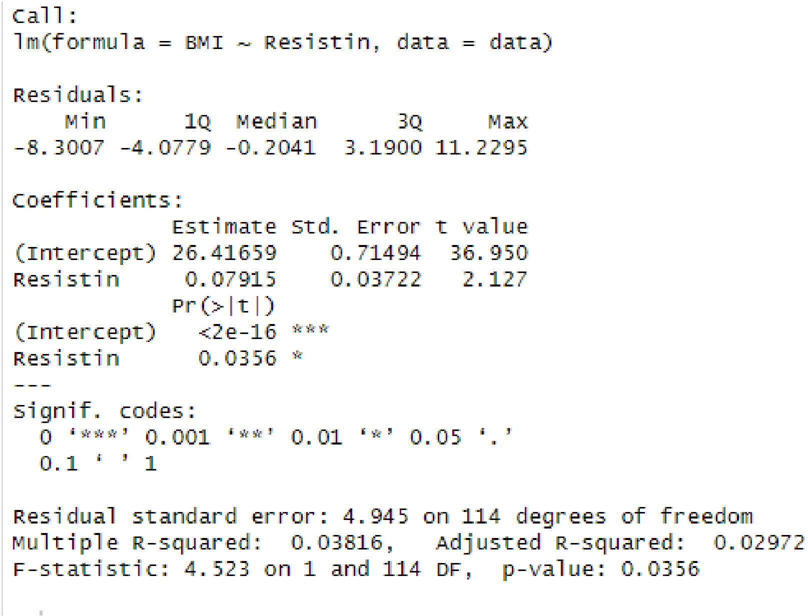


Fig.2 Scatterplot of BMI and Resistin with Regression line

Form – It's a linear form as the data points maintain a straight-line pattern.

Direction – Clearly as the levels increase, even the BMI increase. Hence, it's a positively associated direction.

Strength of Association – This factor can be described by checking how close the data points are associated with each other. Clearly the Data points are strongly associated as they are close and concentrated in the plot.



```
Call:
lm(formula = BMI ~ Resistin, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.3007 -4.0779 -0.2041  3.1900 11.2295 

Coefficients:
            Estimate Std. Error t value
(Intercept) 26.41659    0.71494  36.950
Resistin     0.07915    0.03722   2.127
            Pr(>|t|)    
(Intercept)   <2e-16 ***
Resistin      0.0356 *  
---
Signif. codes:  
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
  0.1 ‘ ’ 1

Residual standard error: 4.945 on 114 degrees of freedom
Multiple R-squared:  0.03816,	Adjusted R-squared:  0.02972 
F-statistic: 4.523 on 1 and 114 DF,  p-value: 0.0356
```

Fig2: Coefficient Summary

a) Overall Model Significance:

The p-value 0.03 is less than 0.05, indicating that Resistin is statistically significant in predicting BMI in this model.

This indicates that, on average, for each one-unit increase in Resistin, the BMI is expected to increase by 0.07915 units.

The overall model is statistically significant (F-statistic: 4.52). This suggests that Resistin is a significant predictor of BMI in this specific model.

R-squared: The multiple R-squared is 0.03816. This value represents the proportion of the variance in BMI that is explained by the model. In this case, only about 3.82% of the variability in BMI is explained by the linear relationship with Resistin.

Model equation: BMI = 26.41 + 0.07R

b) Confidence Intervals for Coefficients:

The confint function provides the 95% confidence intervals for each coefficient in the linear regression model.

We are 95% confident that the true effect of Resistin on BMI (the true change in BMI for each one-unit increase in Resistin) lies between 0.0054 and 0.152

c) <u>Residual Plots:</u> Residual plots for the calculated results based on Resistin.

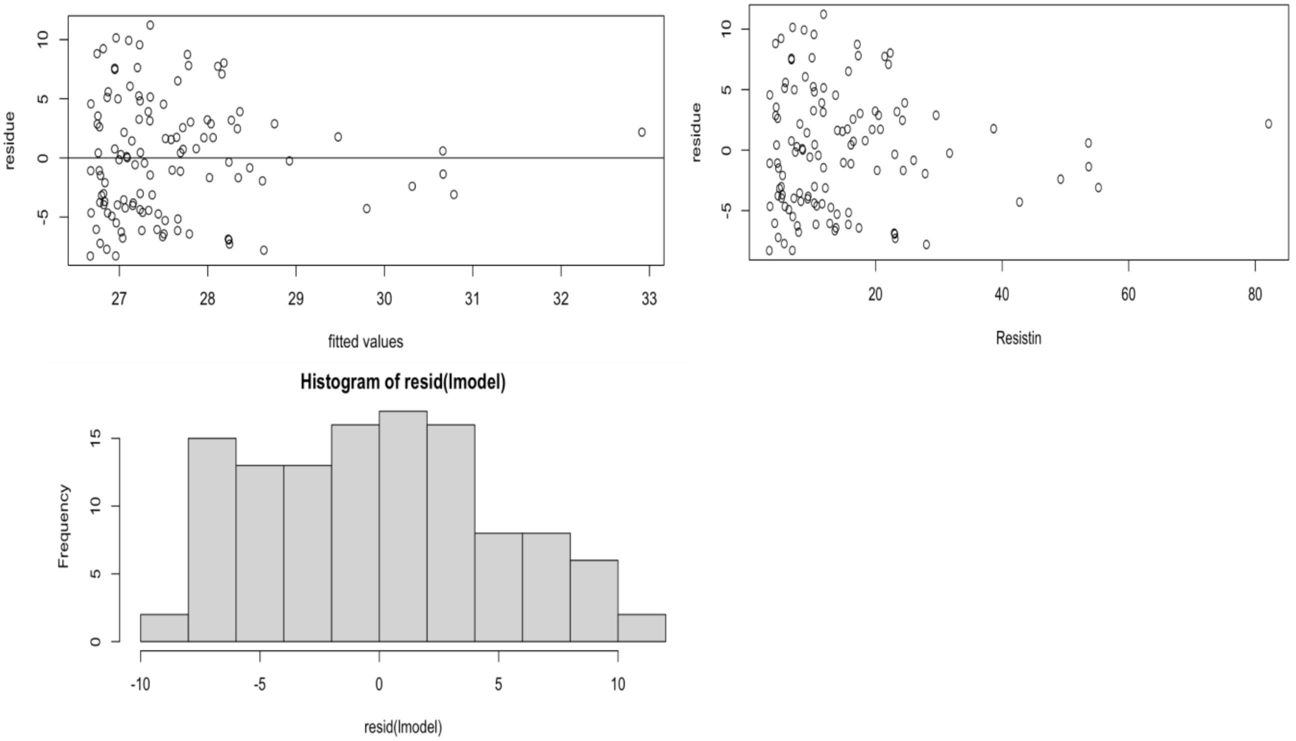


Fig.3 Residual Plots of BMI and Resistin

d) <u>Residual Plot Interpretation:</u>

<u>Linearity:</u> The residuals are randomly scattered around zero thus the variables are linear.

<u>Homoscedasticity (Constant Variance):</u> The spread is consistent as points moving along the x-axis, hence its Homoroscedasticity.

<u>Normality:</u> The histogram of residuals can be used to assess normality. It is right skewed normally distributed.

2) Linear Regression Model - Understanding the association between Adiponectin levels and BMI of the patients.

BMI = B_0 +B_1 X_{hat}

Linear regression results for the impact of Adiponectin levels on the BMI of pateints individually.

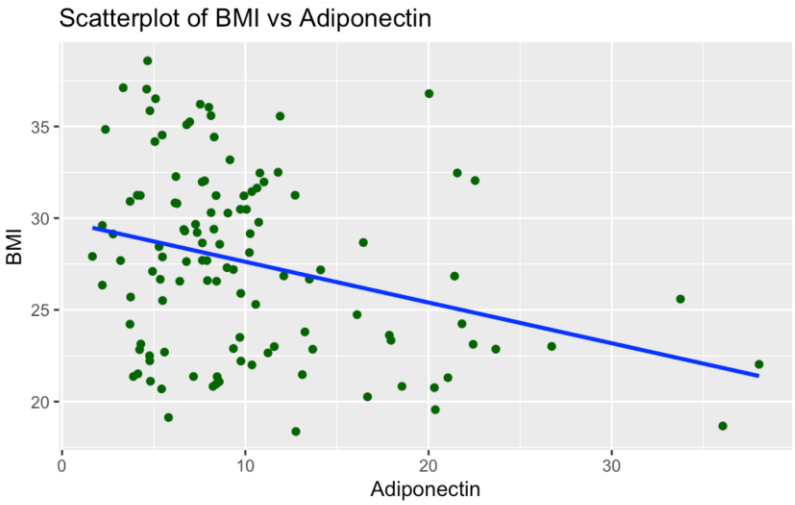


Fig.5 Scatterplot of Adiponectin and BMI with Regression line

The analysis of the relationship between BMI and Adiponectin has produced the following insights:

Form – It's a linear form as the data points maintain a straight-line pattern.

Direction – Clearly as the Adiponectin levels increase, even the BMI decreases. Hence, it's a negatively associated direction.

Strength of Association – This factor can be described by checking how close the data points are associated with each other. Clearly the Data points are more strongly associated as they are close and concentrated in the plot.

a) Overall Model Significance:

The p-value 0.000956 is less than 0.05, indicating that Adiponectin is statistically significant in predicting BMI in this model.
This indicates that, on average, for each one-unit increase in Adiponectin, the BMI is expected to decrease by 0.22208 units.

The overall model is statistically significant (F-statistic: 11.5). This suggests that Adiponectin is a significant predictor of BMI in this specific model.

R-squared: The multiple R-squared is 0.09165. This value represents the proportion of the variance in BMI that is explained by the model. In this case, only about 9.17% of the variability in BMI is explained by the linear relationship with Adiponectin.

Model equation: **BMI = 29.84 - 0.22Ad**

b) Confidence Intervals for Coefficients:

The confint function provides the 95% confidence intervals for each coefficient in the linear regression model.

We are 95% confident that the true effect of Adiponectin on BMI (the true change in BMI for each one-unit increase in Adiponectin) lies between -0.3517 and -0.0923

c) Residual Plots- Residual plots for the calculated results based on Adiponectin levels.

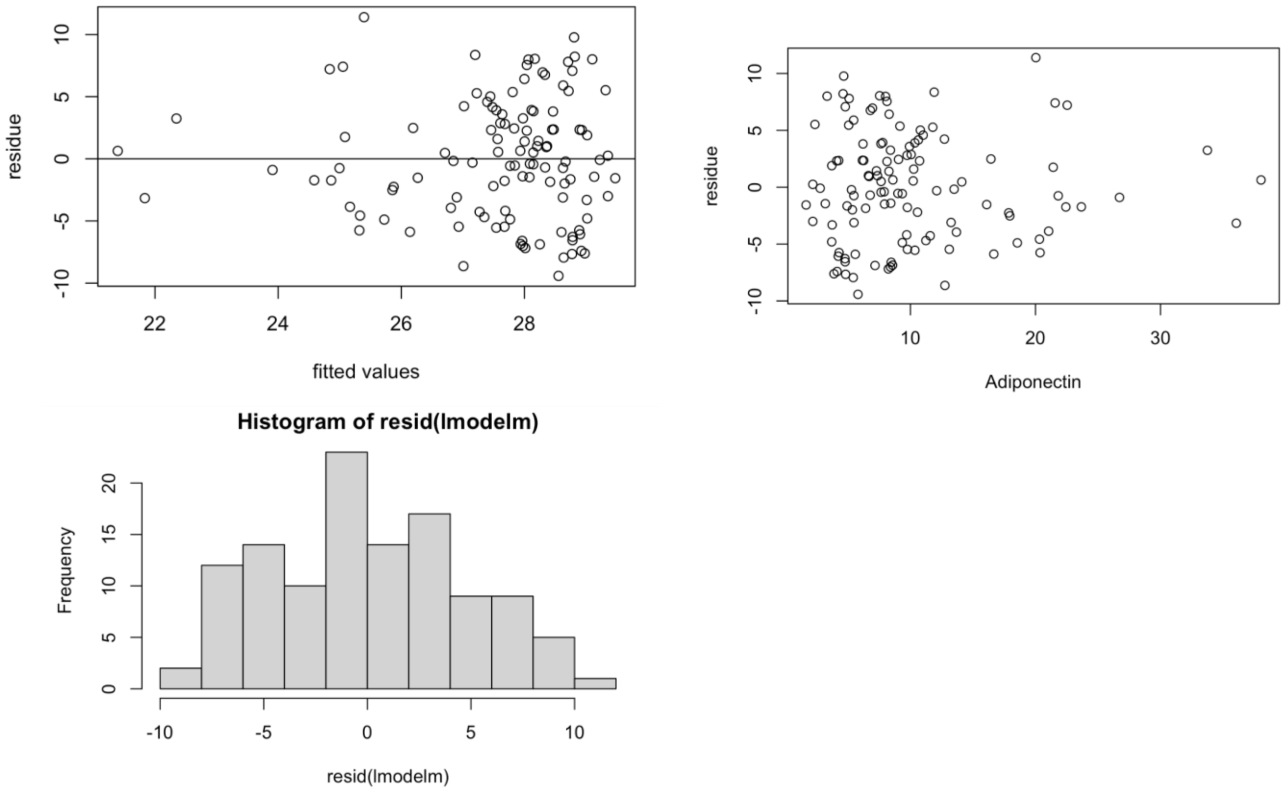


Fig.6 Residual Plots of BMI and Adiponectin levels

d) Residual Plot Interpretation-

Linearity: The residuals are randomly scattered around zero thus the variables are linear.

Homoscedasticity (Constant Variance): The spread is consistent as points moving along the x-axis, hence its Homoscedasticity.

Normality: Its seems bell structured thus normally distributed.

The largest studentized residual is 4.213553, corresponding to observation 1000. The unadjusted p-value for this observation is 2.7408e-05. This suggests that the observation 1000 is statistically significant as an outlier (p-value < 0.05). With the Bonferroni correction, the p-value remains below 0.05, indicating that this observation is still considered a potential outlier after correcting for multiple testing. The Bonferroni-corrected p-value is 0.0276.

3) Multi Regression Model - Understanding the association between Resistin, Adiponectin levels and Age of patients on their BMI.

$Y = B_0 + B_1X_1 \ldots B_kX_k$

Multi regression results for the impact of Resisitin, Adiponectin and age of patients on their BMI altogether.

a) Statistical Significance:

On average, for each one-unit increase in Resistin, the BMI is expected to increase by 0.05024 units. However, the p-value for Resistin is 0.18182, indicating that the effect of Resistin is not statistically significant at the 0.05 significance level.

On average, for each one-unit increase in Adiponectin, the BMI is expected to decrease by 0.20785 units. The p-value for Adiponectin is 0.00338, suggesting that the effect of Adiponectin is statistically significant at the 0.05 significance level.

On average, for each one-unit increase in Age, the BMI is expected to decrease by 0.01685 units. However, the p-value for Age is 0.55580, indicating that the effect of Age is not statistically significant at the 0.05 significance level.

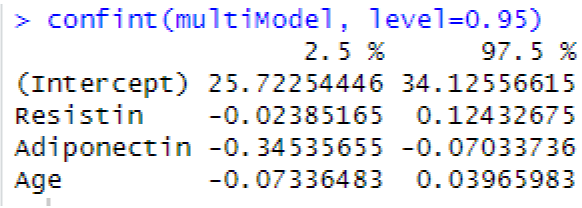
The p-value 0.004528 is less than 0.05 and F-statistic: 4.59 indicating that over all model is statistically significant in predicting BMI in this model.

R-squared: The multiple R-squared is 0.1095. This value represents the proportion of the variability in BMI that is explained by the model with Resistin, Adiponectin, and Age.

Adjusted R-squared: The adjusted R-squared is 0.08568. It adjusts the R-squared for the number of predictors in the model.

Model equation: **BMI = 29.92 + 0.05R - 0.02Ad - 0.01Ag**

b) Confidence Intervals for Coefficients:



```
> confint(multiModel, level=0.95)
                   2.5 %      97.5 %
(Intercept) 25.72254446 34.12556615
Resistin    -0.02385165  0.12432675
Adiponectin -0.34535655 -0.07033736
Age         -0.07336483  0.03965983
```

The confint function provides the 95% confidence intervals for each coefficient in the multi regression model.
We are 95% confident that the true value of the intercept lies between 25.72 and 34.12.
We are 95% confident that the true effect of Resistin on BMI lies between -0.0238 and 0.1243
We are 95% confident that the true effect of Adiponectin on BMI lies between -0.3453 and -0.0703
We are 95% confident that the true effect of Adiponectin on BMI lies between -0.0733 and 0.0396

c) **Residual plots:** Residual plots for the calculated results based on Glucose, MCP1 levels and Age of patients.

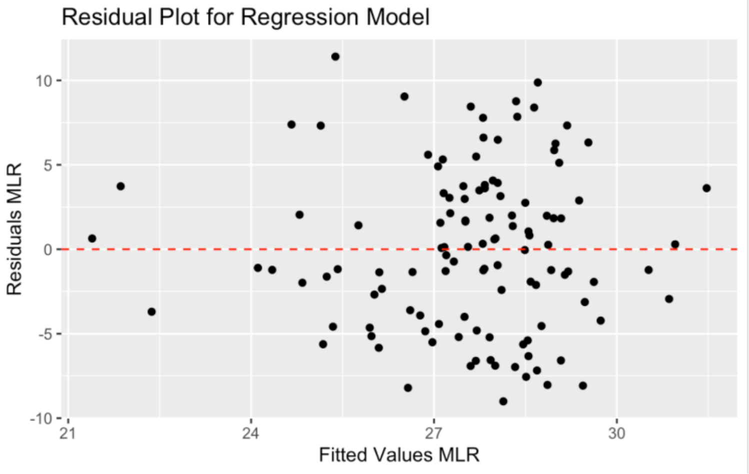
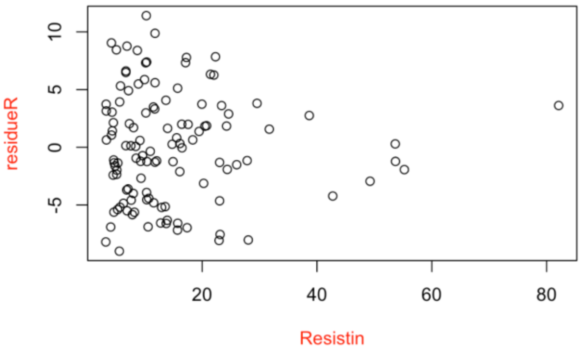
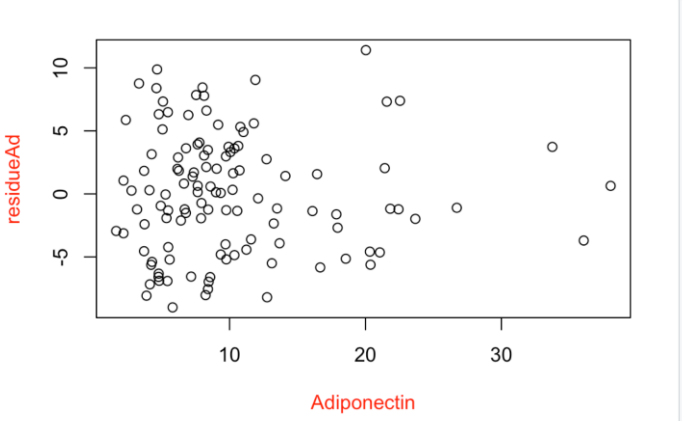


Fig.7 Residual Plot for MLR





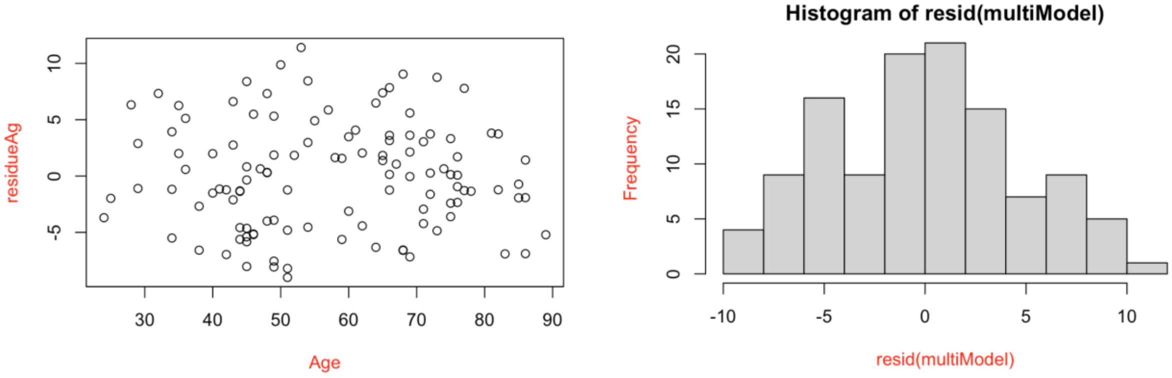


Fig.8 Residual Plot for Resistin, Adiponectin and Age with BMI

d) <u>Residual Plot Interpretation-</u>

Linearity: The residuals are randomly scattered around zero thus the variables are linear.

Heteroscedasticity (Constant Variance): The boxes are widening and narrowing that indicates it is Heteroscedasticity.

Normality: Its bell shaped normally distributed.

6. Conclusions:

Association: All the linear and multi regression models suggest a statistically significant association:

- Between Resistin and BMI
- Between Adiponectin and BMI
- Between Resistin, Adiponectin, Age and BMI

7. Limitations:

Linearity: All the linear and multi regression models assume a linear relationship

- Between Resistin and BMI
- Between Adiponectin and BMI
- Between Resistin, Adiponectin, Age and BMI

Independence: The model assumes independence of observations, so temporal or spatial dependencies may affect the model's validity.

Outliers: Extreme values or outliers in elevation could disproportionately influence the model in a large dataset.