Biostatistical Analysis of Cholesterol Level

Report

After importing and arranging the dataset (by using data.frame), we got the 3 by N structure given below :-

	Drugs	Factors(lifestyle)	LDL[mg/L]
1	Placebo	Exercise	298.7441
2	Placebo	Exercise	300.6534
3	Placebo	Exercise	295.2993
4	Placebo	Exercise	312.1012
5	Placebo	Exercise	309.4755
6	Placebo	Exercise	285.9190
7	Placebo	Exercise	310.2541
8	Placebo	Exercise	301.5346
9	Placebo	Exercise	306.6640
10	Placebo	Exercise	304.7180
11	Placebo	Exercise	300.3269
12	Placebo	Exercise	292.2435
13	Placebo	Exercise	292.7749
14	Placebo	Exercise	304.5919
15	Placebo	Exercise	301.5371
16	Placebo	Exercise	301.2287
17	Placebo	Exercise	297.6162
18	Placebo	Exercise	306.7961
19	Placebo	Exercise	293.3471
20	Placebo	Exercise	302.0243

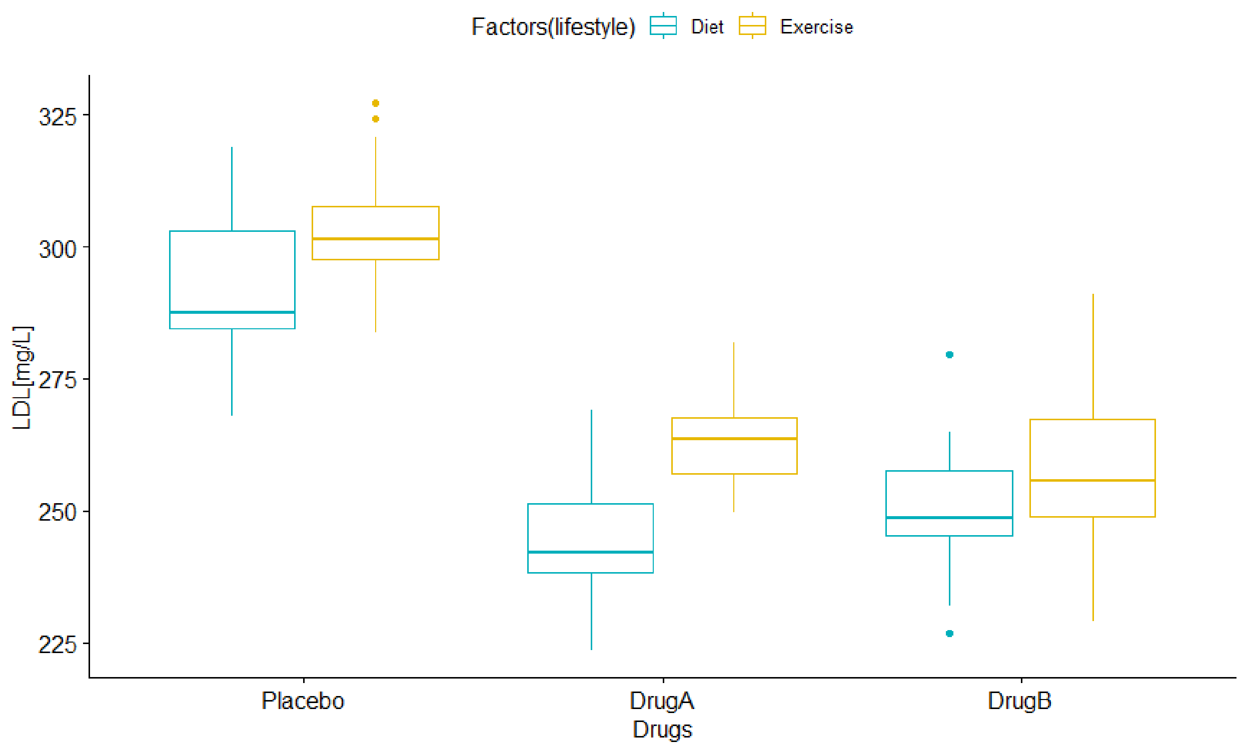
Two-way ANOVA Test

Result

```
                 Df Sum Sq Mean Sq F value   Pr(>F)    
Drugs             2  63491   31746  227.64  < 2e-16 ***
Factor_lifestyle  1   6379    6379   45.74 3.02e-10 ***
Residuals       146  20360     139                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

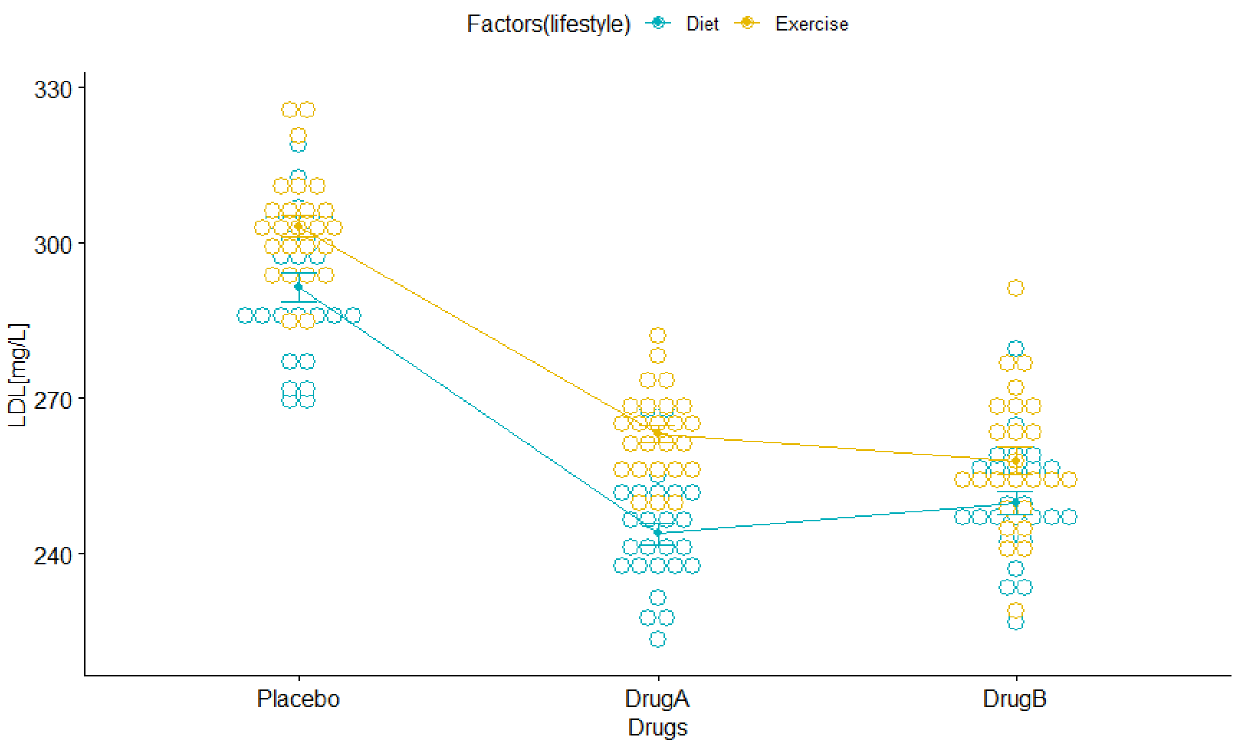
Inference :- After doing two way ANOVA test, we got the above table in which Df is degree of freedom ,the Sum Sq are the corresponding BSS and WSS, the mean squares are the BMS and WMS and in the subsequent columns, we have computed F value and p-Value. From the above ANOVA table, we can conclude that both Drugs and lifestyle factors are statistically significant as p-Value is less than 0.05. i.e.,rejecting the Null Hypothesis. Among both the factors, **Drug is the most significant factor variable i.e., will impact more on cholesterol level**.

Boxplot



Inference: As shown in the above plot, cholesterol level is lower in case of DrugA and DrugB than Placebo that shows DrugA and DrugB impact more on lowering the cholesterol level than Placebo.

Interaction plot



Inference: As shown in the above plot, among the medical drugs, we can see cholesterol level is lower when lifestyle factor is Diet than when it is Exercise.

Post-hoc test

```
        DrugA  DrugB
DrugB   0.88   -
Placebo <2e-16 <2e-16

P value adjustment method: BH
```

Inference: For finding a post-hoc test, pairwise t-test is used, and we got the above result. In the above table we can see the p-value between Placebo-DrugA and Placebo-DrugB is significant i.e., p-Value is less than adjusted p-Value.

Linear regression

Explanatory variable: Weight of an animal species in [mg] (i.e., x)

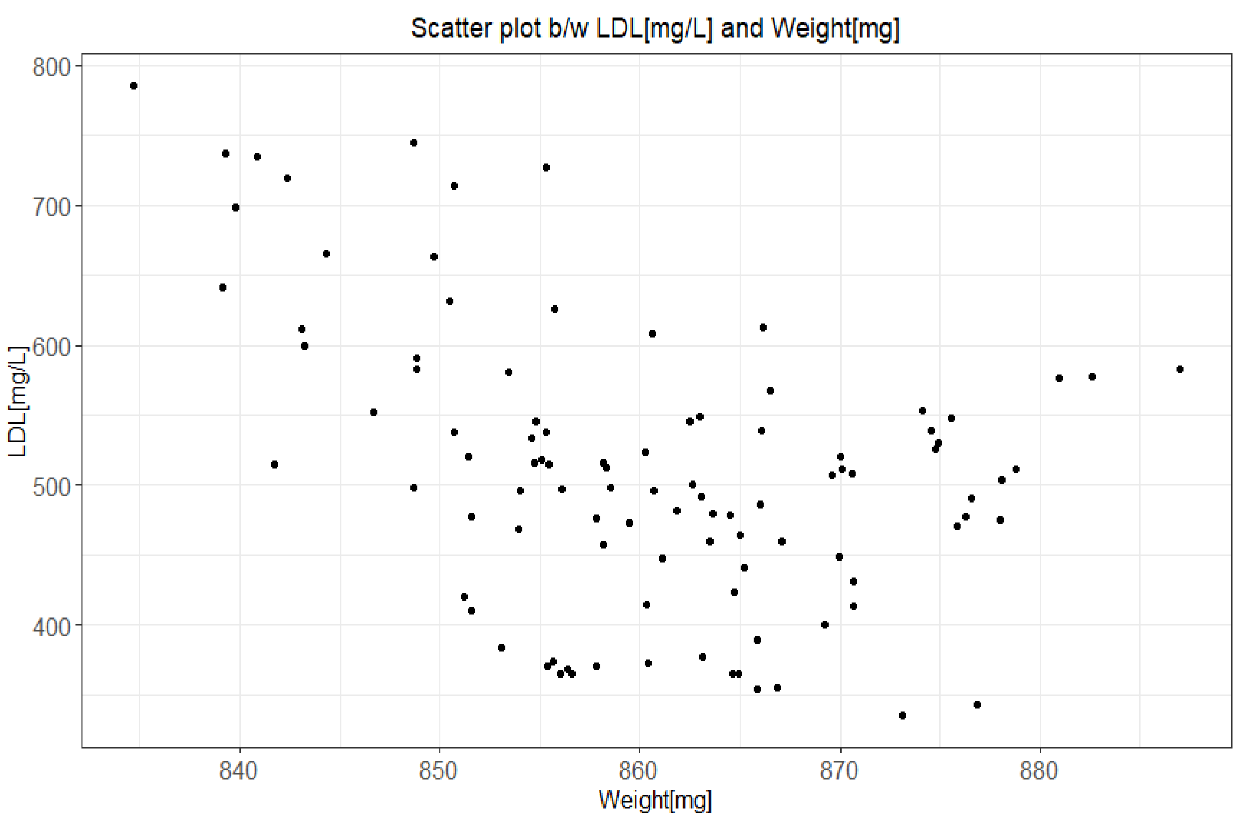
Dependent variable: Cholesterol level LDL (i.e., y)

Mean

- **Explanatory variable** = 860.4183
- **Mean of Dependent variable** = 509.691

Variance

- **Explanatory variable** = 120.2779
- **Dependent variable** = 10261.84



Parameters for the standard linear model

```
Call:
glm(formula = Data_reg$y ~ Data_reg$x, data = Data_reg)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-161.631   -59.508    -6.328    54.278   198.034

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3884.7096   726.8888   5.344 5.92e-07 ***
Data_reg$x    -3.9225     0.8447  -4.643 1.06e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 8497.04)

    Null deviance: 1015922  on 99  degrees of freedom
Residual deviance:  832710  on 98  degrees of freedom
AIC: 1192.5

Number of Fisher Scoring iterations: 2
```

Inference: The estimated effect of weight on cholesterol level is -3.92. That means that for every 1% increase in weight of animal species, there is a correlated 3.92% decrease in cholesterol level. The standard errors for the regression coefficients are 726.88 and 0.8447. The t-statistics are small i.e.,5.344 and -4.643 respectively. The p-values reflect the standard errors and smaller t-statistics.

For standard linear model y = mx + b

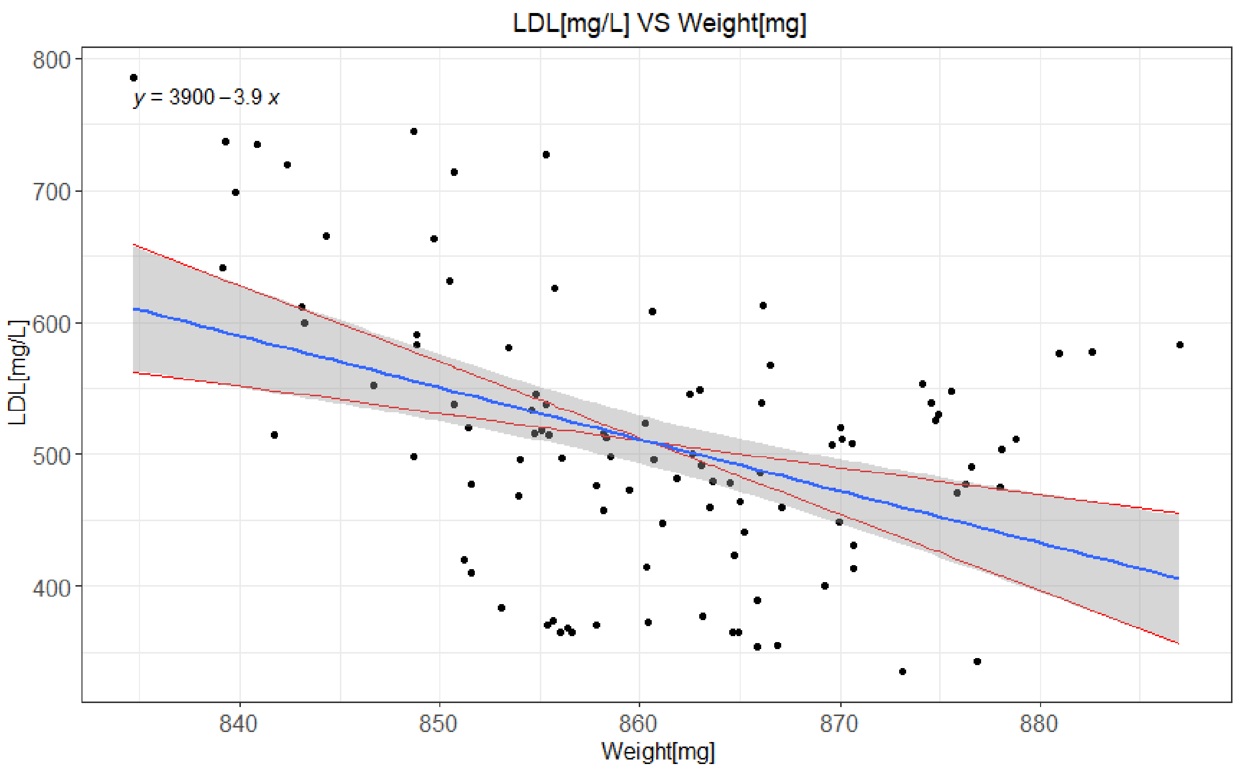
m = -3.922532

b = 3884.71

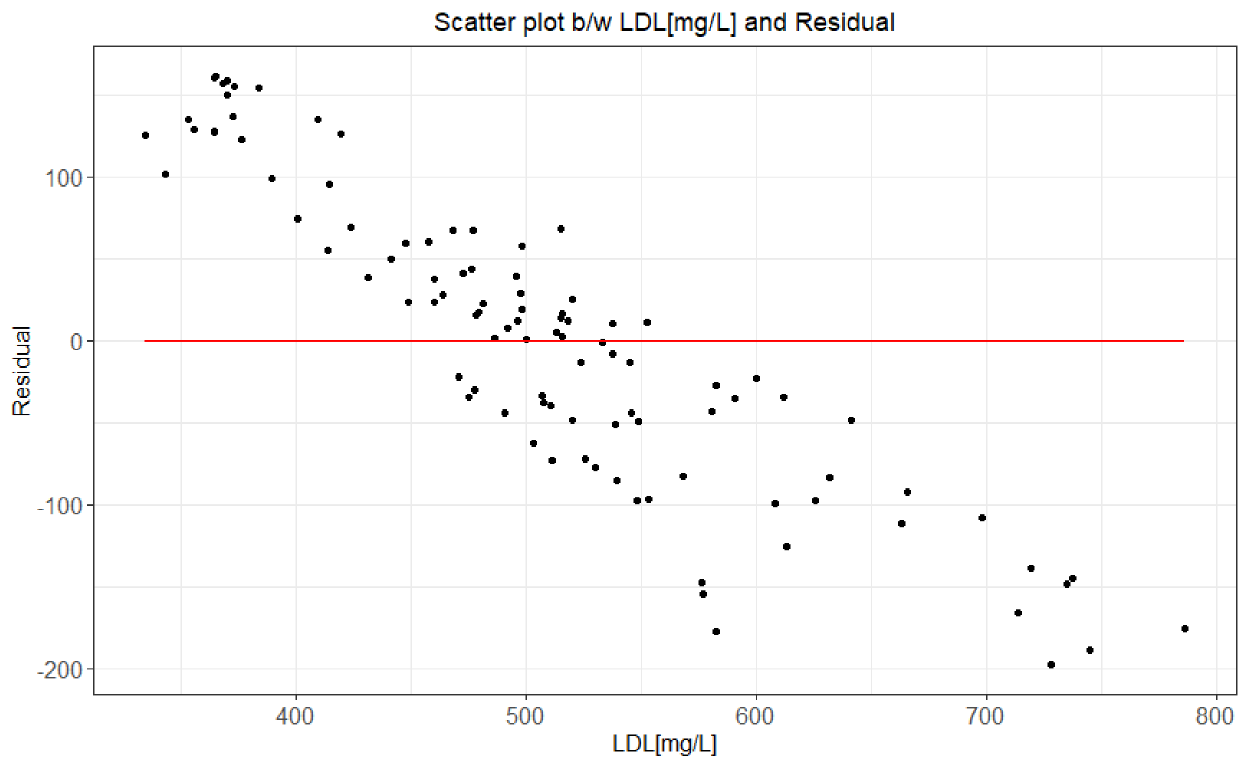
Inference: This linear model is not valid for any value of x beyond the dataset because this model is for this dataset only i.e., it has its own m and b values. For the different dataset there might be a different slope and b intercept which may lead to a bad model.

Confidence Interval

- **For slope(m),** it lies between -5.782775 and -2.062289.
- **For Intercept(b),** it lies between 2283.994 and 5485.425.



Inference: In the above Scatter plot, there are calculated optimal regression lines (blue), upper-limit and lower-limit regression lines (red), Autogenerated slope (superimposed to blue line).



Inference: In the above plot, the scatter plot of the residuals is plotted. Taking the hint from the lecture of linear regression, we can say the spread of **residuals is biased and Heteroscedastic** as average value is not zero in any thin vertical strip, also the spread of the residuals is not equal in any thin strip. This is also because of variance which depends on explanatory variables. Thus, the variance is not equal.