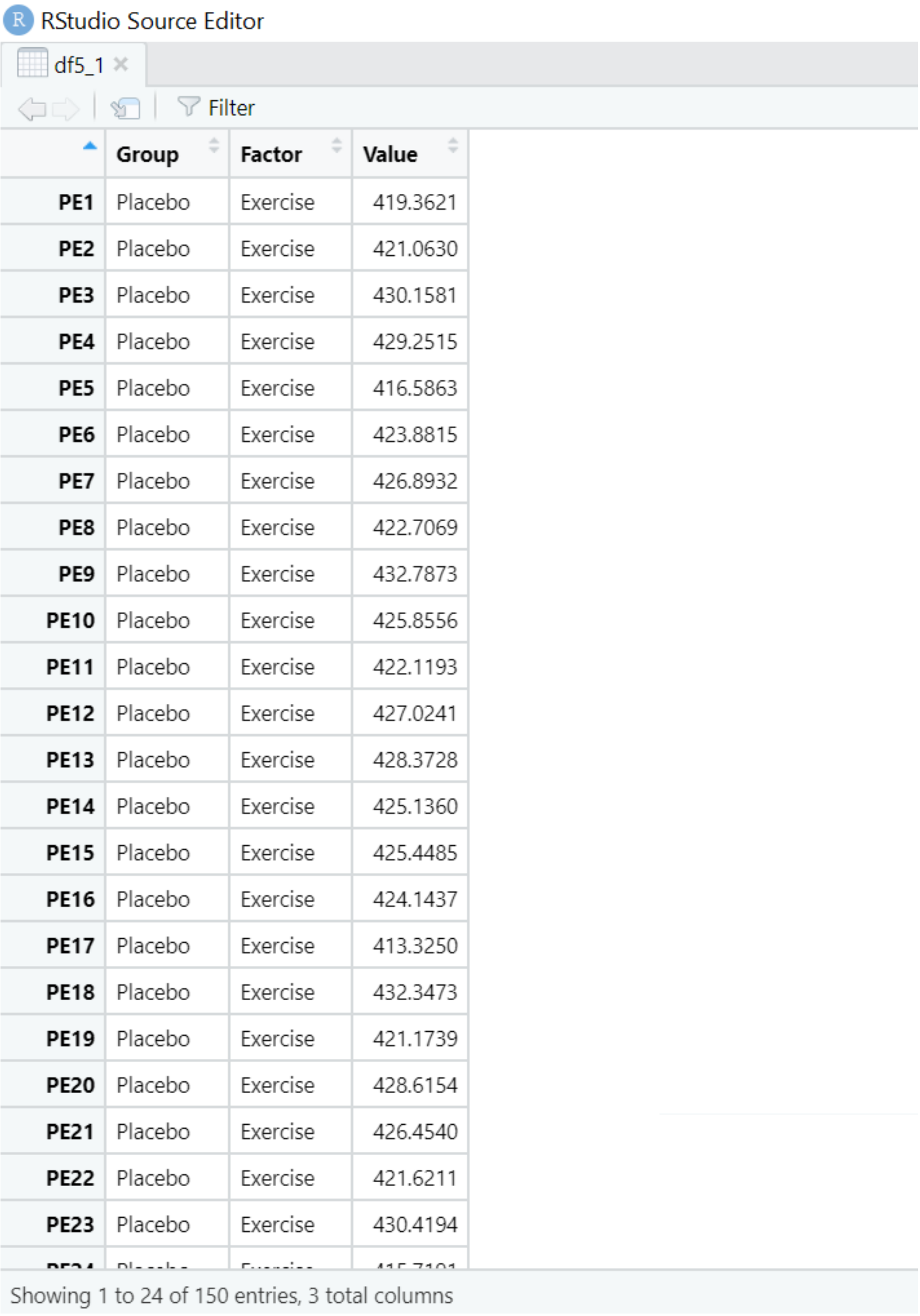
Biostatistical Data Analysis in R

Report

2.1 Two-way ANOVA

1. First few entries of the required data frame (3X150):



	Group	Factor	Value
PE1	Placebo	Exercise	419.3621
PE2	Placebo	Exercise	421.0630
PE3	Placebo	Exercise	430.1581
PE4	Placebo	Exercise	429.2515
PE5	Placebo	Exercise	416.5863
PE6	Placebo	Exercise	423.8815
PE7	Placebo	Exercise	426.8932
PE8	Placebo	Exercise	422.7069
PE9	Placebo	Exercise	432.7873
PE10	Placebo	Exercise	425.8556
PE11	Placebo	Exercise	422.1193
PE12	Placebo	Exercise	427.0241
PE13	Placebo	Exercise	428.3728
PE14	Placebo	Exercise	425.1360
PE15	Placebo	Exercise	425.4485
PE16	Placebo	Exercise	424.1437
PE17	Placebo	Exercise	413.3250
PE18	Placebo	Exercise	432.3473
PE19	Placebo	Exercise	421.1739
PE20	Placebo	Exercise	428.6154
PE21	Placebo	Exercise	426.4540
PE22	Placebo	Exercise	421.6211
PE23	Placebo	Exercise	430.4194
PE24	Placebo	Exercise	415.7101

Showing 1 to 24 of 150 entries, 3 total columns

2.

Two-way ANOVA for Factors and Groups:

```
               Df    Sum Sq   Mean Sq   F value   Pr(>F)
Group           2     31753    15877    153.655    <2e-16 ***
Factor          1      989      989       9.568     0.00237 **
Residuals     146    15086      103
---
Signif. codes:  0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' '  1
```

Two-way ANOVA for Interaction between Factors and Groups:

```
                  Df    Sum Sq   Mean Sq   F value   Pr(>F)
Group              2     31753    15877     713.46    <2e-16 ***
Factor             1      989      989       44.43    5.23e-10 ***
Group:Factor       2     11881    5941      266.96   < 2e-16 ***
Residuals        144     3204      22
---
Signif. codes:  0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' '  1
```

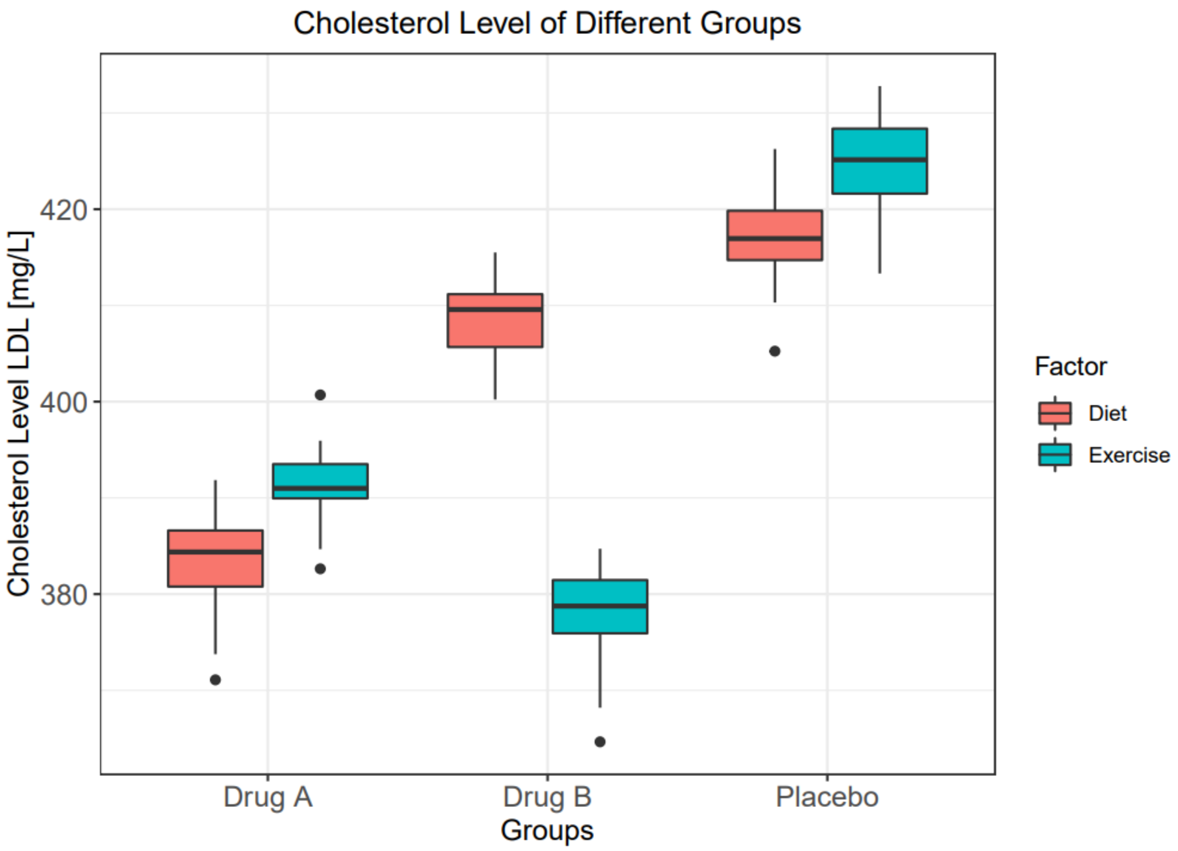
Analysis:

From the first ANOVA table we can conclude that both Group and Factor are statistically significant (p-value < 0.05). Group i.e., the effect of medical drugs, is the most significant variable. These results imply that changing the medical drugs – drug A, drug B, placebo (Groups) or the lifestyle – exercise, diet (Factors), will impact the cholesterol level LDL significantly.

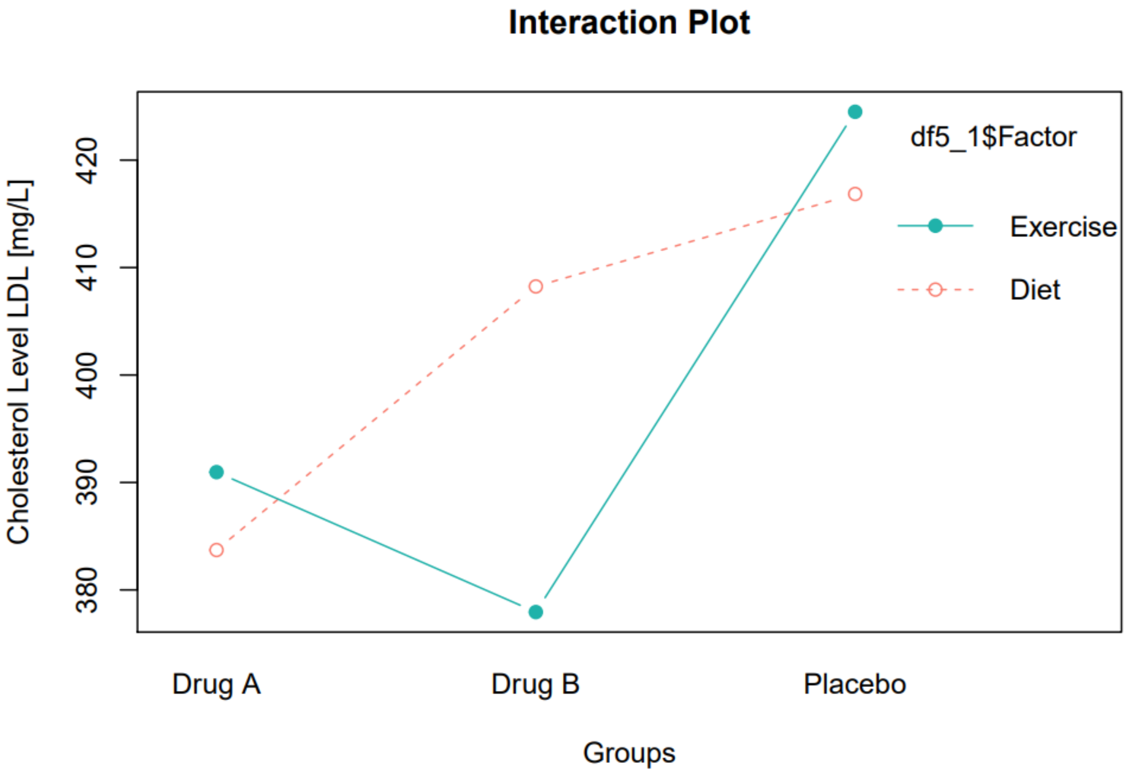
This analysis was done assuming that the 2 variables are independent. Now, taking their interaction into account:

From the second ANOVA table we can conclude that Group and Factor are statistically significant, as well as their interaction. The p-value for interaction between Group and Factor is < 2e-16 (significant) which means that the relationship between lifestyle (exercise, diet) and cholesterol level depend on the medical drug taken (drug A, drug B, placebo) and vice versa.

3. *Boxplot:*



Interaction Plot:



Analysis:

In the interaction plot, the 6 dots represent the group means for various factors. The lines are changing w.r.t. x axis i.e., groups. To check whether they also change w.r.t factors we have to refer the difference between the points. For Drug A and Placebo, the difference is not much but for Drug B there is a significant difference between Cholesterol level LDL of Exercise factor and Diet factor (the same can also be observed from the box plot). This implies that Group is significant, Factor is significant and Factor X Group is also significant.

$G_i \neq 0$
$F_j \neq 0$
$(G \times F)_{ij} \neq 0$

This result can also be seen by conducting two-way ANOVA as done in question 2.

---

4. *Post-hoc test (Tukey Test):*

```
Tukey multiple comparisons of means

95% family-wise confidence level

factor levels have been ordered


Fit: aov(formula = Value ~ Group + Factor, data = df5_1)


$Group

                      diff         lwr          upr          p adj

Drug B-Drug A      5.754264    0.940415    10.56811    0.0145862

Placebo-Drug A    33.336366   28.522516    38.15022    0.0000000

Placebo-Drug B    27.582101   22.768252    32.39595    0.0000000


$Factor

                      diff         lwr         upr       p adj

Diet-Exercise      5.134535    1.853931    8.41514   0.002373
```

Analysis:

diff represents difference between means of the two groups,
lwr, upr represent the lower and the upper end point of the confidence interval at 95%, and
p adj represent the p-value after adjustment for the multiple comparisons.

From the output, it can be seen that Placebo is significantly different from Drug A and Drug B but most significant against the Drug A group. Thus, we can say that Placebo gives a significant effect on cholesterol level LDL. Drug A - Drug B and Exercise - Diet did not show much difference but still have an effect on the cholesterol level LDL.

2.2 Linear Regression

In all the linear regression questions,
y refers to the dependent variable (Cholesterol level LDL [mg/L]) and
x refers to the explanatory variable (Weight of animal species [mg]).

5.

Explanatory Variable (x) – Weight of animal species [mg]:

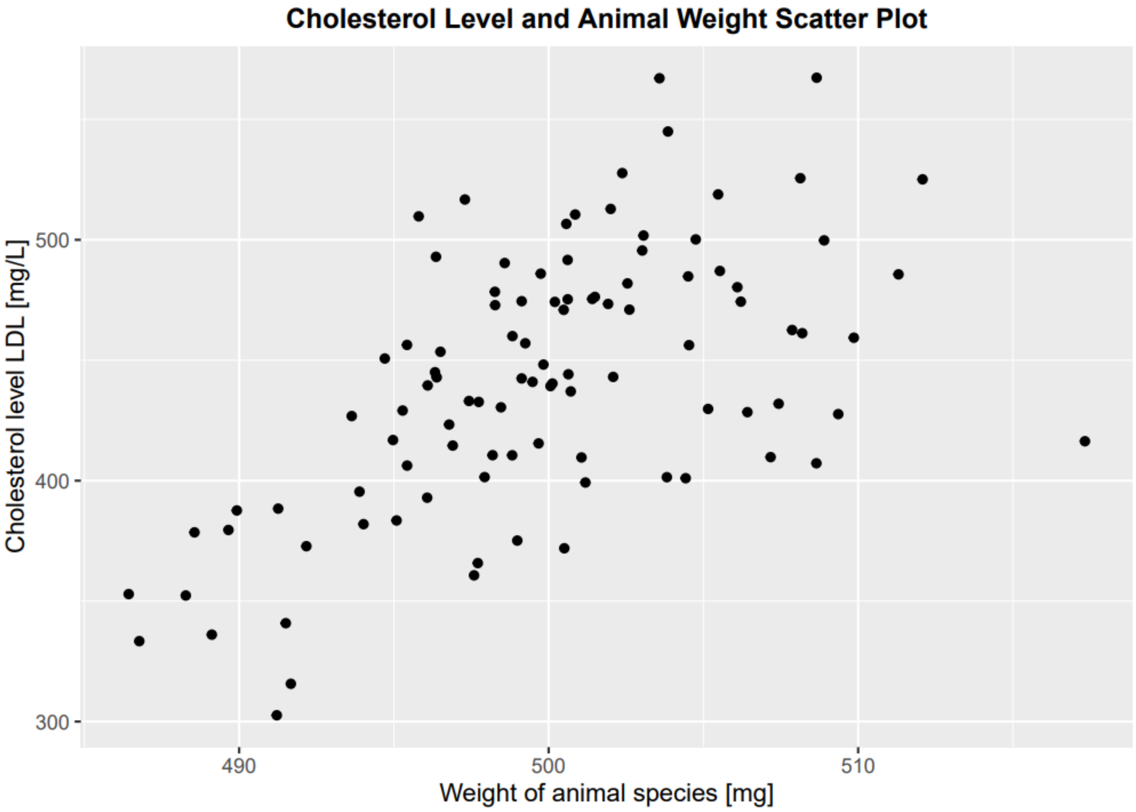
Mean = 499.8605

Variance = 34.51972

Dependent Variable (y) – Cholesterol level LDL [mg/L]:

Mean = 441.902

Variance = 2955.786



6.

```
Call:

glm(formula = y ~ x)


Deviance Residuals:

    Min         1Q       Median     3Q        Max

-119.467   -27.940     0.178     29.432   105.159


Coefficients:

                     Estimate       Std. Error    t value     Pr(>|t|)

(Intercept)     -2246.6439      380.2284     -5.909      5.01e-08 ***

x                      5.3786          0.7606        7.071      2.29e-10 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


(Dispersion parameter for gaussian family taken to be 1977.127)

Null deviance: 292623 on 99 degrees of freedom

Residual deviance: 193758 on 98 degrees of freedom

AIC: 1046.7

Number of Fisher Scoring iterations: 2
```

Parameters-

Standard linear model:
y = mx + b

Estimate of m = 5.3786   (Slope)

Estimate of b = -2246.6439   (y-intercept)

Analysis:

- The *Call* tells us the original call to the glm function.

- The *Deviance Residuals* gives the summary of the residuals – minimum value (-119.467), first quartile (-27.940), median (0.178), third quartile (29.432) and the maximum value (105.159). Ideally, they should be equally distributed around 0 (which is true in this case).

- The *Coefficients* tell us about the least-squares estimates of the fitted line. The Estimate gives the value of intercept and slope. The Standard Error of the estimates and the t-value are also provided along with the p-values. The p-value for x should be < 0.05 for it to be statistically significant (2.29e-10 in this case, thus statistically significant).

  A significant p-value for x (weight of animal species) means that it will give us a reliable guess of y (cholesterol level LDL).

- The *Dispersion Parameter* for this Gaussian family is taken to be 1977.127 which is equal to the residual variance.

- Then we have the *Null Deviance* and the *Residual Deviance*.
  The Null deviance is 292623 on 99 degrees of freedom and
  The Residual deviance is 193758 on 98 degrees of freedom.

  The null deviance shows how well the response is predicted by the model with only the intercept. The residual deviance shows how well the response is predicted by the model when the predictors are included.

- The *AIC* or the Akiake Information Criterion is a measure of goodness of fit that takes into account the ability of the model to fit the data. It can be used to compare different models. Smaller values indicate better fit (1046.7 in this case).

- *Number of Fisher Scoring iterations* tells us how many iterations it took to give the output (2 in this case).

Is the linear model valid for any value of x beyond the dataset?

The linear model y = 5.3786x -2246.6439 is NOT valid for any value of x beyond the dataset. This is because this model is constructed using values for the given dataset. If we change the dataset or include another value in it, then it would cause change in the parameters i.e., m and b. Thus, different data values would lead to different models.

7.

y = mx + b
slope = m
y-intercept = b

Confidence Interval with α = 5% (and df=98):

CI_mmax = 6.888012

CI_mmin = 3.869172

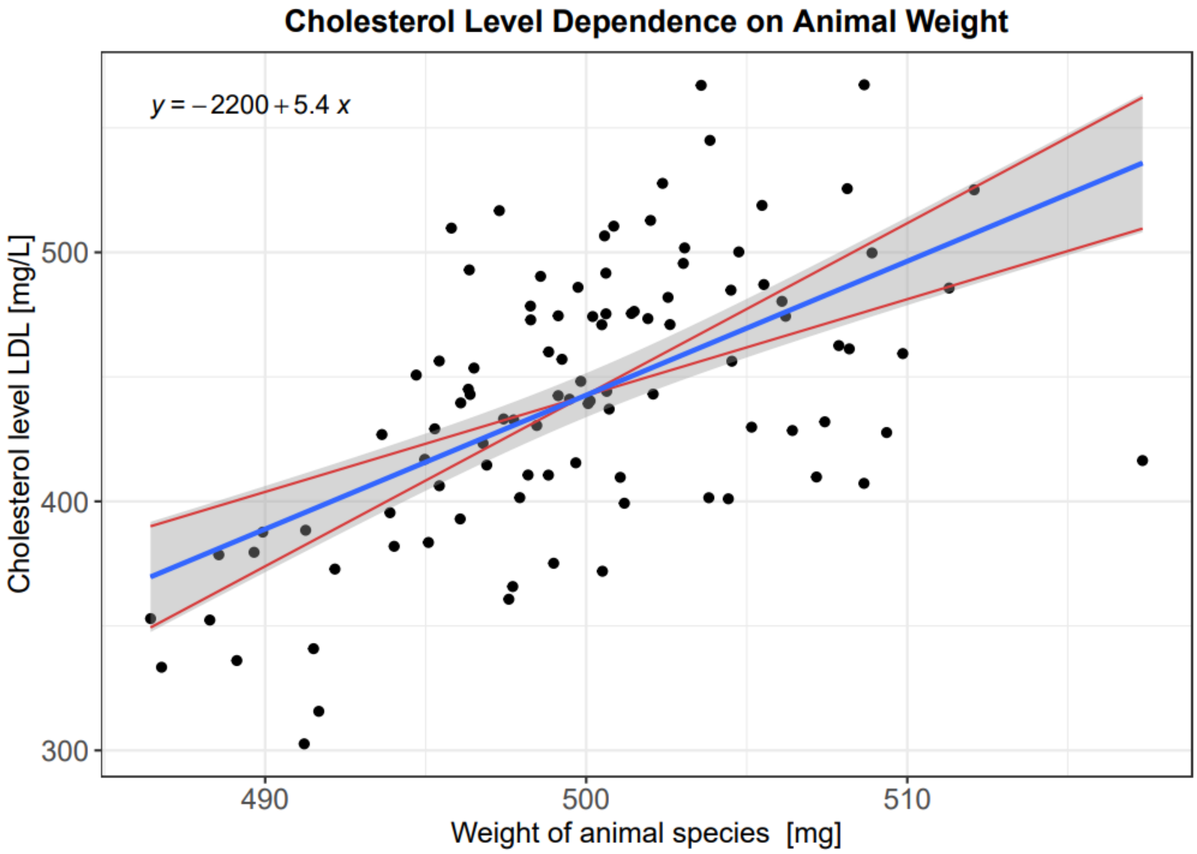
CI_bmax = -1492.093

CI_bmin = -3001.195

Confidence interval for slope m: [3.869172, 6.888012]

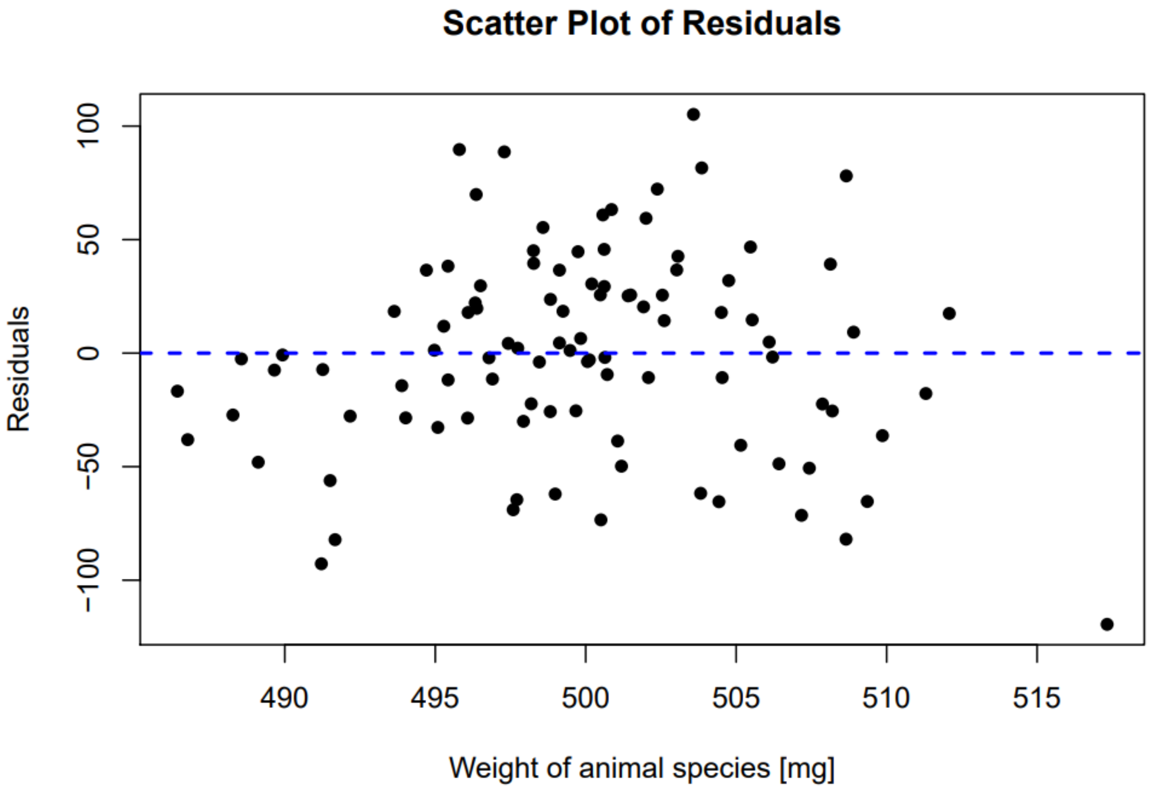
Confidence interval for y-intercept b: [-3001.195, -1492.093]

8.



Scatter plot with optimal regression line (blue) and confidence interval regression lines (red)

9.



Analysis:

The residual plot is biased and heteroscedastic.

(i) Biased: For the plot to be unbiased, the average value in any thin vertical strip should be zero. As we can see from the plot, for the left extreme (for explanatory variable < 493 (approx.)), the average value is not zero.

(ii) Heteroscedastic: For the plot to be homoscedastic, the spread of the residuals in any thin strip should be equal. It can be seen from the plot that this is not the case.