STA 5207: Homework 2

Due: Wednesday, September 20 by 11:59 PM

Include your R code as an appendix at the end of your homework. Do not include your code in your answers unless the question explicitly tells you to include your code. Your answers to each exercise should be self-contained without code so that the grader can determine your solution without reading your code or deciphering its output.

Exercise 1 (Using `lm` for Estimation) [35 points]

For this exercise we will use the `faithful` dataset. This is a default dataset in `R` , so there is no need to load it (when using `R`). Otherwise, you can find the data in `faithful.csv` on Canvas. The dataset contains 272 measurements of the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park in Wyoming. You should use `?faithful` to learn about the dataset's background. The variables in the dataset are

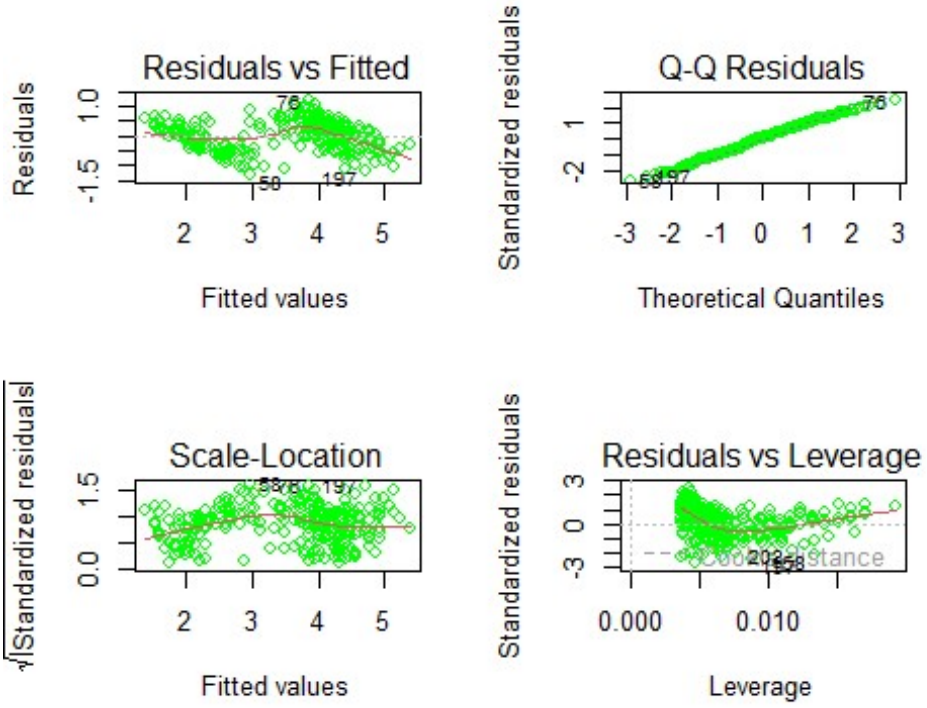
- `eruptions`: Duration of the eruption in minutes.

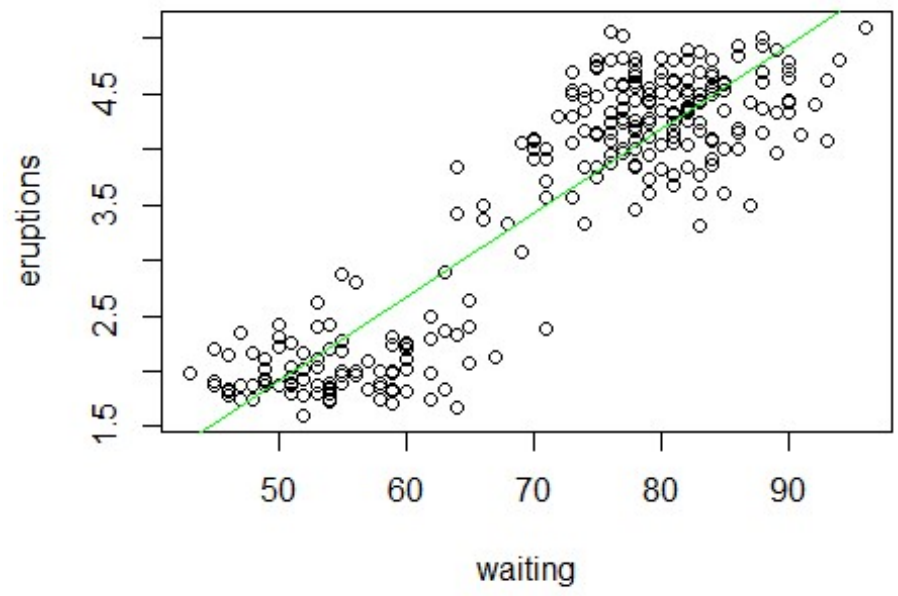
- `waiting`: Waiting time before the next eruption in minutes.

Suppose we would like to predict the duration of an eruption of the Old Faithful geyser based on the waiting time before the eruption.

1. (5 points) Give the simple linear regression model for this data set and the assumptions about the error terms.

Linearity: The relationship between the independent and dependent variables is linear. The model implies that the relationship may be expressed by a straight line Independent: The values of one observation's error terms are unrelated to the value of another observation's error terms. Normality: Because the error terms have a normal distribution, we can generate a Q-Q plot. A 45-degree ref line may be seen on that plot. Homoscedasticity: the residual spread should be roughly the same across the entire range of plot residuals vs fitted line.





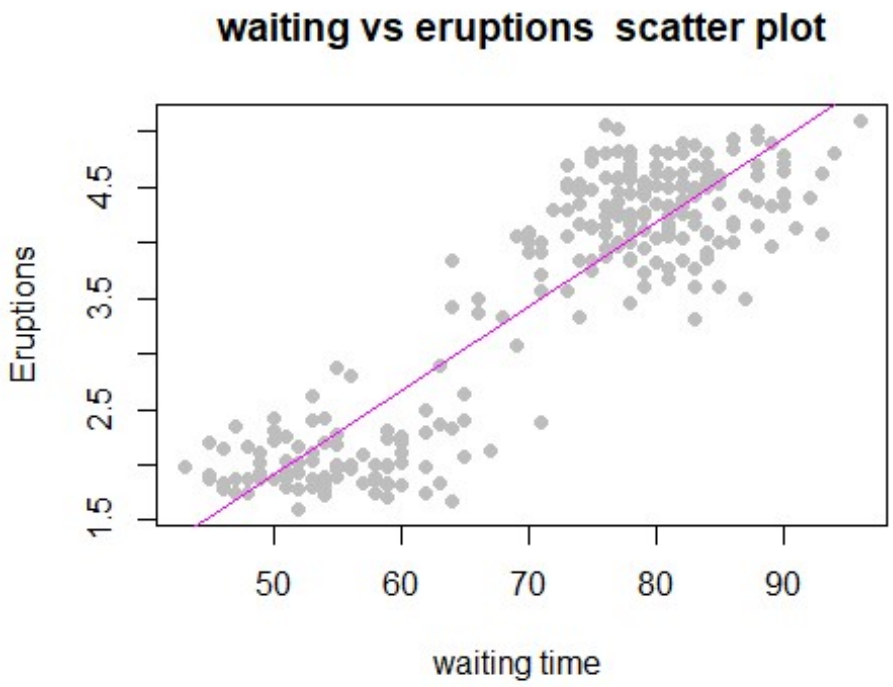
2. (4 points) What is the interpretation of β_1 *in the context of this problem*.

β1 signifies the approximated alteration in the eruption duration for every one-minute shift in waiting time.

3. (8 points) Give the estimated regression equation. What is the interpretation of $\hat{\beta}_1$ *in the context of this problem*?

Y = -1.874 + 0.07562(x) The slope coefficient indicates that for every additional minute of waiting time, the expected increase in the duration of the eruption is approximately 0.07562 minutes.

4. (4 points) Create a scatter plot of the data and add the fitted regression line. Make sure your plot is well labeled and is somewhat visually appealing.



5. (5 points) Use your model to predict the duration of an eruption based on a waiting time of **80** minutes. Do you feel confident in this prediction? Briefly explain.

Eruptions = -1.874016 + 0.07562795 * waiting ie -1.874016 + 0.07562795 *80 4.17622 minutes is the answer.I'm quite confident in our analysis because the Simple Linear Regression (SLR) meets the requirements of Linearity and Normality assumptions, and the R-squared (R2) value is a strong 0.81. However, we should investigate the presence of any outliers to ensure the robustness of our results.

6. (5 points) Use your model to predict the duration of an eruption based on a waiting time of **120** minutes. Do you feel confident in this prediction? Briefly explain.

Eruptions = -1.874016 + 0.07562795 * waiting ie -1.874016 + 0.07562795 *120 7.201338 minutes is the answer.I'm quite confident about this because our Simple Linear Regression (SLR) model meets the assumptions of linearity and normality, and the R-squared value is a robust 0.81. However, we still need to verify if there are any outliers.

7. (2 points) Report the residual standard error (RSE) for the model.

Residual standard error: 0.4965129 . can be found in the summary

8. (2 points) Give the value and interpretation of R^2 for the model.

The R^2(R2) value, which ranges from 0 to 1, assesses how well a regression model fits the data. In this case, the R2 value is 0.8115, indicating a strong level of fit.

Exercise 2 (Comparing Models With R^2) [30 points]

For this exercise, we will use the data stored in `goalies.csv`. It contains career data for all 716 players in the history of the National Hockey League to play goaltender through the 2014-2015 season. The variables in the dataset are:

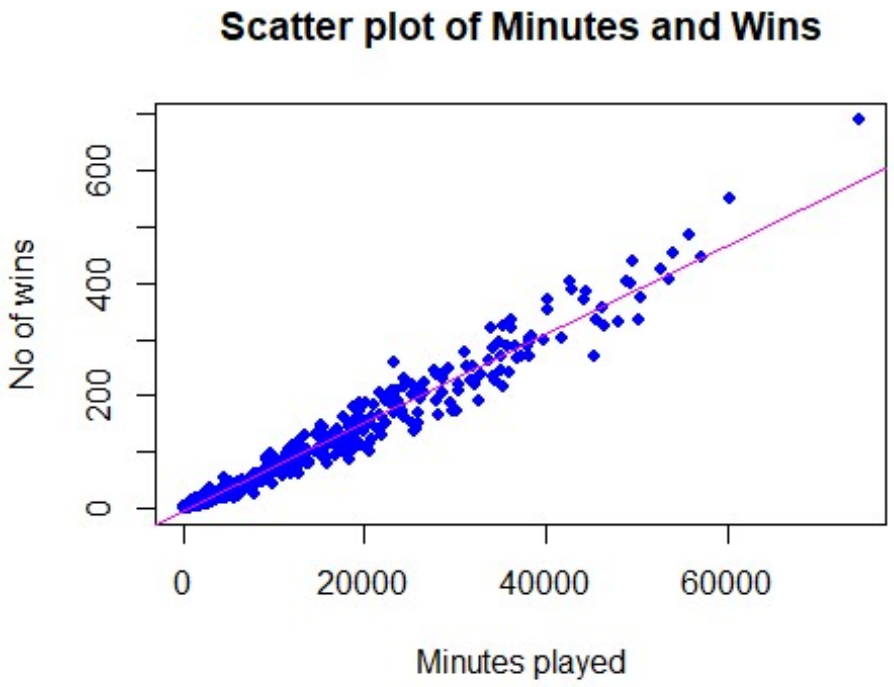
- `Player`: NHL Player Name
- `First`: First year of NHL career
- `Last`: Last year of NHL carerr
- `GP`: Games Played
- `GS` : Games Started
- `W`: Wins
- `L`: Losses
- `TOL`: Ties/Overtime/Shootout Losses
- `GA`: Goals Against
- `SA`: Shots Against
- `SV`: Saves
- `SV_PCT`: Save Percentages
- `GAA`: Goals Against Average
- `SO`: Shutouts
- `MIN`: Minutes
- `G`: Goals (that the player recorded, not opponents)
- `A`: Assists (that the player recorded, not opponents)
- `PTS`: Points (that the player recorded, not opponents)
- `PIM`: Penalties in Minutes

1. (6 points) Fit a model with "wins" as the response and "minutes" as the predictor. Report the value of R^2 for this model. Also provide a scatter plot of the fitted regression line.]

The model's R2 value stands at 0.9712.

```
## Rows: 716 Columns: 19
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (1): Player
## dbl (18): First, Last, GP, GS, W, L, TOL, GA, SA, SV, SV_PCT, GAA, SO,
MIN, ...
##
```

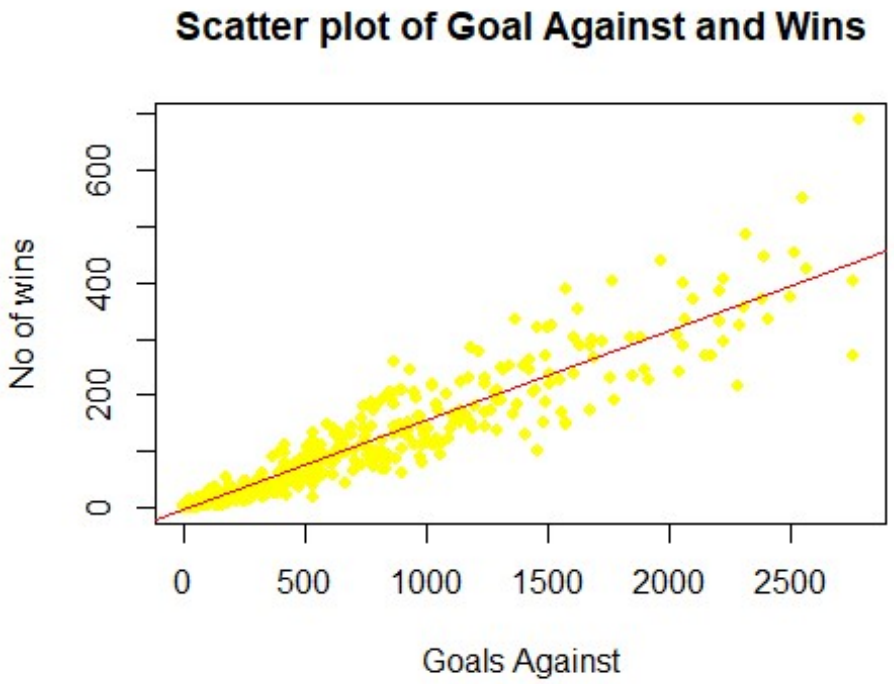
```
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.

## [1] 0.9711568
```



2. (6 points) Fit a model with "wins" as the response and "goals against" as the predictor. Report the value of R^2 for this model. Also provide a scatter plot of the fitted regression line.

The model's R2 value stands at 0.9008.

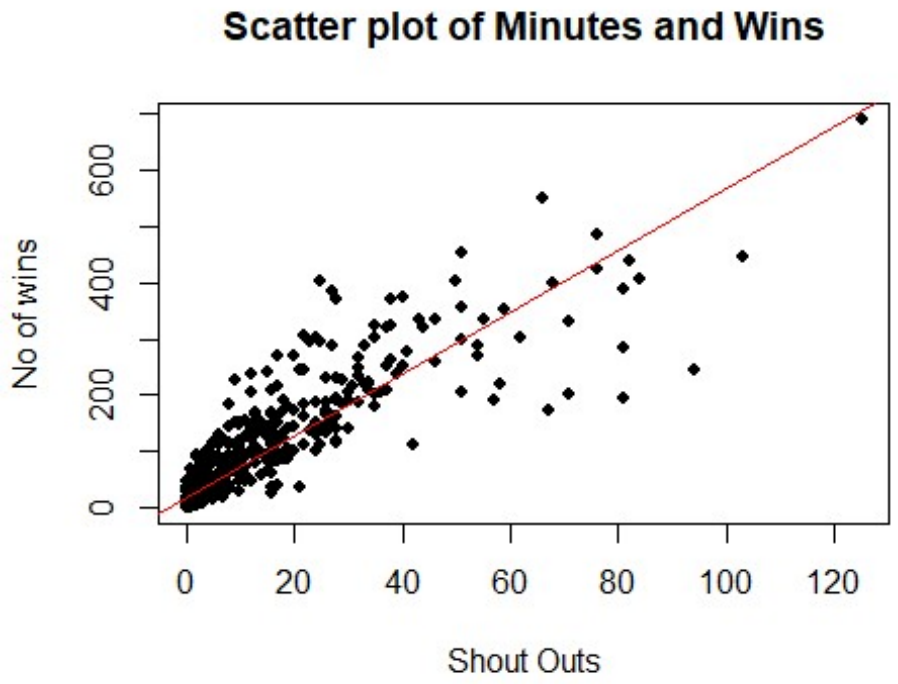
```
## [1] 0.9007736
```



3. (6 points) Fit a model with "wins" as the response and "shutouts" as the predictor. Report the value of R^2 for this model. Also provide a scatter plot of the fitted regression line.

The model's R2 value stands at 0.7932

```
## [1] 0.7934997
```



4. (12 points) If we use R^2 to measure each model's goodness-of-fit, which of the three predictors fits the data better? Briefly explain.

To determine which predictor aligns best with the data among "Wins vs Minutes," "Wins vs Goals Against" (GA), and "Wins vs Shutouts," we should look at the R-squared (R^2) values for each model. Here are the R^2 values:

R^2 for Wins vs Minutes: 0.9712 R^2 for Wins vs Goals Against: 0.9008 R^2 for Wins vs Shutouts: 0.7932 In this context, "Minutes" (MIN) has the highest R^2 value (0.9712), indicating that it provides the closest fit to the data.

Exercise 3 (Using `lm` for Inference) [35 points]

For this exercise, we will use the `cats` dataset from the `MASS` package. To load the dataset in `R`, run `data(cats, package='MASS')`. You can also find that data in `cats.csv` on Canvas. The dataset contains the following variables:

- `Sex`: The gender of the cat.
- `Bwt`: The body weight of the cat in kilograms.
- `Hwt`: The weight of the cat's heart in kilograms.

To read more about the dataset type `?cats` in RStudio.

1. (7 points) Fit the following simple regression model with the cat's heart weight as the response and its body weight as the predictor:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Use a t-test to test the significance of the regression. Report the following:

- The null and alternative hypothesis
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.05$
- A conclusion in the context of the problem.

Ans: The null hypothesis posited that there is no significant linear correlation between a cat's body weight and its heart weight, with the alternative hypothesis suggesting the opposite - a significant linear connection. The test statistic, which stands at 16.11939, produced a remarkably low p-value of 6.969045e-34.

Given that this p-value is substantially less than the conventional significance level of 0.05, we have compelling statistical grounds to reject the null hypothesis (H0). This outcome provides robust support for the alternative hypothesis (H1). In the context of this analysis, where the p-value is infinitesimally small compared to the 0.05 threshold, it strongly indicates the presence of a substantial linear association between a cat's body weight and its heart weight.

To sum it up, within the framework of the simple linear regression model utilizing a cat's heart weight (Hwt) as the outcome and its body weight (Bwt) as the predictor variable, our findings unequivocally demonstrate the existence of a statistically significant and meaningful linear relationship between these two variables.

2. (8 points) Use an F-test to test the significance of the regression. Report the following **and** discuss how they compare to the answers from the t-test in part 1:
   - The null and alternative hypothesis
   - The value of the test statistic
   - The p-value of the test
   - A statistical decision at $\alpha = 0.05$
   - A conclusion in the context of the problem.

Ans: Null Hypothesis: There is no substantial linear connection between a cat's body weight and its heart weight, expressed as H0: β1 = 0. Alternative Hypothesis: A significant linear association exists between a cat's body weight and its heart weight, denoted as H1: β1 ≠ 0. The computed test statistic yields a value of 259.83. The p-value associated with the test is exceedingly small, being less than 2.2e-16. When conducting a statistical analysis with a significance level (α) of 0.05, the decision rule is such that if the p-value is smaller than α,

the null hypothesis is rejected; conversely, if it is greater, we retain the null hypothesis. In the context of this problem, where the p-value is far below 0.05, it strongly indicates the existence of a significant linear connection between a cat's body weight and its heart weight.

In conclusion, based on our analysis using a simple linear regression model that employs a cat's heart weight (Hwt) as the outcome and its body weight (Bwt) as the predictor variable, we find compelling evidence to reject the null hypothesis. The F-test p-value is less than the 0.05 significance level (Alpha = 0.05), reinforcing our decision to reject the null hypothesis.

3. (5 points) Give the 99% confidence interval for β_1. Give an interpretation of the interval in the context of the problem.

The 99% confidence interval for β_1 is: 3.380656 4.687469

4. (5 points) Give the 90% confidence interval for β_0. Give an interpretation of the interval in the context of the problem.

The 99% confidence interval for β0 is:-1.502834 0.7895096

5. (5 points) Report a 95% confidence interval to estimate the mean heart weight for body weights of 2.5 and 3.0 kilograms. Which of the intervals is wider? Why?

For a body weight of 2.5 kilograms, the 95% confidence interval is:9.464902 9.992087 For a body weight of 3 kilograms, the 95% confidence interval is:11.46995 12.02110 The interval of 3 kg is wider than the interval of 2.5 kg body weight

6. (5 points) Report a 95% prediction interval to predict the heart weight for body weights of 2.5 and 4.0 kilograms.

For a body weight of 2.5 kilograms, the 95% Prediction interval is:6.845352 12.611637 For a body weight of 4 kilograms, the 95% Prediction interval is:12.83018 18.7290

Code Appendix

```
#library(faraway)
#colnames(faithful)
#View(faithful)
model= lm(eruptions~waiting,data = faithful)
par(mfrow=c(2,2))
plot(model,col='green')

par(mfrow=c(1,1))
plot(eruptions~waiting,data = faithful)
abline(lm(eruptions~waiting,data = faithful),col='green')
library(faraway)
#colnames(faithful)
#View(faithful)
lm(eruptions~waiting,data = faithful)
summary(model)
```

```r
intercept = coefficients(model)[1]
slope = coefficients(model)[2]
model=lm(eruptions~waiting,data = faithful)
plot(faithful$eruptions~faithful$waiting,xlab = 'waiting time ',ylab =
'Eruptions',main = 'waiting vs eruptions  scatter plot',pch = 19,cex =1,col =
'grey')
abline(model$coefficients ,col="magenta")
predict(model,newdata=data.frame(waiting = 80))
predict(model,newdata=data.frame(waiting = 120))
summary(model)$sigma
summary(model)$r.squared
library(readr)
goalies <- read_csv("C:/Users/rajitha/Downloads/goalies.csv")
#View(goalies)
#colnames(goalies)
par(mfrow=c(1,1))
model_2 = lm(W~MIN,data = goalies)
summary(model_2)$r.squared
plot(W~MIN,data = goalies,xlab = 'Minutes played',ylab = 'No of wins',main =
'Scatter plot of Minutes and Wins',pch= 20,cex=1.25,col = 'blue')
abline(model_2$coefficients,col='magenta')


model_3 = lm(W~GA,data = goalies)
summary(model_3)$r.squared
plot(W~GA,data = goalies,xlab = 'Goals Against',ylab = 'No of wins',main =
'Scatter plot of Goal Against and Wins',pch= 20,cex=1.25,col = 'yellow')
abline(model_3$coefficients,col='red')

model_4 = lm(W~SO,data = goalies)
summary(model_4)$r.squared
plot(W~SO,data = goalies,xlab = 'Shout Outs',ylab = 'No of wins',main =
'Scatter plot of Minutes and Wins',pch= 20,cex=1.25,col = 'black')
abline(model_4$coefficients,col='red')

data(cats, package='MASS')
#View(cats)
#colnames(cats)
#sum(is.na(cats))

model_5=lm(Hwt~Bwt,data = cats)
summary_model= summary(model_5)
#plot(Hwt~Bwt,data = cats)
summary_model$coefficients[2,"t value"]
summary_model$coefficients[2,"Pr(>|t|)"]

model_ = lm(Hwt~Bwt,data = cats)
null_model = lm(Hwt~1,data = cats)
annova_res = anova(null_model, model_)
```

```r
annova_res['F']
annova_res['Pr(>F)']

lm= lm(Hwt~Bwt,data = cats)
confidence_intervel_1 = confint(lm ,level = 0.99)
confidence_intervel_1
lm_model = lm(Hwt~Bwt,data = cats)
confidence_intervel_2 = confint(lm_model ,level = 0.90)
confidence_intervel_2
predicted_value<-predict(lm_model,newdata = data.frame(Bwt=c(2.5,3)), level =
0.95,interval = "confidence")
predicted_value[1,c("fit", "lwr", "upr")]
predicted_value[2,c("fit", "lwr", "upr")]
predicted_value1<-predict(lm_model,newdata = data.frame(Bwt=c(2.5,4)), level
= 0.95,interval = "prediction")
predicted_value1[1,c("fit", "lwr", "upr")]
predicted_value1[2,c("fit", "lwr", "upr")]
```