R Assignment - 4

Jasmeet Singh Saini - 0758054

2023-03-26

Question 1 - Red "Die" No. 40?

Required Data: *red_dye40.csv*

The safety of Red No. 40 (a dye) was tested using a mouse study. Groups of mice were provided with none of the dye (control group), a low, medium, or high amount and the time to death was recorded. This data is contained in the red_dye40.csv file.

Firstly, let's loading the data set and check the summary of *red_dye40.csv*.

```
# Importing the dataset
red_dye_dataset <- read.csv("red_dye40.csv")
# View the summary
summary(red_dye_dataset)
```

```
##     control            low            medium           high
##  Min.   : 70.00   Min.   :49.00   Min.   :30.00   Min.   : 34.00
##  1st Qu.: 85.00   1st Qu.:63.00   1st Qu.:58.25   1st Qu.: 45.00
##  Median : 93.00   Median :70.00   Median :79.50   Median : 56.50
##  Mean   : 91.36   Mean   :69.89   Mean   :71.50   Mean   : 65.25
##  3rd Qu.:101.00   3rd Qu.:77.00   3rd Qu.:89.25   3rd Qu.: 92.75
##  Max.   :103.00   Max.   :89.00   Max.   :97.00   Max.   :102.00
##                   NA's   :2       NA's   :1       NA's   :3
```

The data is seem to be inconsistent, as three of the variable has NA values, that is, low, medium and high. The "control" variable does not have any NA values. To make it look good, we will remove **NA** values from the dataset.

Now, let's remove the NA values:

```
low_na  <- which(is.na(red_dye_dataset$low))
med_na  <- which(is.na(red_dye_dataset$medium))
high_na <- which(is.na(red_dye_dataset))

control <- red_dye_dataset$control
low     <- red_dye_dataset$low[-low_na]
medium  <- red_dye_dataset$medium[-med_na]
high    <- red_dye_dataset$high[-high_na]
```

a) Is there any difference between the time to death? Conduct a hypothesis test.

Step 1 : Hypothesis & Assumptions

Step 1a: Hypothesis

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

$$H_0 : \text{There is no significant difference between the time to death}$$
$$\text{vs.}$$
$$H_A : \text{There is a significant difference between the time to death}$$

Step 1b: Assumptions

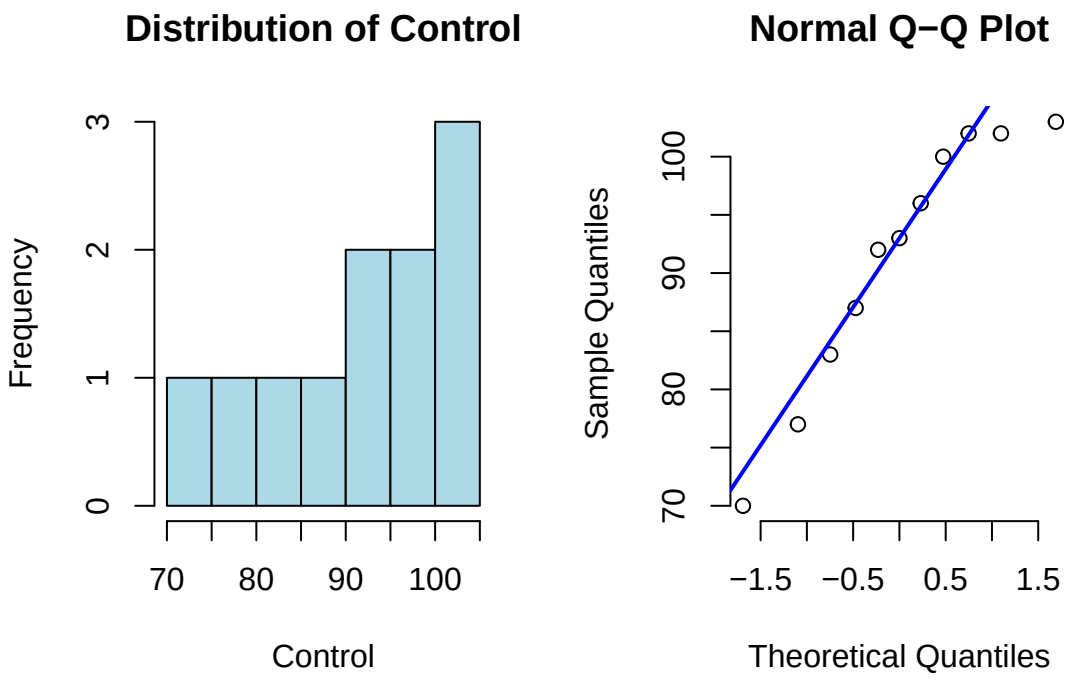
We need to check the assumptions of:

- **Independence** : The data was randomly sampled and the sample size is less than 10% of the population size. Thus, the condition is satisfied.
- **Normality** : We need to check if the distribution is normal across all groups.
- **Variance** : The variability across the groups should be about equal.

We will check for each of the variable:

Now, we will check for Normality of "control" variable:

```
par(mfrow=c(1,2))
hist(red_dye_dataset$control, main = "Distribution of Control"
     , xlab = "Control", col = "lightblue")
qqnorm(red_dye_dataset$control, pch = 1, frame = FALSE)
qqline(red_dye_dataset$control, col = "blue", lwd = 2)
```

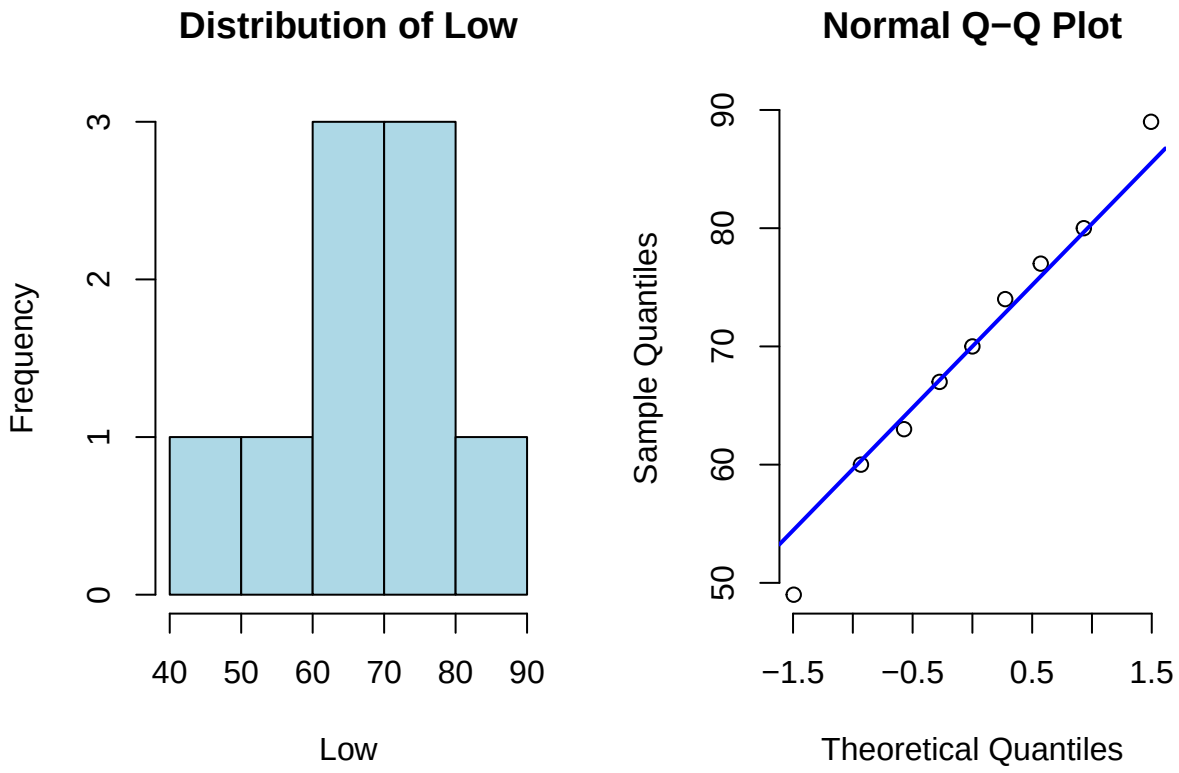


```
# Shapiro-Wilk's test,
# To check whether or not a sample fits a normal distribution
shapiro.test(red_dye_dataset$control)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  red_dye_dataset$control
## W = 0.90927, p-value = 0.2391
```

Secondly, we will check for Normality of "low" variable:

```
par(mfrow=c(1,2))
hist(red_dye_dataset$low, main = "Distribution of Low"
     , xlab = "Low", col = "lightblue")
qqnorm(red_dye_dataset$low, pch = 1, frame = FALSE)
qqline(red_dye_dataset$low, col = "blue", lwd = 2)
```



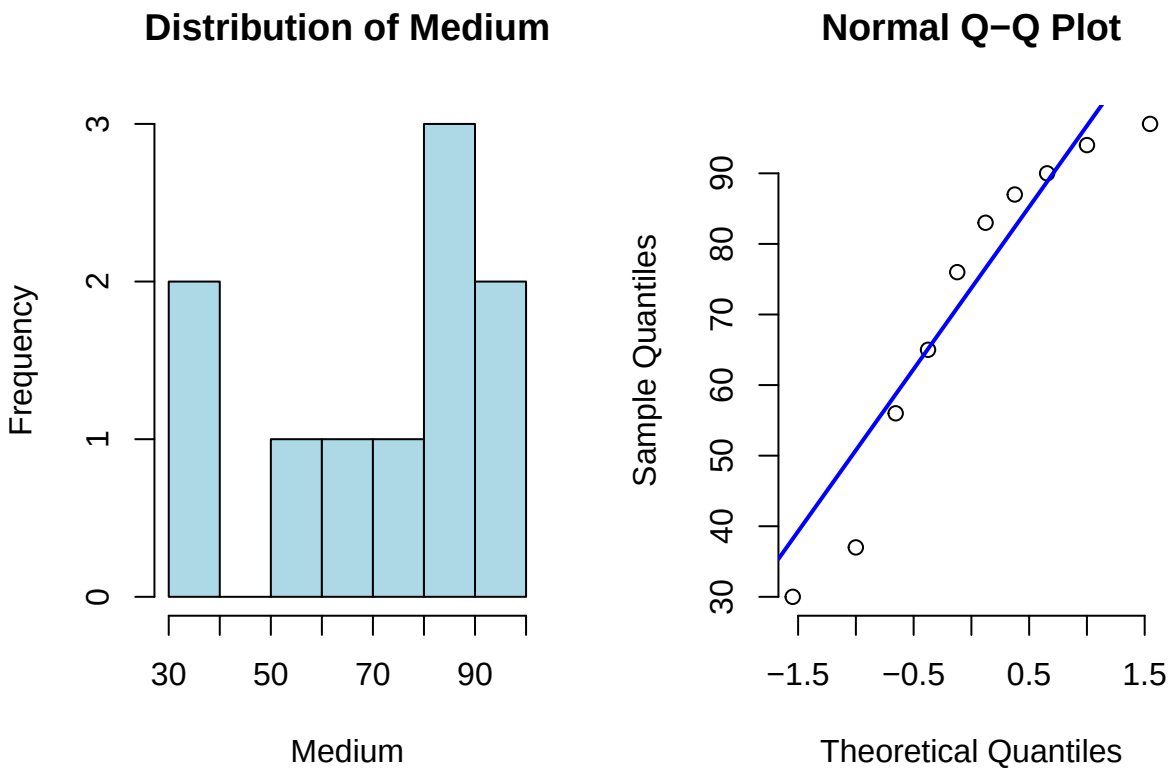

```
# Shapiro-Wilk's test,
# We will check p-value
shapiro.test(red_dye_dataset$low)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  red_dye_dataset$low
## W = 0.996, p-value = 0.9999
```

Now, we will check for Normality of "medium" variable:

```
par(mfrow=c(1,2))
hist(red_dye_dataset$medium, main = "Distribution of Medium"
     , xlab = "Medium", col = "lightblue")
qqnorm(red_dye_dataset$medium, pch = 1, frame = FALSE)
qqline(red_dye_dataset$medium, col = "blue", lwd = 2)
```

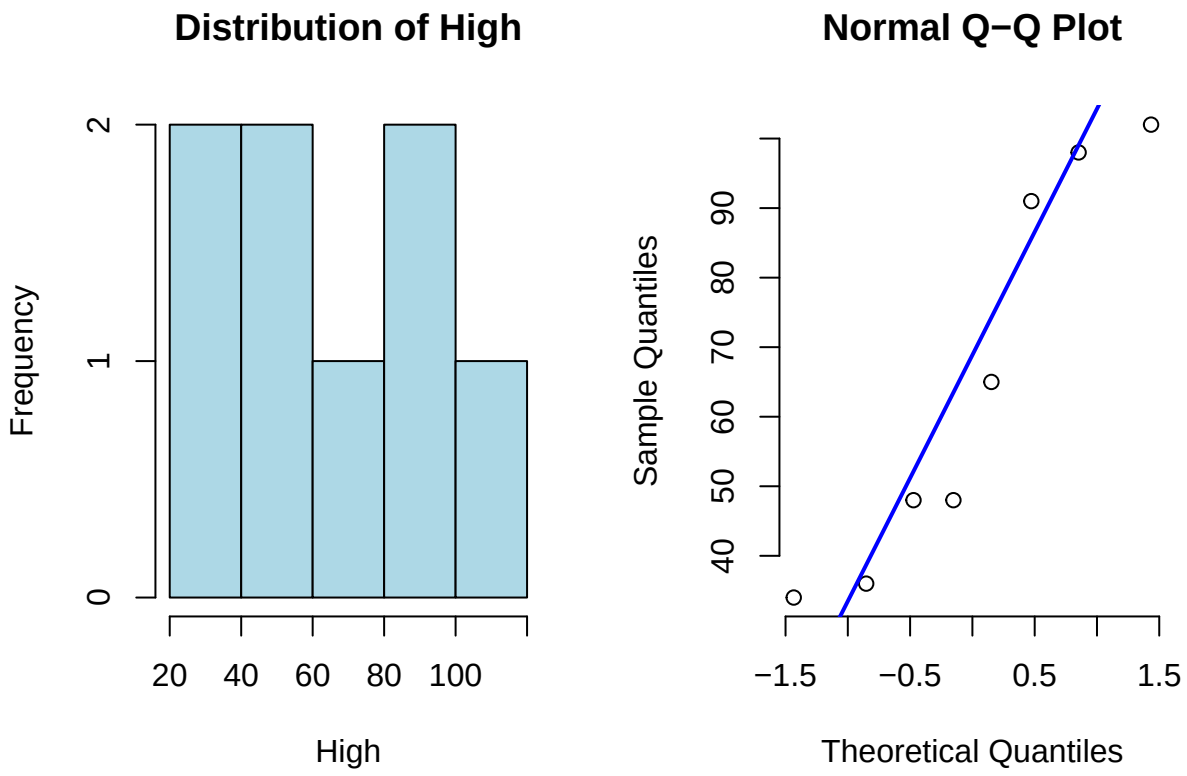



```
# Shapiro-Wilk's test
shapiro.test(red_dye_dataset$medium)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  red_dye_dataset$medium
## W = 0.89469, p-value = 0.1914
```

Lastly, we will check for Normality of "high" variable:

```
par(mfrow=c(1,2))
hist(red_dye_dataset$high, main = "Distribution of High"
     , xlab = "High", col = "lightblue")
qqnorm(red_dye_dataset$high, pch = 1, frame = FALSE);
qqline(red_dye_dataset$high, col = "blue", lwd = 2);
```

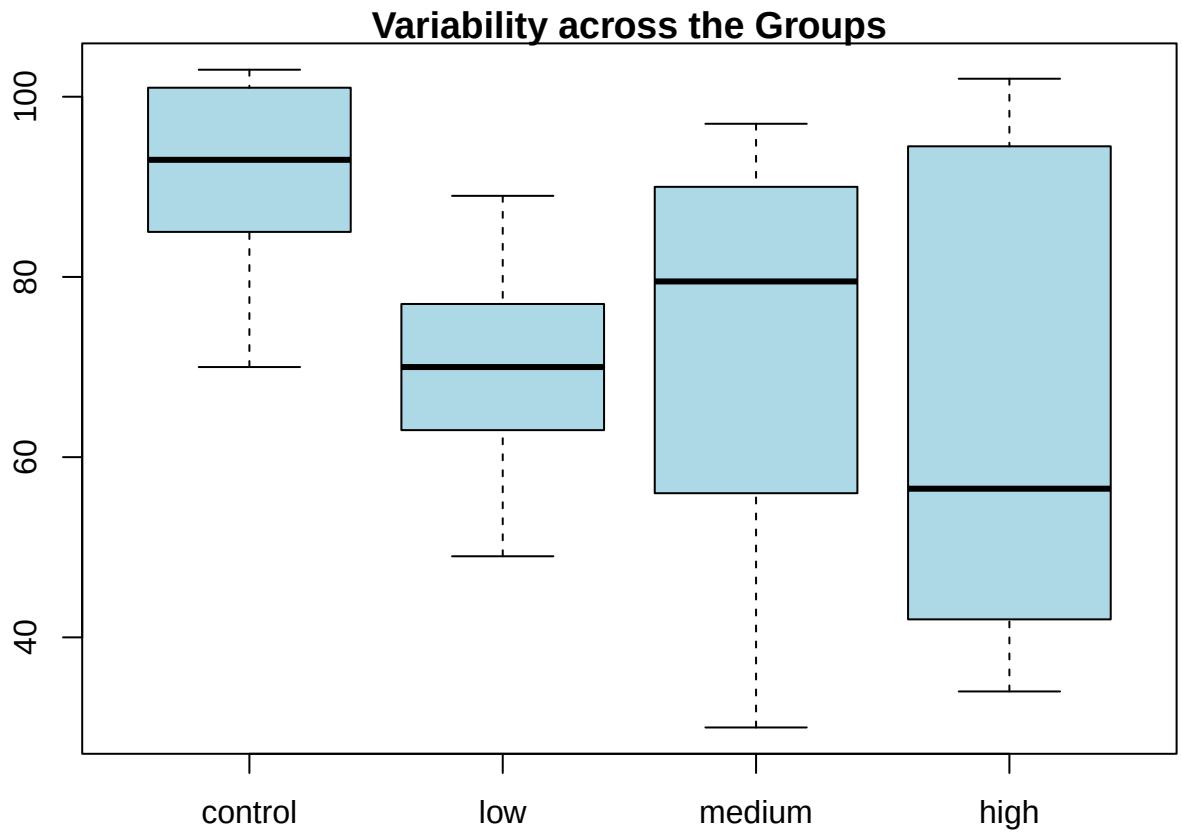



```
shapiro.test(red_dye_dataset$high)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  red_dye_dataset$high
## W = 0.86966, p-value = 0.1496
```

Here, the $p - value > 0.05$ for all the groups in Shapiro-Wilk's test and QQ-plot shoes nearly normal distribution. Hence, the condition for normality is met and we can assume the normality in all the groups.

```
par(mar = c(3, 3, 1, 1))
boxplot(red_dye_dataset, cex.axis = 1.0
        , main = "Variability across the Groups", col = "lightblue")
```



It can be seen that, the variances across each group is nearly normal. Hence, all the assumptions are met and now, we can use parametric approach.

Step 2 : Test Statistic and p-value

Now, we need to analyze the result of **One-way ANOVA Test** to make our decision.

```
dataframe <- data.frame(Group = c(rep("control", length(control))
                                  ,rep("low", length(low))
                                  , rep("medium", length(medium))
                                  , rep("high", length(high)))
                        , Time = c(control, low, medium, high))
```

Now, let's test the ANOVA and get the summary:

```
aov_eq <- aov(Time ~ Group, data = dataframe)
summary(aov_eq)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Group        3   4052  1350.7    3.55 0.0245 *
## Residuals   34  12937   380.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3 observations deleted due to missingness
```

Hence, the test statistic is **3.55** and p-value is **0.0245**.

Step 3 : Statistical Decision

Therefore, we **reject** Null Hypothesis, H_0 because p-value is less than α, level of significance (that is, p-value < 0.05).

Step 4 : Conclusion

There is enough evidence to support the alternative hypothesis, and therefore we conclude that there is a significant difference between time to death, is accurate.

b) If there is a difference, which groups are different?

To check whether there is a difference of mean in all groups, we will preform Post-hoc test (Tukey HSD - Statistical Test for Differences).

```
posthoc <- TukeyHSD(aov_eq)
posthoc
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Time ~ Group, data = dataframe)
##
## $Group
##                       diff       lwr       upr     p adj
## high-control   -26.113636 -50.59372 -1.633551 0.0328848
## low-control    -21.474747 -45.15437  2.204879 0.0868094
## medium-control -19.863636 -42.88286  3.155592 0.1109953
## low-high         4.638889 -20.96086 30.238636 0.9609302
## medium-high      6.250000 -18.74014 31.240141 0.9056532
## medium-low       1.611111 -22.59544 25.817667 0.9978960
```

Here, the "*p adj*" values of "low-control", "medium-control", "low-high", "medium-high" and "medium-low" groups is **0.0868094**, **0.1109953**, **0.9609302**, **0.9056532** and **0.9978960** respectively, which is more than $\alpha = 0.05$. Hence, these groups have a difference in means.

The "*p adj*" value of "high-control" group is **0.0328848**, which is significantly less than $\alpha = 0.05$. Hence, we can say that there is **no difference of means in "high-control" group**.

Question 2 - Ground Water

Required Data: *groundwater.csv*

Water treatment plants add bicarbonate to water in order to keep microorganisms in the system happy and healthy. The data in *groundwater.csv* is a sample of pH is measured on a logarithmic scale from 0 to 14 and bicarbonate levels are measured in parts per million (ppm)

First, let's import the data set:

```
groundwater_dataset <- read.csv("groundwater.csv")
# View the summary
summary(groundwater_dataset)
```

```
##        pH         Bicarbonate
##  Min.   :6.700   Min.   : 35.0
##  1st Qu.:7.300   1st Qu.:107.0
##  Median :7.600   Median :147.0
##  Mean   :7.662   Mean   :142.8
##  3rd Qu.:8.000   3rd Qu.:186.5
##  Max.   :8.800   Max.   :262.0
```

Parametric

a) Use a linear model and parametric method to determine if there is a relationship between bicarbonate levels and pH in the water. (Still check and comment on assumptions, but use a parametric method regardless!)

Step 1 : Hypothesis & Assumptions

Step 1a: Hypothesis

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

H_0 : There is no significant relationship between bicarbonate levels and pH in the water

vs.

H_A : There is a significant relationship between bicarbonate levels and pH in the water

To determine if there is a relationship between bicarbonate levels and pH in the water, we need to fit in a linear model first:

To check the relationship between bicarbonate levels and pH in the water, we need to fit a linear model. This is given by:

$$pH = \beta_o + \beta_1 * bicarbonate$$

```
linear_model <- lm(pH ~ Bicarbonate, data = groundwater_dataset)
linear_model
```

```
##
## Call:
## lm(formula = pH ~ Bicarbonate, data = groundwater_dataset)
##
## Coefficients:
## (Intercept)  Bicarbonate
##    8.097595    -0.003052
```

Hence, the intercept, β_0 is **8.097595** and the slope β_1 is **-0.003052**. The model is:

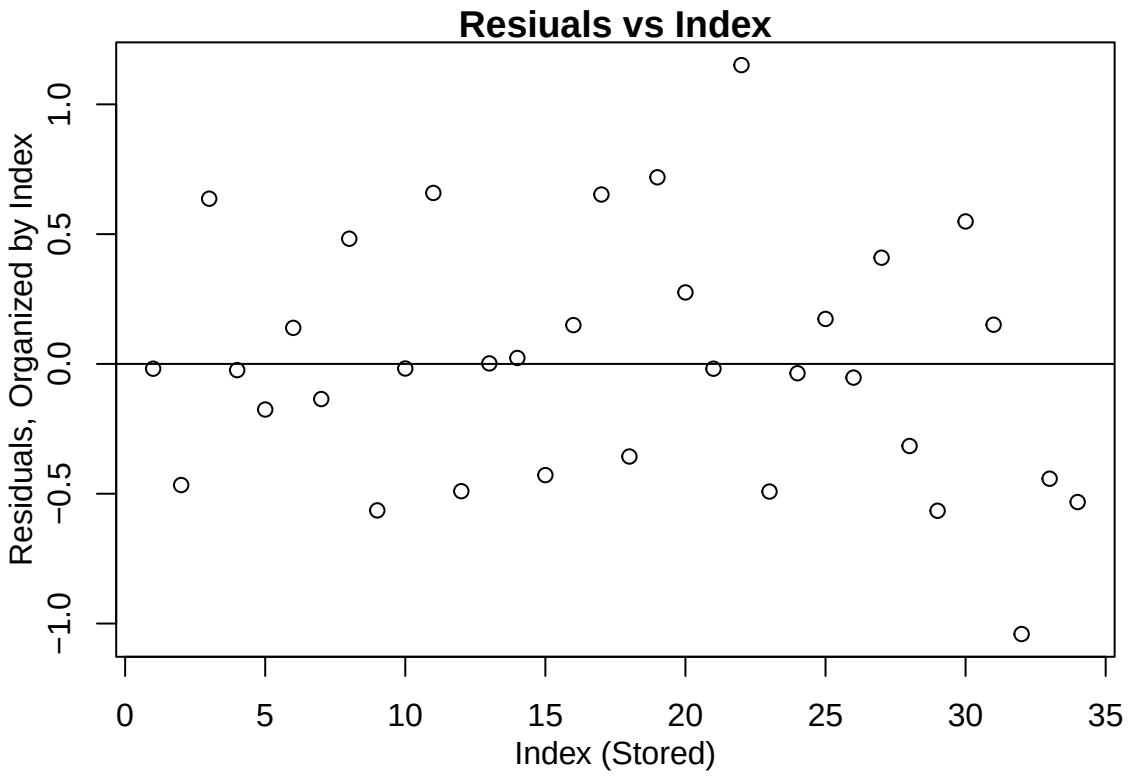
$$pH = 8.097595 + (-0.003052) * Bicarbonate$$

Step 1b: Assumptions We need to check the assumptions of:

- **Independence** : The data was randomly sampled and the sample size is less than 10% of the population size. Thus, the condition is satisfied.
- **Linearity** : There should be a linear relationship between the two variables, that is, pH and Bicarbonate.
- **Normality** : The residuals should be nearly normal distributed and show normality.
- **Variance** : The variability across the groups should be about equal (equal variance of residuals).

Now, we will check the assumptions. Firstly, let's check for independence:

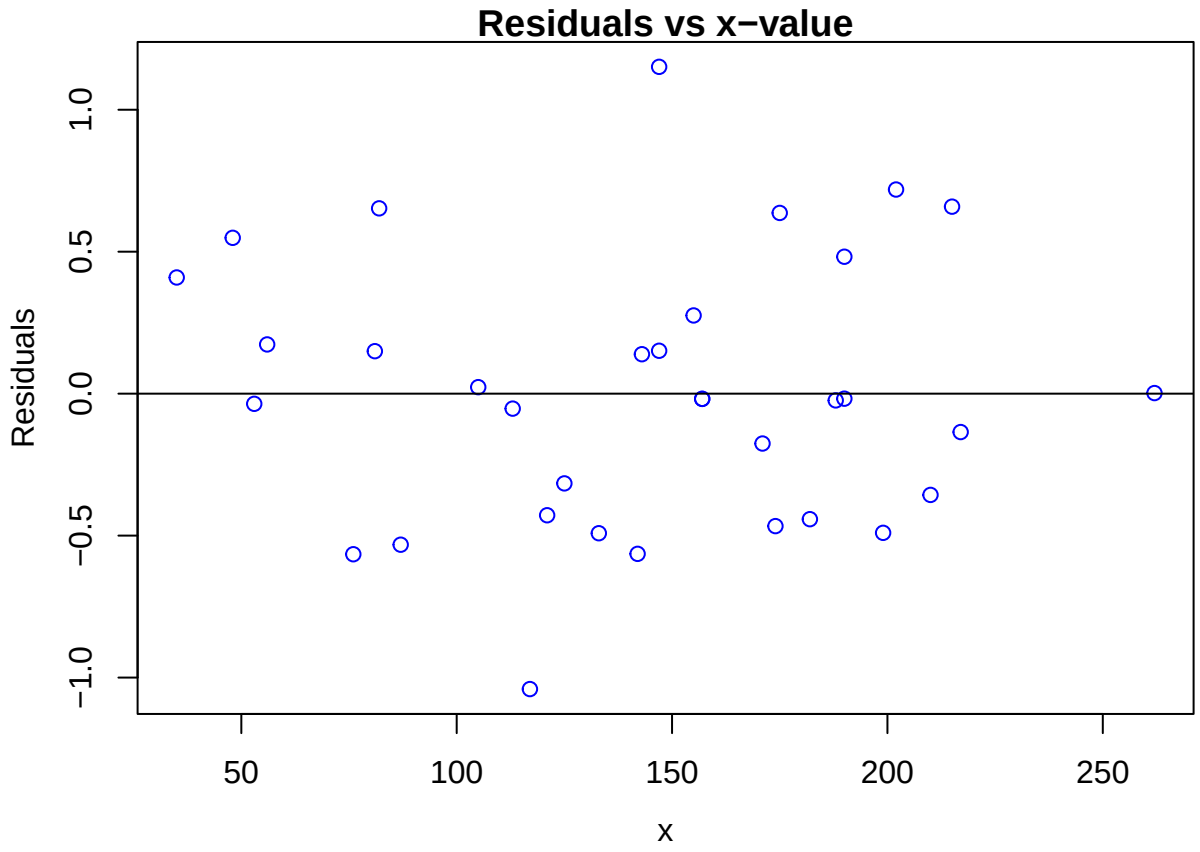
```
par(mar = c(3,3,1, 1), mgp = c(2, 1, 0))
plot(x = 1:length(linear_model$residuals), y = linear_model$residuals
     , xlab = "Index (Stored)",ylab = "Residuals, Organized by Index", main = "Resiuals vs Index")
abline(h = 0)
```



It is independent, as no serial correlation can be observed.

Now, we will check the assumptions of linearity among pH and Bicarbonate:

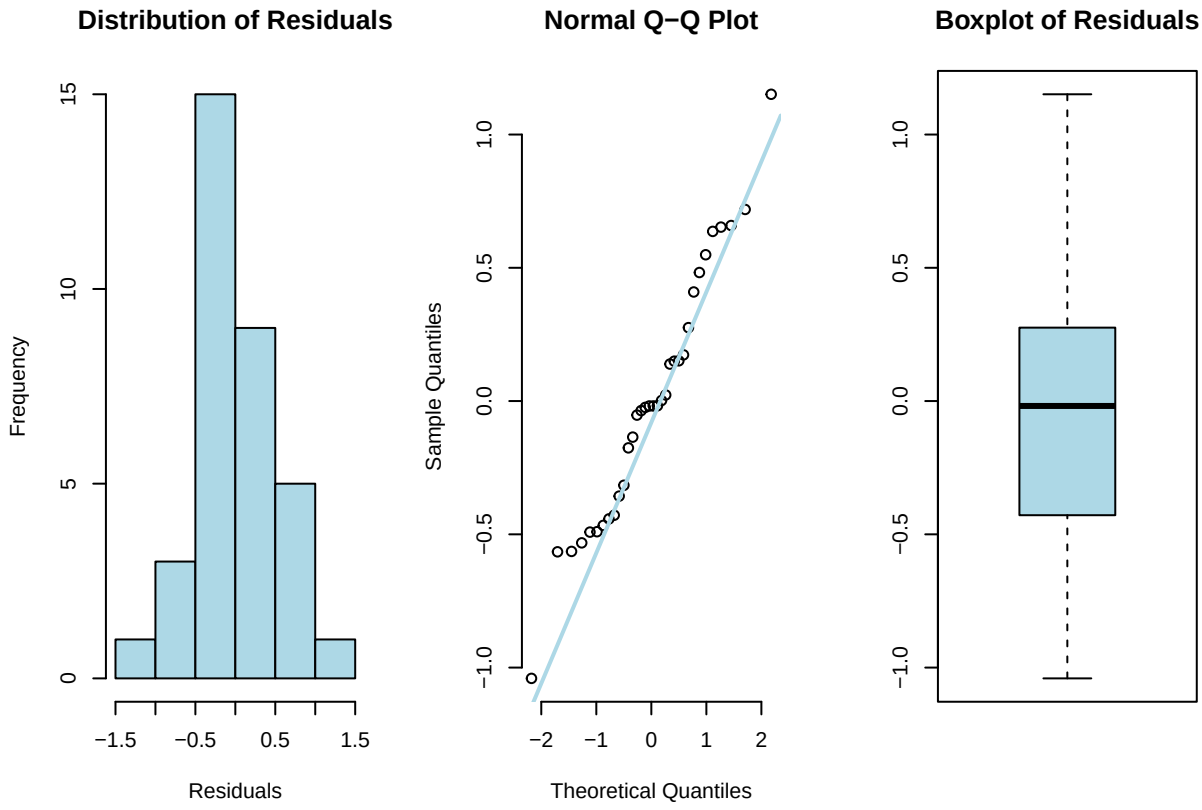
```
par(mar = c(4,4,1, 1), mgp = c(2.5, 1, 0))
plot(x = groundwater_dataset$Bicarbonate, y = linear_model$residuals
     , xlab = "x", ylab = "Residuals", col = "blue"
     , main = "Residuals vs x-value")
abline(h = 0)
```



Here, it can be seen that the residuals vs x are nearly linear. Hence, the assumption is met.

Now, we will check the assumptions of normality:

```
par(mfrow=c(1,3))
hist(linear_model$residuals, main = "Distribution of Residuals"
     , col = "lightblue", xlab = "Residuals")
qqnorm(linear_model$residuals, pch = 1, frame = FALSE);
qqline(linear_model$residuals, col = "lightblue"
         , lwd = 2);
boxplot(linear_model$residuals,col = "lightblue"
          , main = "Boxplot of Residuals")
```



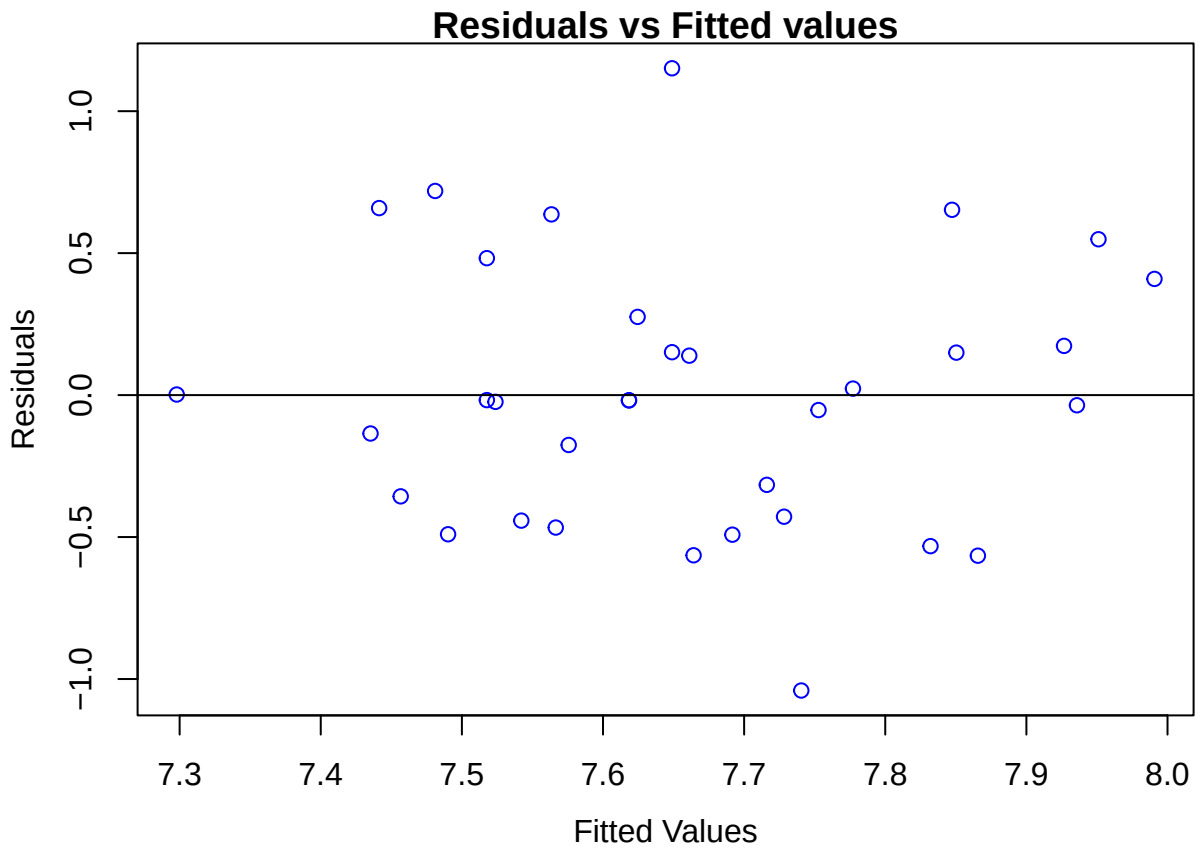

```
shapiro.test(linear_model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  linear_model$residuals
## W = 0.97231, p-value = 0.5281
```

Here, it can be seen that the distribution is nearly normal and is confirm by Shapiro-Wilk's test, as p-value > 0.05. Hence, the assumption of normality is met.

Finally, we will check the assumptions of variability:

```
par(mar = c(4,4,1, 1), mgp = c(2.5, 1, 0))
plot(x = linear_model$fitted.values, y = linear_model$residuals
     , col = "blue"
     , xlab = "Fitted Values", ylab = "Residuals"
     , main = "Residuals vs Fitted values")
abline(h = 0)
```



The data across the line are nearly constant, as the fitted values increase the residuals are constant. Hence, the condition of variability is met.

Step 2 : Test Statistic and p-value

Now, we can proceed with the parametric approach to test the hypothesis.

```
summary(linear_model)
```

```
##
## Call:
## lm(formula = pH ~ Bicarbonate, data = groundwater_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04049 -0.41037 -0.01841  0.24995  1.15107
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.097595   0.228714  35.405   <2e-16 ***
## Bicarbonate -0.003052   0.001495  -2.042   0.0495 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.479 on 32 degrees of freedom
```

```
## Multiple R-squared:  0.1153, Adjusted R-squared:  0.08762
## F-statistic: 4.169 on 1 and 32 DF,  p-value: 0.04948
```

Here, we get the test statistic as **4.169** and p-value as **0.04948**.

Step 3 : Statistical Decision

Therefore, we **reject** Null Hypothesis, H_0 because p-value is less than α, level of significance (that is, p-value < 0.05).

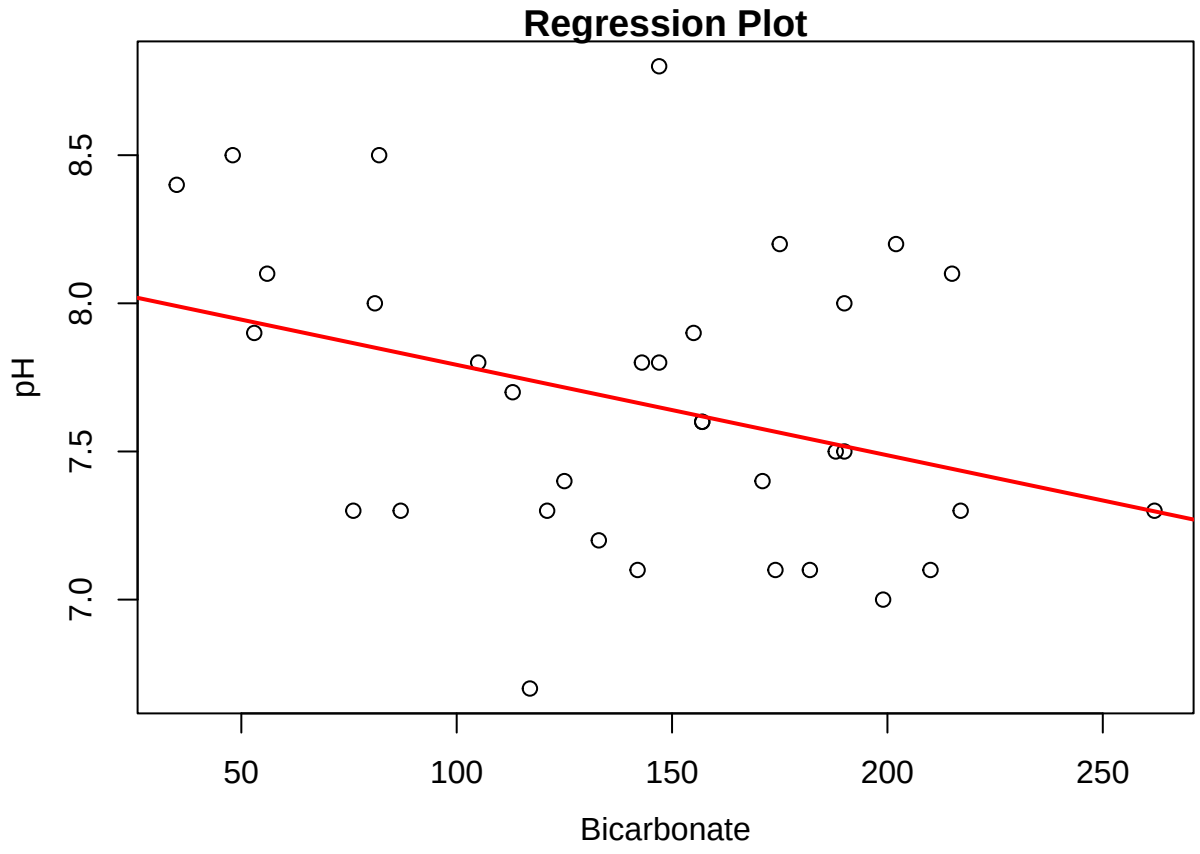
Step 4 : Conclusion

There is enough evidence to support the alternative hypothesis, and therefore we conclude that there exists a significant relationship between bicarbonate levels and pH of the water, is accurate.

b) Plot the data and regression line. Include appropriate labels and title. Bonus for including confidence and prediction intervals on this plot1 [2 marks]

The regression plot is given below:

```
par(mar = c(4,4,1,1), mgp = c(2.5, 1, 0))
plot(groundwater_dataset$Bicarbonate, groundwater_dataset$pH
     , main = "Regression Plot", xlab = "Bicarbonate", ylab = "pH")
abline(linear_model, col="red", lw=2)
```



c) Provide a 95% parametric confidence interval for the true slope parameter in the linear model.

A 95% parametric confidence interval for the true slope parameter in the linear model is

```
confint(linear_model, level = 0.95, parm = "Bicarbonate")
```

```
##                      2.5 %       97.5 %
## Bicarbonate -0.006096981 -7.33803e-06
```

Here, we get the true slope confidence intervals as **-0.006096981** and **-7.33803e-06**.

Intervals [-0.006096981, -7.33803e-06].

Non-Parametric

d) Use a linear model and non-parametric method to determine if there is a relationship between bicarbonate levels and pH in the water.

Step 1 : Hypothesis & Assumptions

Step 1a: Hypothesis

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

H_0 : There is no significant relationship between bicarbonate levels and pH in the water

vs.

H_A : There is a significant relationship between bicarbonate levels and pH in the water

Step 1b: Assumptions

We need to check the assumptions of (assumptions is done in Part a) uisng diagnostic plots):

- **Independence** : The data was randomly sampled and the sample size is less than 10% of the population size. Thus, the condition is satisfied.
- **Linearity** : There should be a linear relationship between the two variables, that is, pH and Bicarbonate.
- **Normality** : The residuals should be nearly normal distributed and show normality.
- **Variance** : The variability across the groups should be about equal (equal variance of residuals).

Step 2 : Test Statistic and p-value

Now, we can proceed with the non-parametric approach and calculate the test statistic and p-value.

```
# simulation
set.seed(0758054)
nsim <- 5000
coef_sim <- numeric(nsim)
sim_df <- groundwater_dataset

for (i in 1:nsim){
```

```
  sim_df$pH <- sample(groundwater_dataset$pH, size = nrow(groundwater_dataset), replace = FALSE)
  sim_lm <- lm(pH ~  Bicarbonate, data = sim_df)
  coef_sim[i] <- sim_lm$coefficients[2]
}
```

Now, calculating the p-value:

```
distance <- abs(linear_model$coefficients[2] - 0)
pval <- (length(which(abs(coef_sim - 0) >= distance)) + 1) / (nsim + 1)
pval
```

```
## [1] 0.04939012
```

Hence, the p-value we get using the non-parametric approach is **0.04939**.

Step 3 : Statistical Decision

Therefore, we **reject** Null Hypothesis, H_0 because p-value is less than α, level of significance (that is, p-value < 0.05).

Step 4 : Conclusion

There is enough evidence to support the alternative hypothesis, and therefore we conclude that there exists a significant relationship between bicarbonate levels and pH of the water, is accurate.

e) Provide a 95% non-parametric confidence interval for the true slope parameter in the linear model.

A 95% non-parametric confidence interval for the true slope parameter in the linear model, we will use bootstrapping.

```
library(boot)
coefficient_wrapper <- function(df2, index){
  lin_model <- lm(Bicarbonate ~ pH, data = df2[index, ])
  return(lin_model$coefficients[2])
}

# Slope Parameter
coefficient_bs <- boot(groundwater_dataset, statistic = coefficient_wrapper, R = 5000)
boot.ci(coefficient_bs, conf = 0.95, type = "bca")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = coefficient_bs, conf = 0.95, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   (-78.93,  -0.15 )
## Calculations and Intervals on Original Scale
```

Hence, the confidence intervals for the true slope parameter for 95% is **-78.93** and **-0.15**.
Intervals [-78.93, -0.15]

Quesiton 3 - Fawning Over Models

Required Data: *antelope.csv*

The data provided in the *antelope.csv* file contains:

- **the spring fawn count (divided by 100)**
- **the size of the adult antelope population (divided by 100)**
- **annual precipitation (in inches)**
- **winter severity index (1=mild, 5=severe)**

Firstly, let's importing the dataset

```
antelope_dataset <- read.csv("antelope.csv")
# View the summary
summary(antelope_dataset)
```

```
##       fawn           adult           precip         severity
##  Min.   :1.900   Min.   :6.800   Min.   :10.60   Min.   :1.000
##  1st Qu.:2.075   1st Qu.:7.725   1st Qu.:11.10   1st Qu.:2.000
##  Median :2.350   Median :8.600   Median :11.90   Median :3.000
##  Mean   :2.525   Mean   :8.450   Mean   :12.04   Mean   :2.875
##  3rd Qu.:2.975   3rd Qu.:9.300   3rd Qu.:12.75   3rd Qu.:3.250
##  Max.   :3.400   Max.   :9.700   Max.   :14.10   Max.   :5.000
```

a) Fit the full model (i.e., fawn ~ adult + precip + severity);

- **Provide diagnostic plots, the summary, model formula, and brief discussion of the model fit.**

Let's fit the multiple regression model, $fawn = \beta_o + \beta_1 adult + \beta_2 precip + \beta_3 severity$

where,

β_0 = Intercept,

β_1 = Slope of *adult*,

β_2 = Slope of *precip*, and

β_3 = Slope of *severity*

```
multi_reg_model <- lm(fawn ~ adult + precip + severity, data = antelope_dataset)
summary(multi_reg_model)
```

```
##
## Call:
## lm(formula = fawn ~ adult + precip + severity, data = antelope_dataset)
##
## Residuals:
##        1        2        3        4        5        6        7        8
```

```
## -0.11533 -0.02661  0.09882 -0.11723  0.02734 -0.04854  0.11715  0.06441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.92201    1.25562  -4.716   0.0092 **
## adult        0.33822    0.09947   3.400   0.0273 *
## precip       0.40150    0.10990   3.653   0.0217 *
## severity     0.26295    0.08514   3.089   0.0366 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1209 on 4 degrees of freedom
## Multiple R-squared:  0.9743, Adjusted R-squared:  0.955
## F-statistic: 50.52 on 3 and 4 DF,  p-value: 0.001229
```

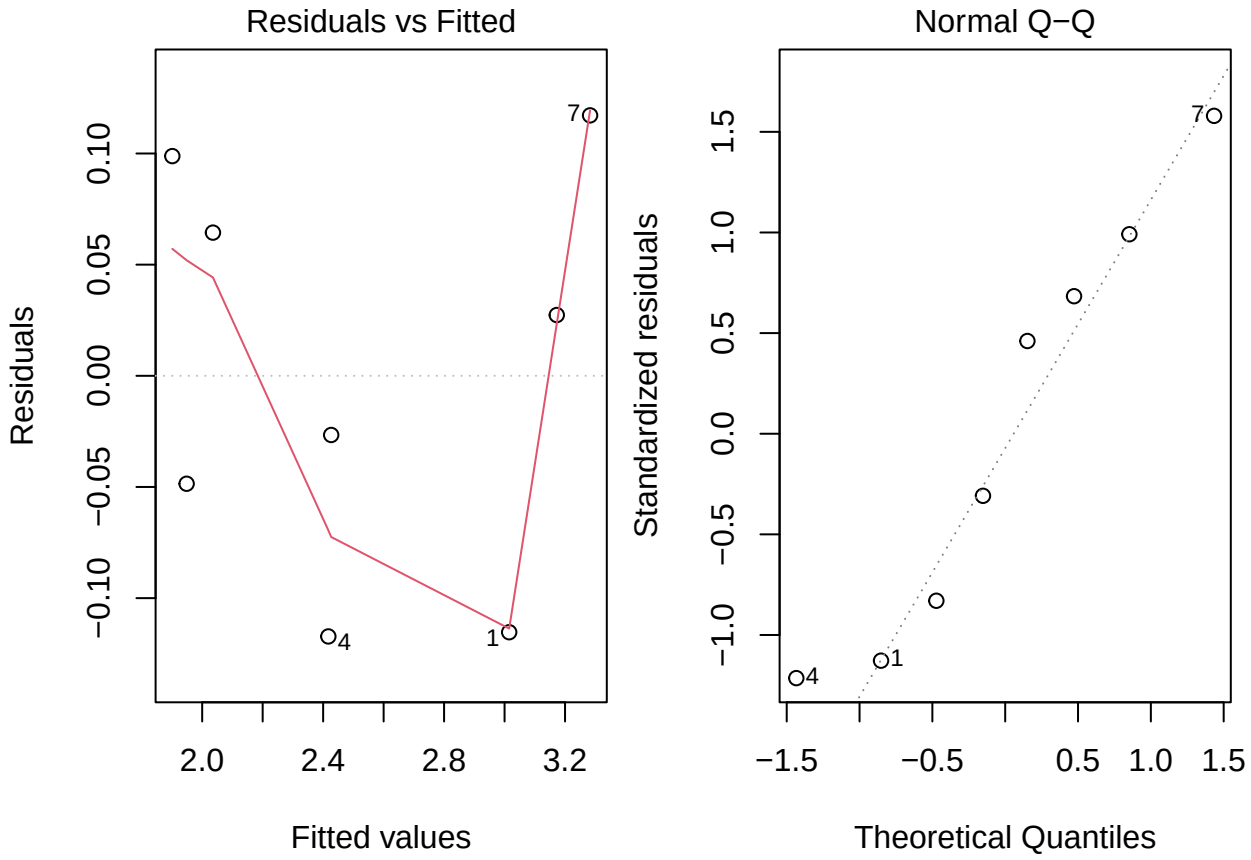
Model Formula

$$fawn = -5.92201 + (0.33822)adult + (0.40150)precip + (0.26295)severity$$

Diagnostic Plots

The Plot of Residual vs Fitted and Normal Q-Q is given below:

```
par(mfrow = c(1, 2), mar = c(4,4,1.5, 0.5))
plot(multi_reg_model, which = 1:2)
```



Summary and Model Fit

The intercept of the model (β_0) is **-5.92201** and the slope of *adult* (β_1), *precip* (β_2) and *severity* (β_3) is **0.33822**, **0.40150** and **0.26295** respectively. The diagnostic plots show nearly normal distribution across the residuals and there is nearly constant variance across the residuals. The Residual Standard error is **0.1209** and Adjusted R-squared is **0.955**, indicating that the model explains a large proportion of the variability in the data.

b) Fit the quadratic polynomial model: fawn ~ adult + adult2

- **Provide diagnostic plots, the summary, model formula, and brief discussion of the model fit.**
- **Also include the model plotted on top of the original data used in this model.**

Let's fit the polynomial regression model, $fawn = \beta_o + \beta_1 adult + \beta_2 adult^2$

where,

β_0 = intercept,

β_1 = Slope of *adult*, and

β_2 = Slope of $adult^2$

Let's fit the model,

```
ploy_reg_model <- lm(fawn ~ adult + I(adult^2), data = antelope_dataset)
summary(ploy_reg_model)
```

```
##
## Call:
## lm(formula = fawn ~ adult + I(adult^2), data = antelope_dataset)
##
## Residuals:
##        1        2        3        4        5        6        7        8
##  0.04501 -0.06238  0.06814 -0.03554 -0.04676 -0.05437  0.04451  0.04140
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.90135    1.88703   6.837 0.001022 **
## adult       -3.07710    0.46078  -6.678 0.001138 **
## I(adult^2)   0.21577    0.02778   7.767 0.000566 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06428 on 5 degrees of freedom
## Multiple R-squared:  0.9909, Adjusted R-squared:  0.9873
## F-statistic: 272.8 on 2 and 5 DF,  p-value: 7.86e-06
```

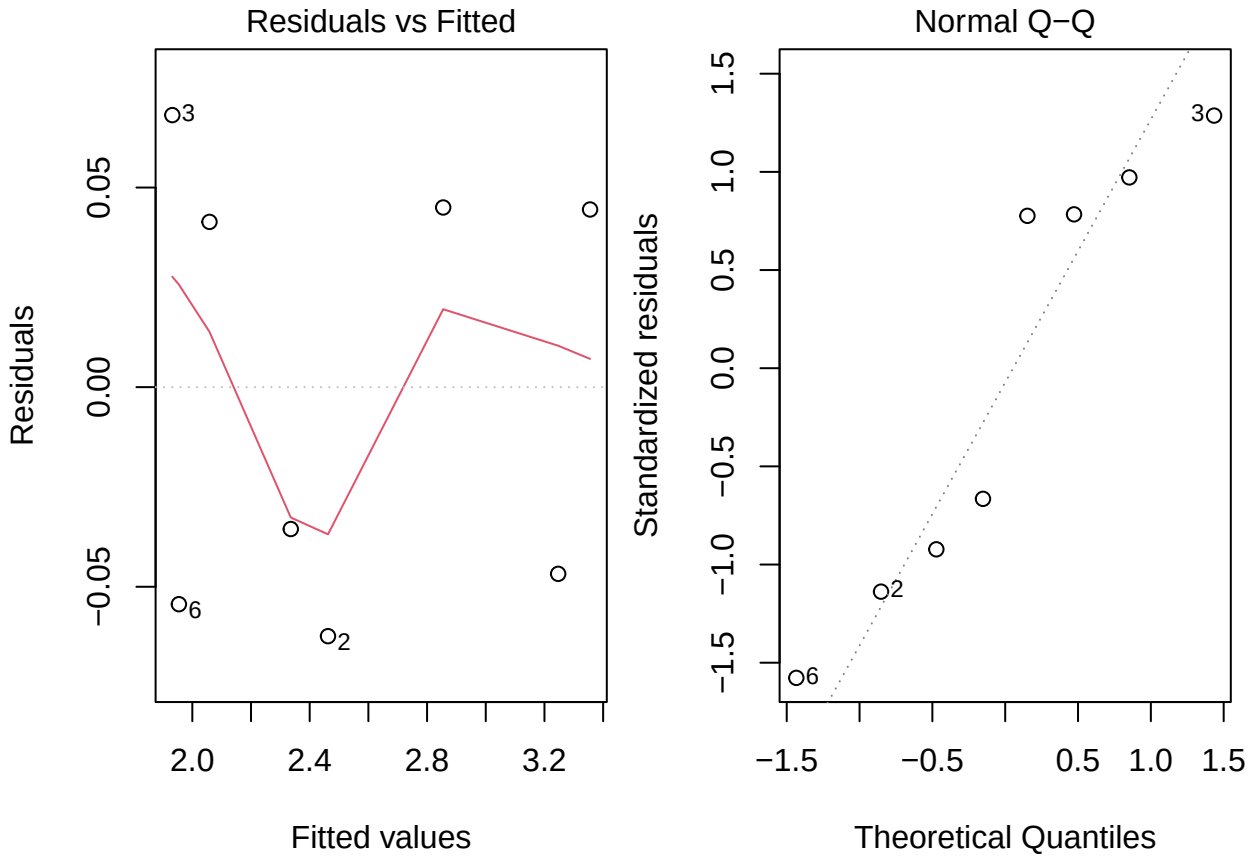
Model Formula

$$fawn = 12.90135 + (-3.07710)adult + (0.21577)adult^2$$

Diagnostic Plots

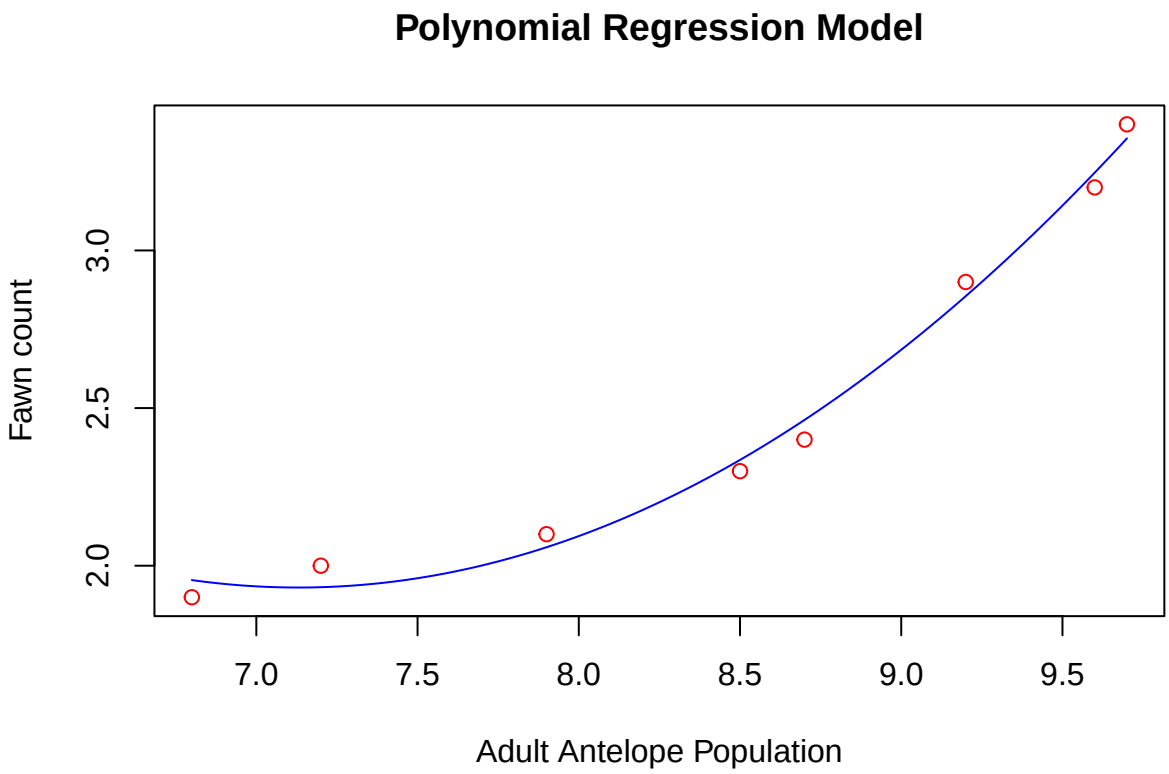
The Plot of Residual vs Fitted and Normal Q-Q is given below:

```
par(mfrow = c(1, 2), mar = c(4,4,1.5, 0.5))
plot(ploy_reg_model, which = 1:2)
```



Plotting the model on top of data

```
plot(fawn ~ adult, data = antelope_dataset
     , xlab = "Adult Antelope Population", ylab = "Fawn count", col= "red"
     , main = "Polynomial Regression Model")
curve(predict(ploy_reg_model, newdata = data.frame(adult = x, adult2 = x^2))
      , add = TRUE, col = "Blue")
```



Summary and Model Fit

The intercept of the model (β_0) is **12.90135** and the slope of *adult* (β_1) and $adult^2$ (β_1) is **-3.07710** and **0.21577** respectively. The diagnostic plots show nearly normal distribution across the residuals and there is nearly constant variance across the residuals. The Residual Standard error is **0.06428** and Adjusted R-squared is **0.9873**, indicating that the model explains a large proportion of the variability in the data.

c) Which model is better? Why?

The **Polynomial Model** is a better model than the Multiple Regression model because it has a **higher adjusted R-squared** value, indicating that it explains more of the variability in the data. Large Adjusted R-squared implies more explanatory power. Also, the polynomial model has **less predictor variables** (only "adult") as compared to Multiple regression model ("adult", "precip" and "severity") which makes it easier to interpret and use. Besides, the polynomial model has **lower Residual standard error** indicating a better model fit than multiple regression model. Lastly, the **small p-value** of Polynomial model indicate a better significance of the model as compared to the Multiple regression model.