Assignment 2 - Probability and Statistics in Data Analysis

Here is the solution to each question from the exercises provided in assignment 2 of the Probability and Statistics in Data Analysis course

Exercise 1

Question a

Provide a 99% confidence interval for Cholesterol values

Solution

99% Confidence Interval for Cholesterol values: 180.6816 185.5204

Question b

Provide a 95% confidence interval for Cholesterol values after receiving drug A and B, respectively

Solution

95% Confidence Interval for Cholesterol values after recieving the Drug A: 178.6626 183.3694

95% Confidence Interval for Cholesterol values after recieving the Drug B: 182.4306 187.9414

Question c

Provide a 90% confidence interval for the mean difference in Cholesterol values after receiving drug A and drug B, respectively

Solution

90% Confidence Interval for the mean difference in Cholesterol values after receiving drug A and drug B: -7.164971 -1.175029

Question d

Examine the following hypothesis test:

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A < \mu_B$$

where μ_A and μ_B are the mean Cholesterol values after receiving drug A and B, respectively. The level of significance is α = 0.05

Solution

Welch Two Sample t-test

data: drug_a_data Cholesterol and drug_b_data Cholesterol
t = -2.3126, df = 95.66, p-value = 0.01144
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -1.175029
sample estimates:
mean of x mean of y
181.016 185.186

Since the p-value=0.01144 is less than 0.05 the null hypothesis is rejected and the mean cholesterol level for drug A is significantly less than for drug B

Question e

Provide a hypothesis test (α = 0.01) for the equality of variances of Glucose levels after receiving drug A and drug B, respectively

Solution

F test to compare two variances

data: drug_a_data Glucose and drug_b_data Glucose
F = 1.0484, num df = 49, denom df = 49, p-value = 0.8694
alternative hypothesis: true ratio of variances is not equal to 1
99 percent confidence interval:
0.4961326 2.2152165
sample estimates:
ratio of variances
1.048352

Since the p-value=0.8694 is greater than 0.01 there is no enough evidence of a difference in variances

Question f

At a significance level of 5%, test if there is a statistically significant side effect on Glucose levels

Solution

Welch Two Sample t-test

data: drug_a_data$\boxed{\text{Glucose and drug_b_data}}$Glucose
t = 1.5297, df = 97.945, p-value = 0.1293
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.7431617 5.7431617
sample estimates:
mean of x mean of y
92.118 89.618

Since the p-value=0.1293 is greater than 0.05, there is no enough evidence for statistically significant side effect on Glucose levels between the drugs

Question g

Provide a 95% confidence interval for the proportion of volunteers who had Myalgia symptoms

Solution

95% Confidence Interval for the proportion of volunteers who had Myalgia symptoms:
0.02860529 0.13891973

Question h

Test if the proportion of volunteers who had Myalgia symptoms is statistically greater than 5% at a significance level of 5%

Solution

Exact binomial test

data: count(with_myalgia_data)$nandcount(data)$n
number of successes = 7, number of trials = 100, p-value = 0.234
alternative hypothesis: true probability of success is greater than 0.05
95 percent confidence interval:
0.03331192 1.00000000
sample estimates:
probability of success
0.07

Since the p-value = 0.234 is greater than 0.05 there is no evidence the proportion exceeds 5%

Question i

Test if the drug and the presence of Myalgia symptoms are independent (α = 0.05)

Solution

Fisher's Exact Test for Count Data

data: contingency_table
p-value = 0.1117
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.003194961 1.328490415
sample estimates:
odds ratio
0.1520682

Since p-value = 0.1117 is greater than 0.05, there is no statistically significant association between Drug type and Myalgia symptoms at the α=0.05 level

Question j

Provide a 95% confidence interval for the mean difference $\mu_1 - \mu_2$, where μ_1 and μ_2 are the mean Glucose levels for volunteers with and without Myalgia symptoms, respectively

Solution

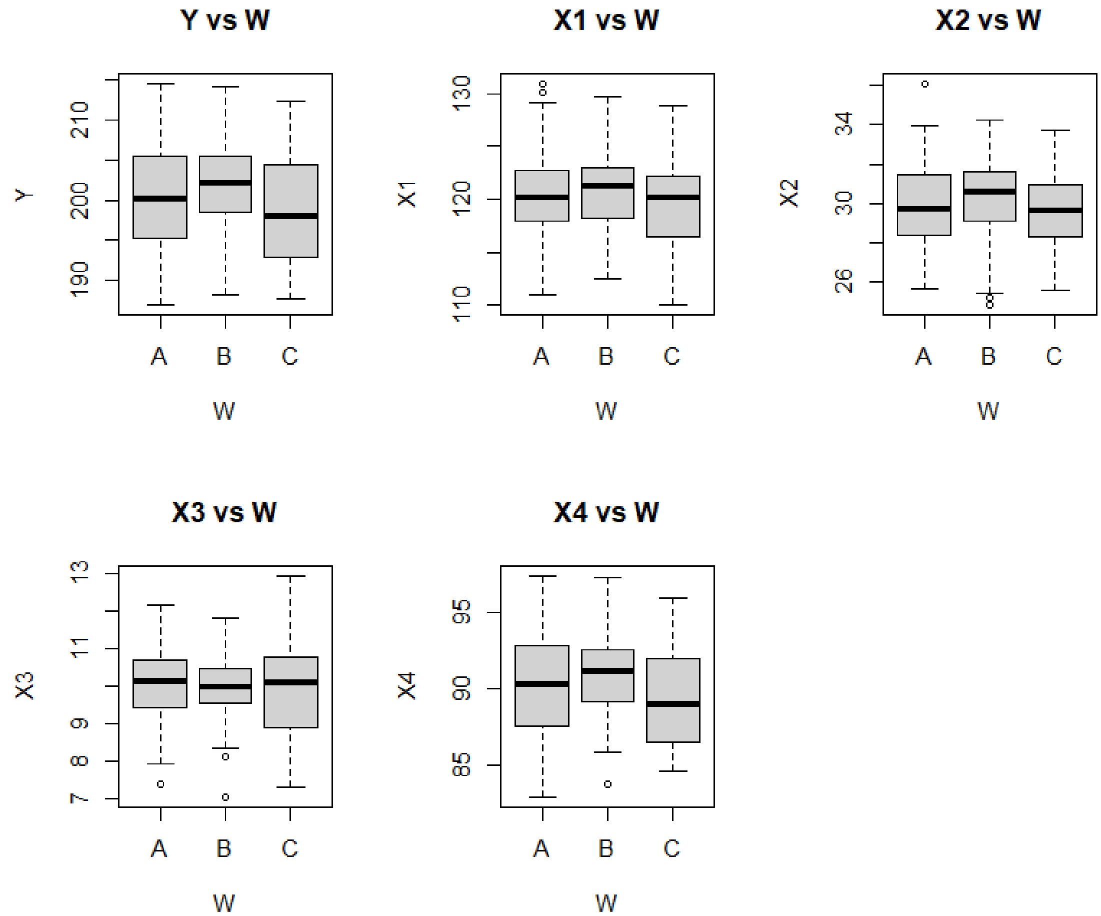
95% Confidence Interval for the mean difference in Glucose levels with and without Myalgia symptoms: -12.62472 4.73655

Exercise 2

Question a-i

Provide a graphical representation of each continuous variable versus the categorical variable

Solution



Question a-ii

Provide the ANOVA output

Solution

Anova on Y-W

```
             Df Sum Sq Mean Sq F value Pr(>F)  
W             2    333  166.71   4.352 0.0141 *
Residuals   197   7546   38.31                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Because the p-value < 0.05, significant differences exist between groups

Anova on X1-W

```
             Df Sum Sq Mean Sq F value Pr(>F)  
W             2   76.3   38.13    2.42 0.0915 .
Residuals   197 3104.1   15.76                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Because the p-value > 0.05, no significant differences exist between groups

Anova on X2-W

```
             Df Sum Sq Mean Sq F value Pr(>F)
W             2   17.0   8.489   2.079  0.128
Residuals   197  804.3   4.083
```

Because the p-value > 0.05, no significant differences exist between groups

Anova on X3-W

```
             Df Sum Sq Mean Sq F value Pr(>F)
W             2   0.28  0.1397   0.133  0.876
Residuals   197 207.24  1.0520
```

Because the p-value > 0.05, no significant differences exist between groups

Anova on X4-W

```
             Df Sum Sq Mean Sq F value Pr(>F)  
W             2   75.8   37.89   4.171 0.0168 *
Residuals   197 1789.6    9.08                 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the p-value < 0.05, significant differences exist between groups

Question a-iii

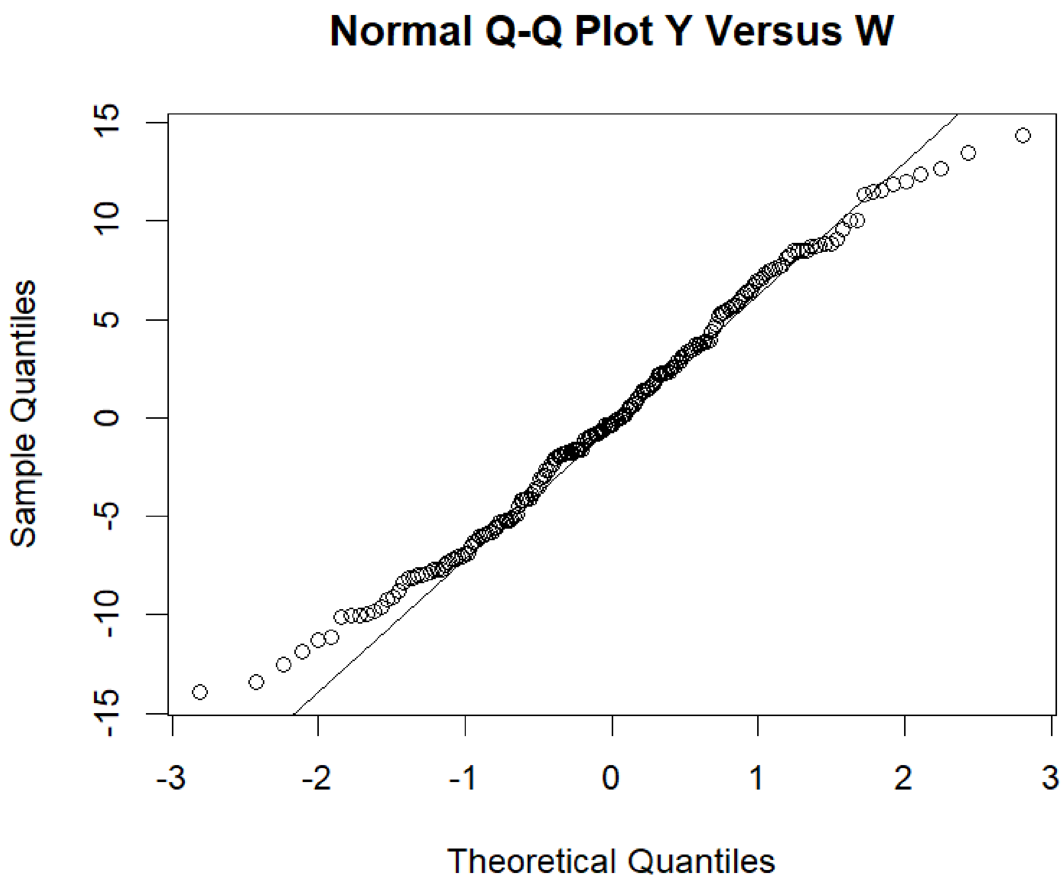
Check the assumptions

Solution

1. Normality Assumption

Y versus W

- Q-Q plot

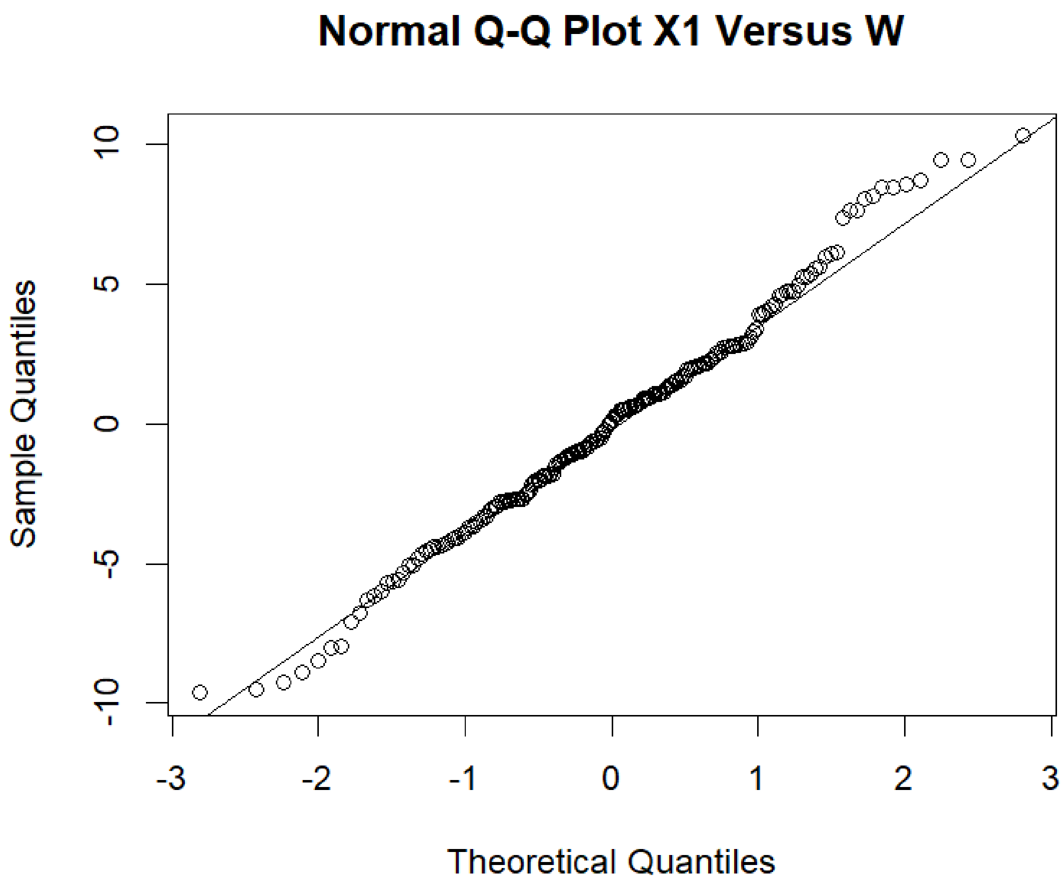


- Shapiro-Wilk normality test

```
        Shapiro-Wilk normality test

data:  residuals(anova_Y)
W = 0.98923, p-value = 0.1374
```

<u>X1 versus W</u>

- Q-Q plot

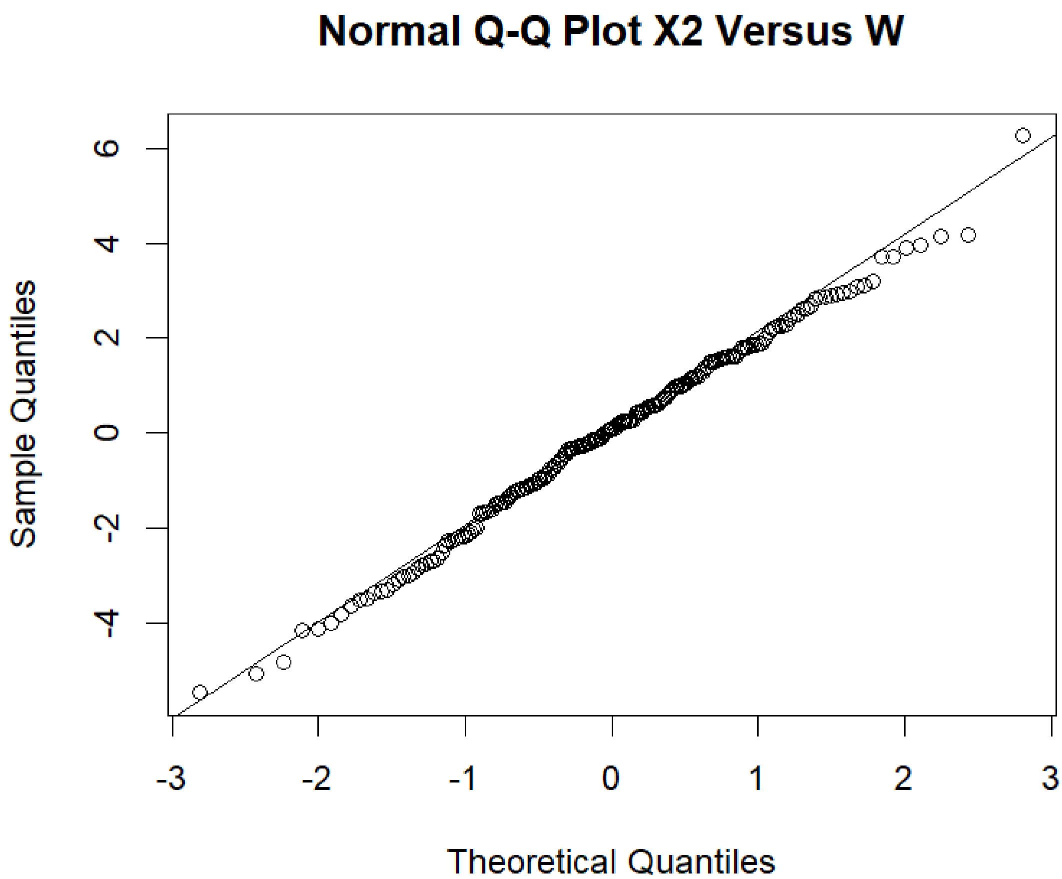


- Shapiro-Wilk normality test

```
        Shapiro-Wilk normality test

data:  residuals(anova_X1)
W = 0.99123, p-value = 0.268
```

X2 versus W

- Q-Q plot

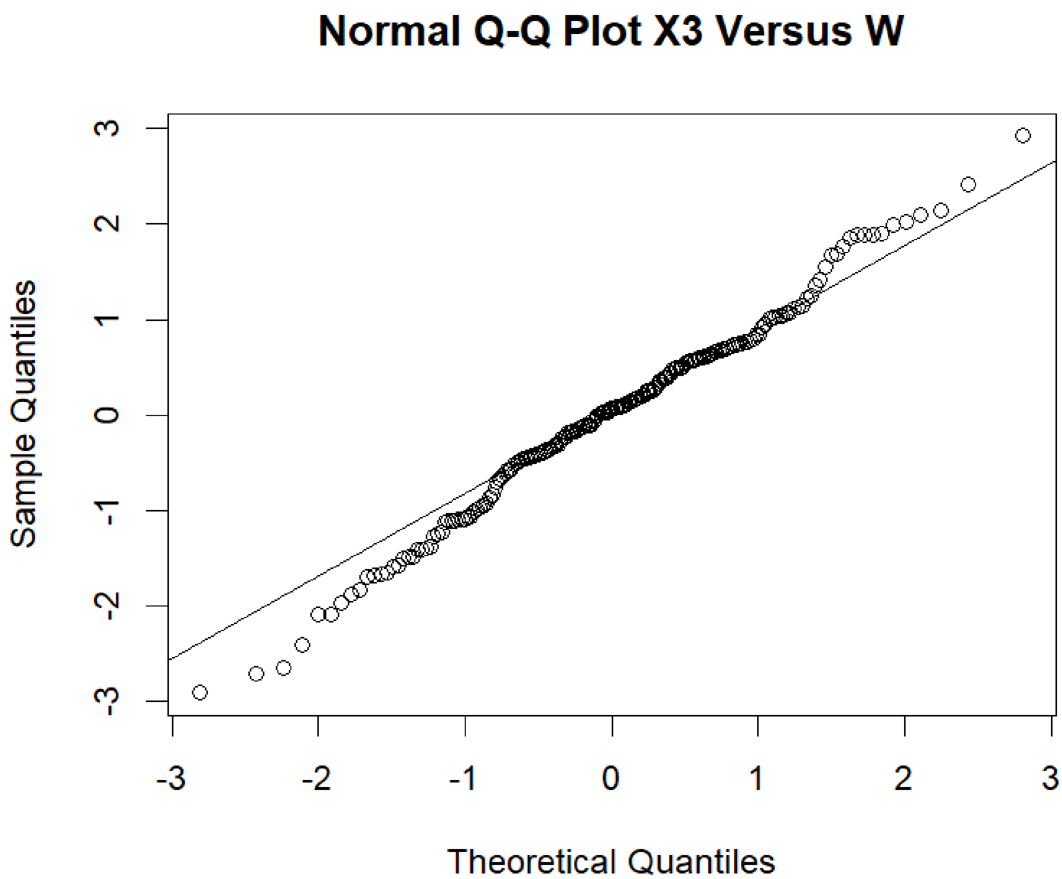


- Shapiro-Wilk normality test

```
	Shapiro-Wilk normality test

data:  residuals(anova_X2)
W = 0.99539, p-value = 0.8049
```

X3 versus W

- Q-Q plot

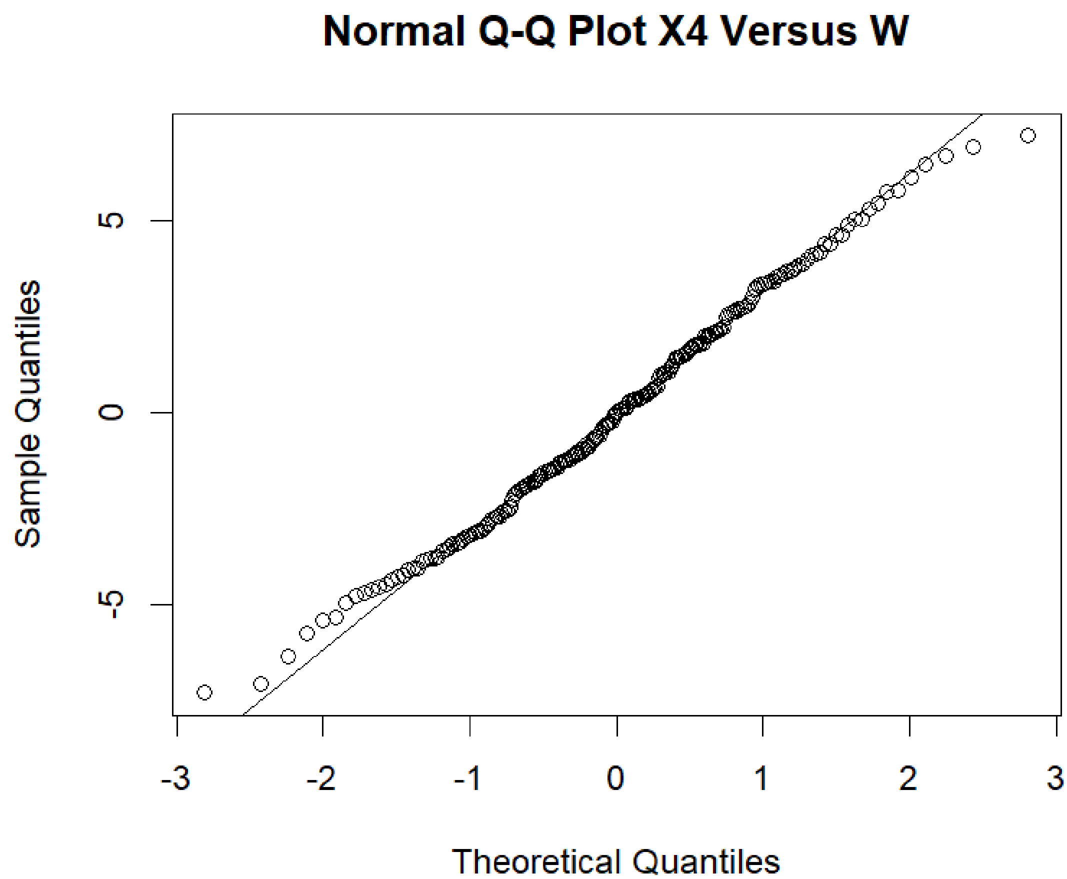


- Shapiro-Wilk normality test

```
        Shapiro-Wilk normality test

data:  residuals(anova_X3)
W = 0.99108, p-value = 0.2555
```

X4 versus W

- Q-Q plot



- Shapiro-Wilk normality test

```
        Shapiro-Wilk normality test

data:  residuals(anova_X4)
W = 0.99272, p-value = 0.4243
```

We observe that the null hypothesis of normality of the residuals holds for all the combinations since the significance level α=0.05 is less than the p-value in all cases

2. <u>Homogeneity of Variances Assumption</u>

<u>Y versus W</u>

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   2  3.6897 0.02672 *
      197
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because, the p-value is less than 0.05, the assumption of homogeneity of variances is violated

X1 versus W

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  1.0945 0.3367
      197
```

Because, the p-value is greater than 0.05, the assumption of homogeneity of variances is not violated

X2 versus W

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2   1.081 0.3412
      197
```

Because, the p-value is greater than 0.05, the assumption of homogeneity of variances is not violated

X3 versus W

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   2  7.4498 0.0007605 ***
      197
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because, the p-value is less than 0.05, the assumption of homogeneity of variances is violated

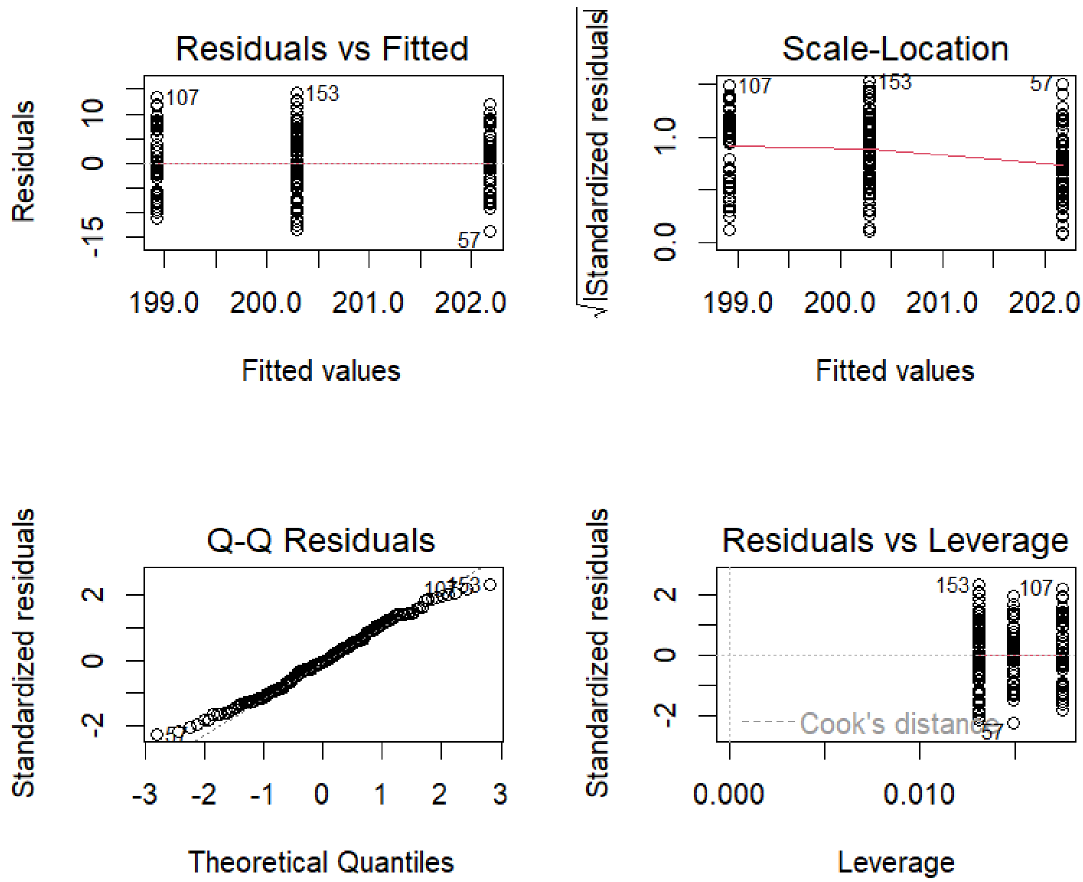
X4 versus W

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  2.2203 0.1113
      197
```

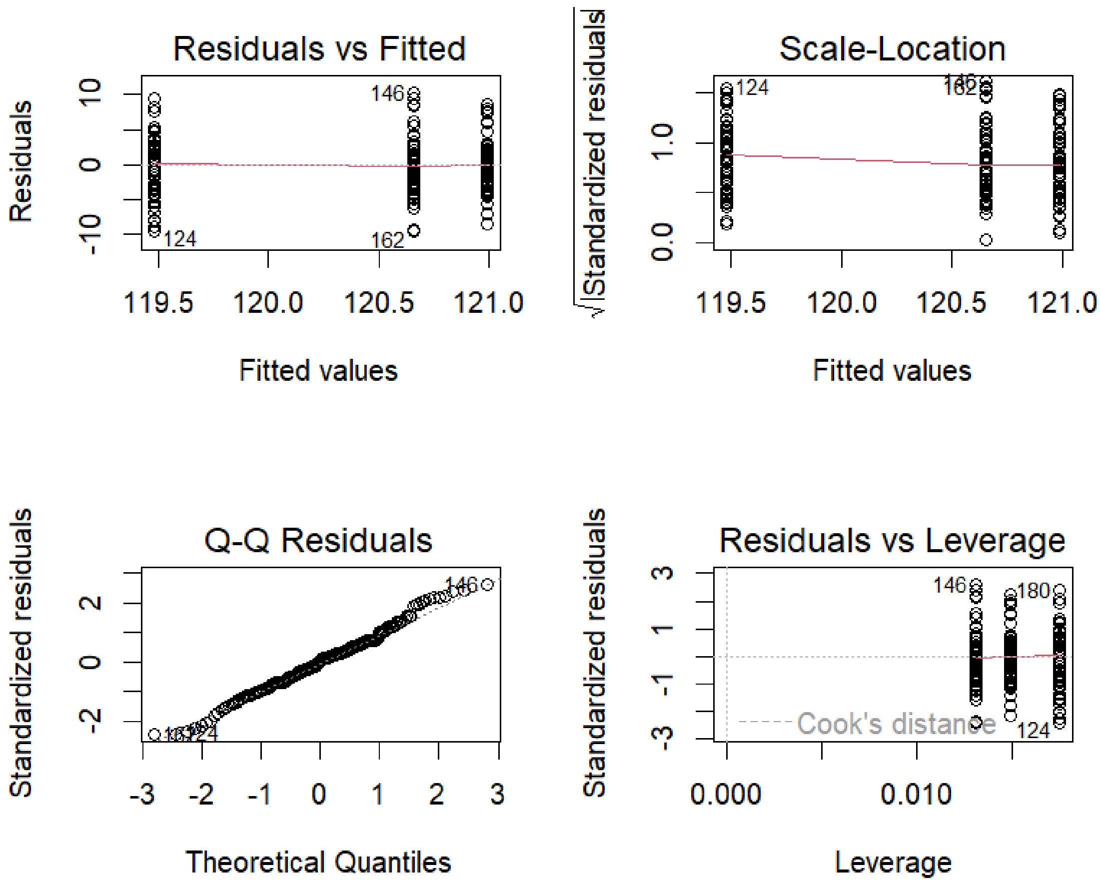
Because, the p-value is greater than 0.05, the assumption of homogeneity of variances is not violated

Provide also diagnostic plots to verify the above interpretations:

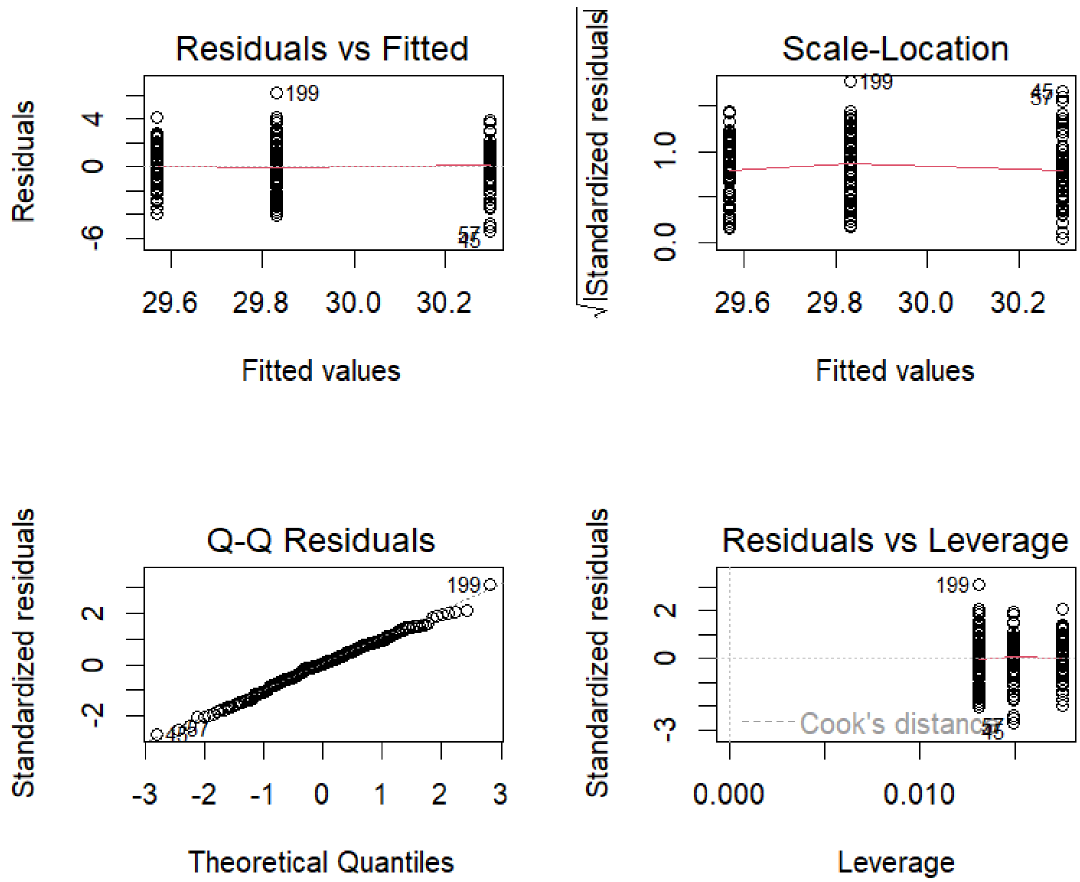
Y versus W


Residuals vs Fitted
Residuals
Fitted values
Scale-Location
√|Standardized residuals|
Fitted values
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

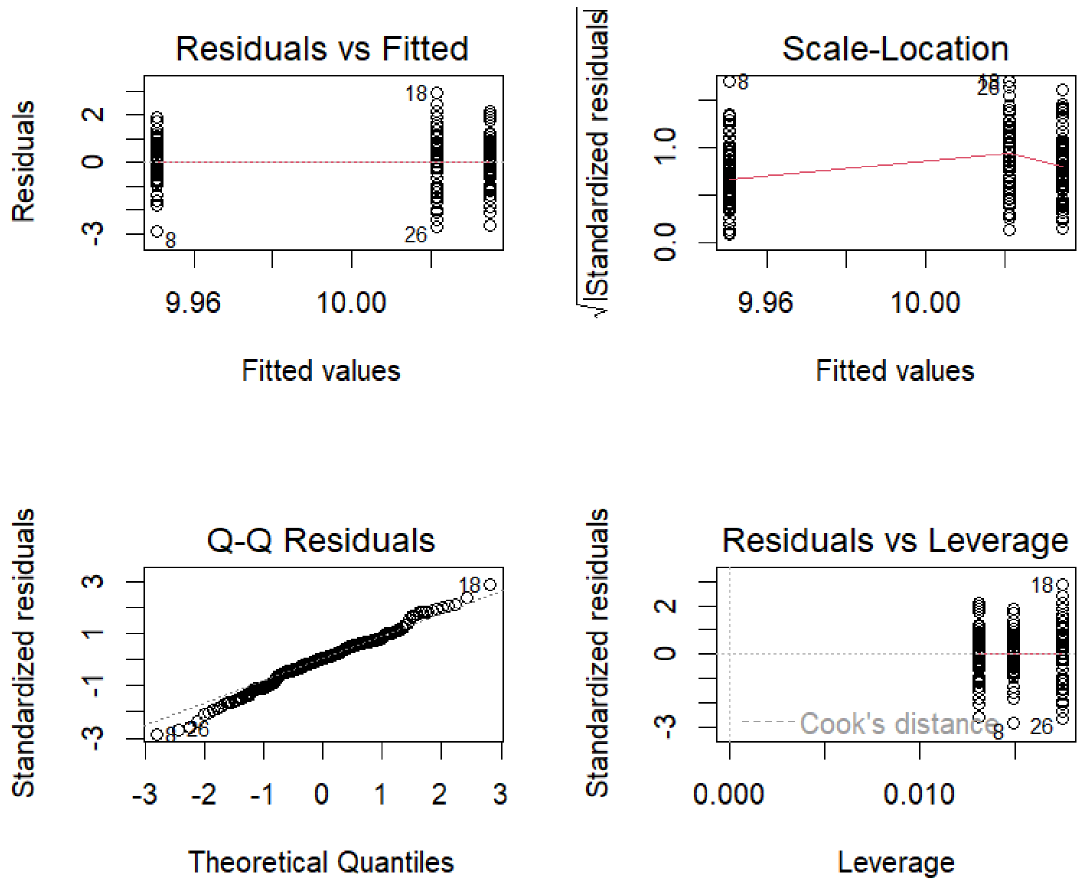
X1 versus W


Residuals vs Fitted
Residuals
Fitted values
146
124
162
Scale-Location
√|Standardized residuals|
Fitted values
124
146
162
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
146
124
Residuals vs Leverage
Standardized residuals
Leverage
146
180
124
Cook's distance

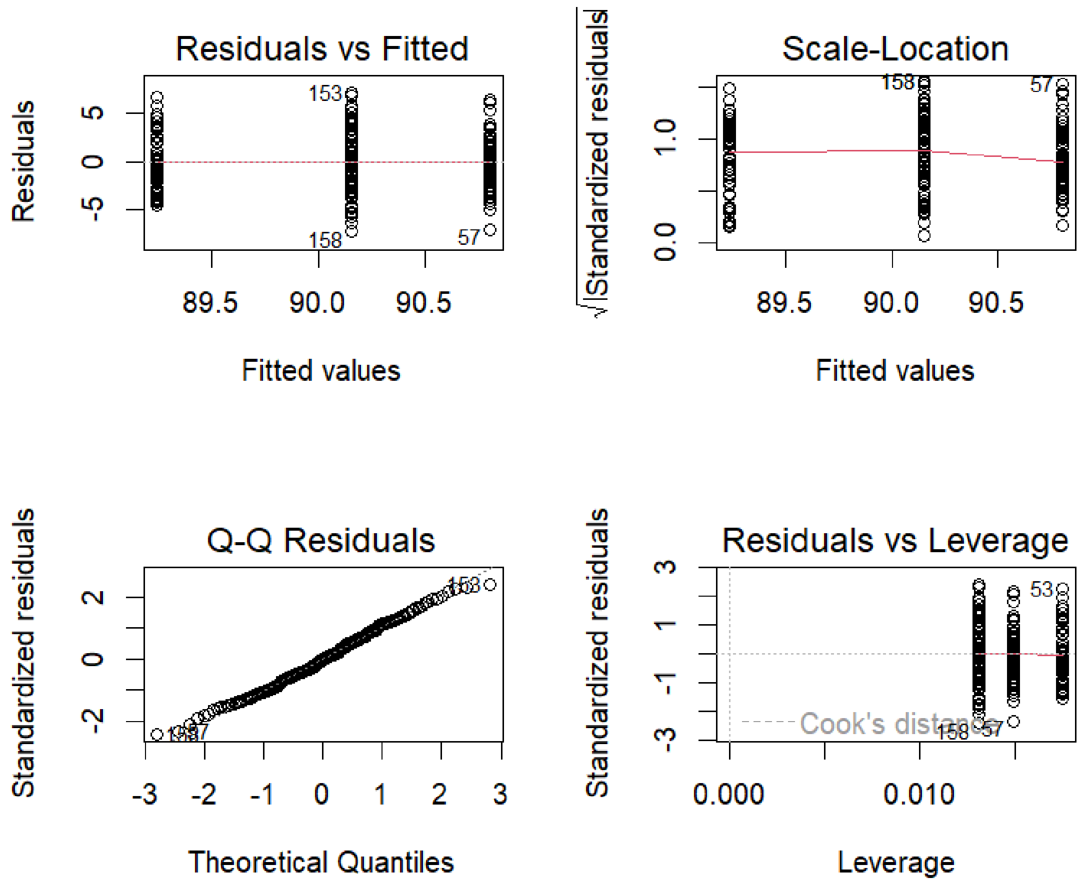
X2 versus W


Residuals vs Fitted
Residuals
Fitted values
199
57
45
Scale-Location
√|Standardized residuals|
Fitted values
199
45
57
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
199
45
57
Residuals vs Leverage
Standardized residuals
Leverage
199
Cook's distance
57
45

X3 versus W


Residuals vs Fitted
Residuals
Fitted values
Scale-Location
√|Standardized residuals|
Fitted values
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

X4 versus W

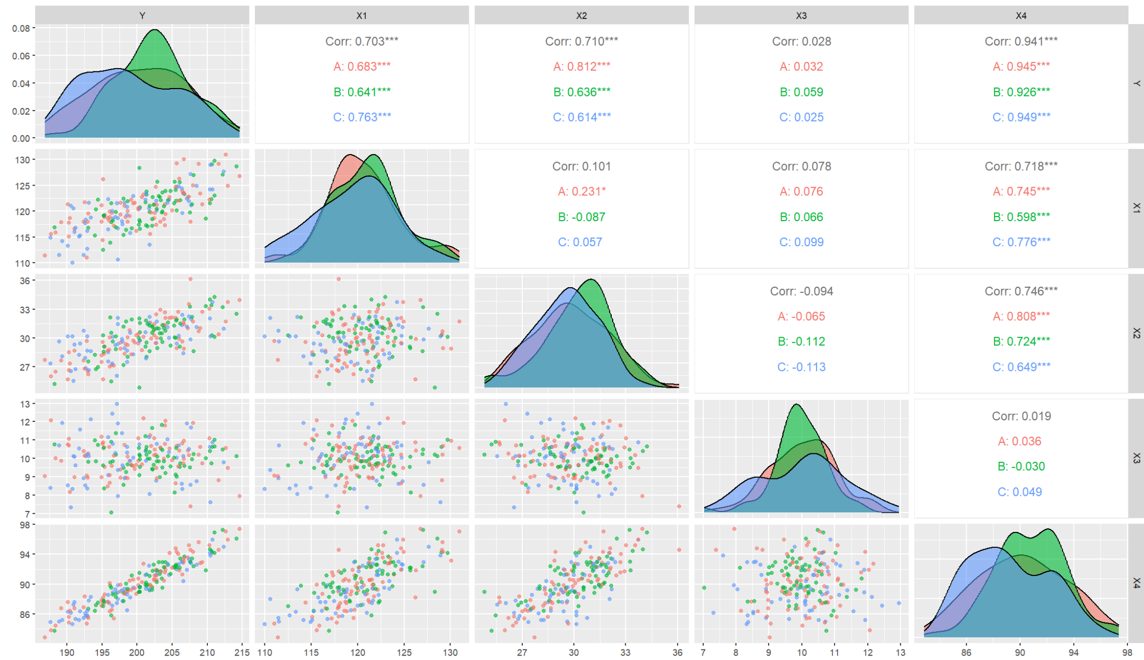


Question b

Provide a scatter-plot matrix of Y, X1, X2, X3, and X4, annotating the different levels of W in each plot using a different color

Solution

Below we can see the scatter-plot matrix where shows the correlations between variables in the upper triangle, displays the scatter plots in the lower triangle and the density plots in the diagonal



Question c

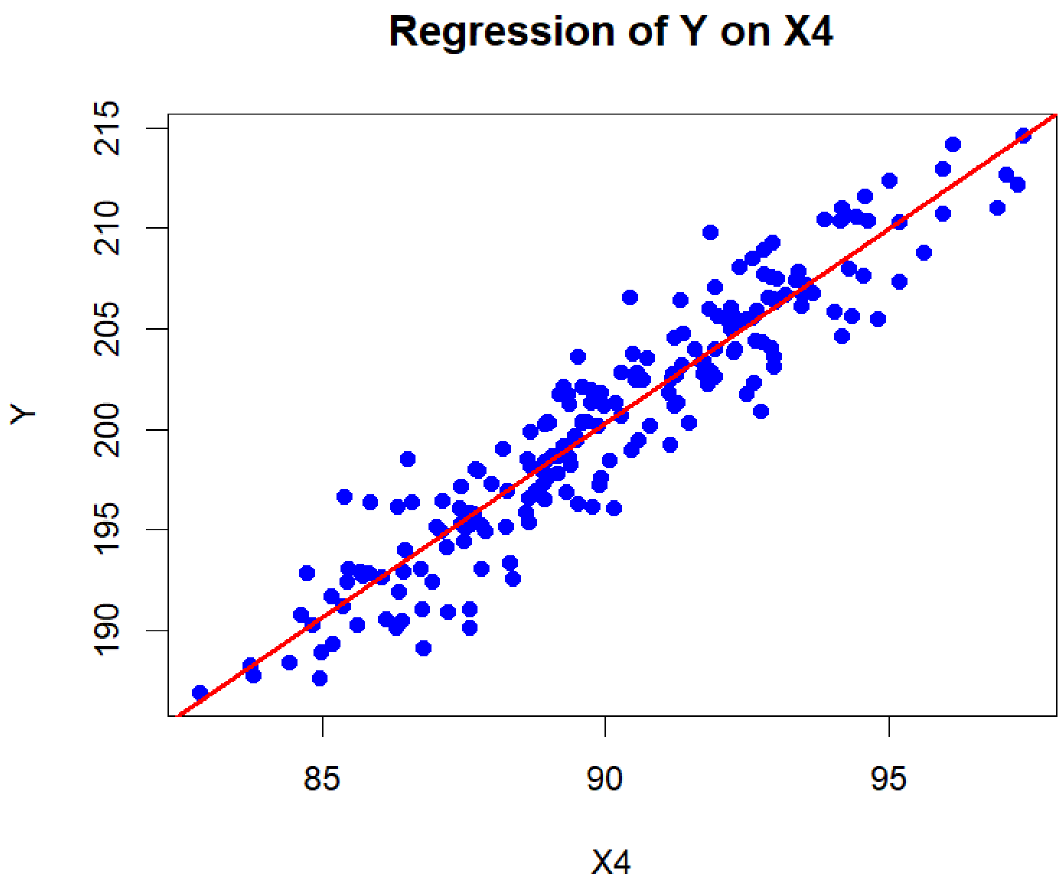
Run the regression model of Y on X4

Solution

- Regression model

```
Call:
lm(formula = Y ~ X4, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5133 -1.3818  0.1039  1.4803  5.9044

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.1973     4.4449   5.894  1.6e-08 ***
X4            1.9347     0.0493  39.243  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.129 on 198 degrees of freedom
Multiple R-squared:  0.8861,	Adjusted R-squared:  0.8855
F-statistic:  1540 on 1 and 198 DF,  p-value: < 2.2e-16
```

- Regression line



Question d

Run the regression model of Y on all the remaining variables (X1, X2, X3, X4, W), including the non-additive terms (i.e., interactions of the continuous predictors with the categorical variable)

Solution

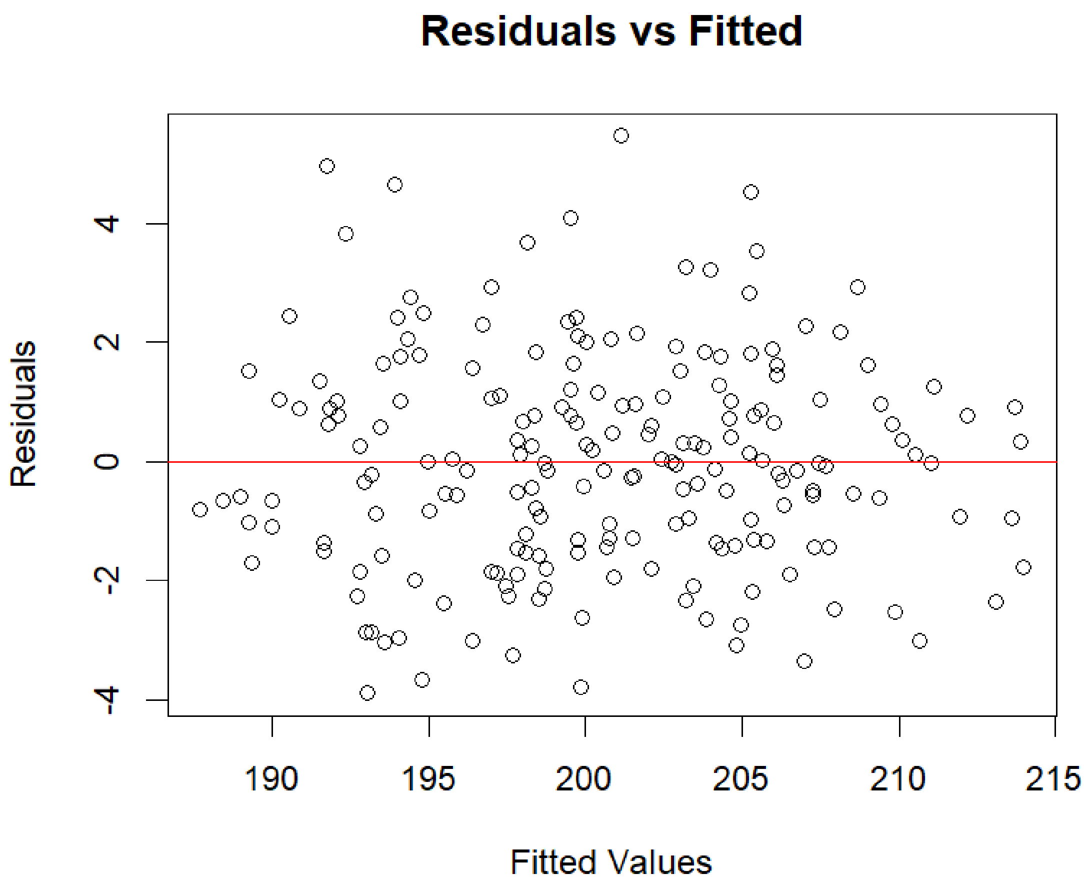
```
Call:
lm(formula = Y ~ (X1 + X2 + X3 + X4) * W, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8807 -1.3656 -0.0337  1.0723  5.4653 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  28.3612     7.1589   3.962 0.000106 ***
X1            1.1682     0.2570   4.545 9.90e-06 ***
X2            2.7008     0.5276   5.119 7.64e-07 ***
X3            0.3221     0.2313   1.393 0.165391    
X4           -0.5859     0.5015  -1.168 0.244184    
WB           -8.2392    11.6561  -0.707 0.480544    
WC          -24.4132    10.7774  -2.265 0.024658 *  
X1:WB        -0.2119     0.3432  -0.617 0.537741    
X1:WC        -0.4392     0.3618  -1.214 0.226304    
X2:WB        -0.9233     0.7186  -1.285 0.200463    
X2:WC        -1.3562     0.7368  -1.841 0.067257 .  
X3:WB         0.2838     0.3743   0.758 0.449266    
X3:WC        -0.3090     0.3076  -1.005 0.316355    
X4:WB         0.6572     0.6797   0.967 0.334848    
X4:WC         1.3478     0.7030   1.917 0.056730 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.879 on 185 degrees of freedom
Multiple R-squared:  0.9171,	Adjusted R-squared:  0.9108 
F-statistic: 146.2 on 14 and 185 DF,  p-value: < 2.2e-16
```

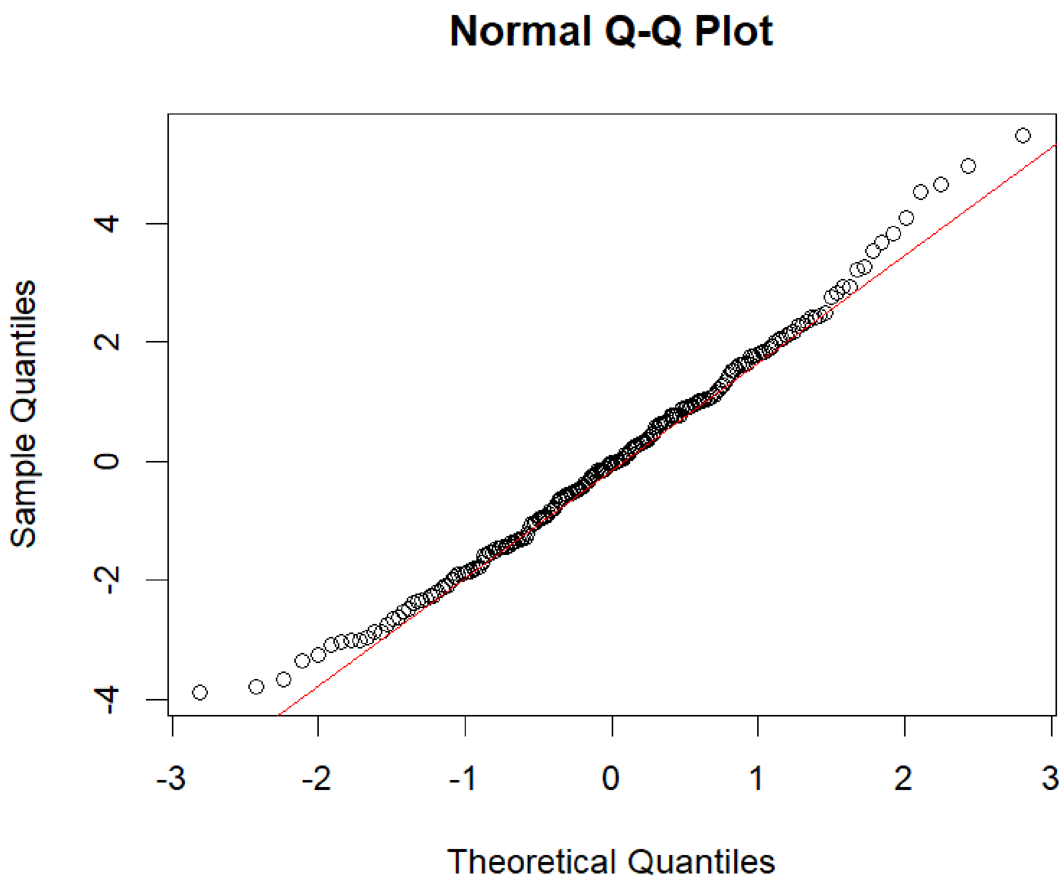
Question e

Examine the regression assumptions and provide alternatives if any of them fail

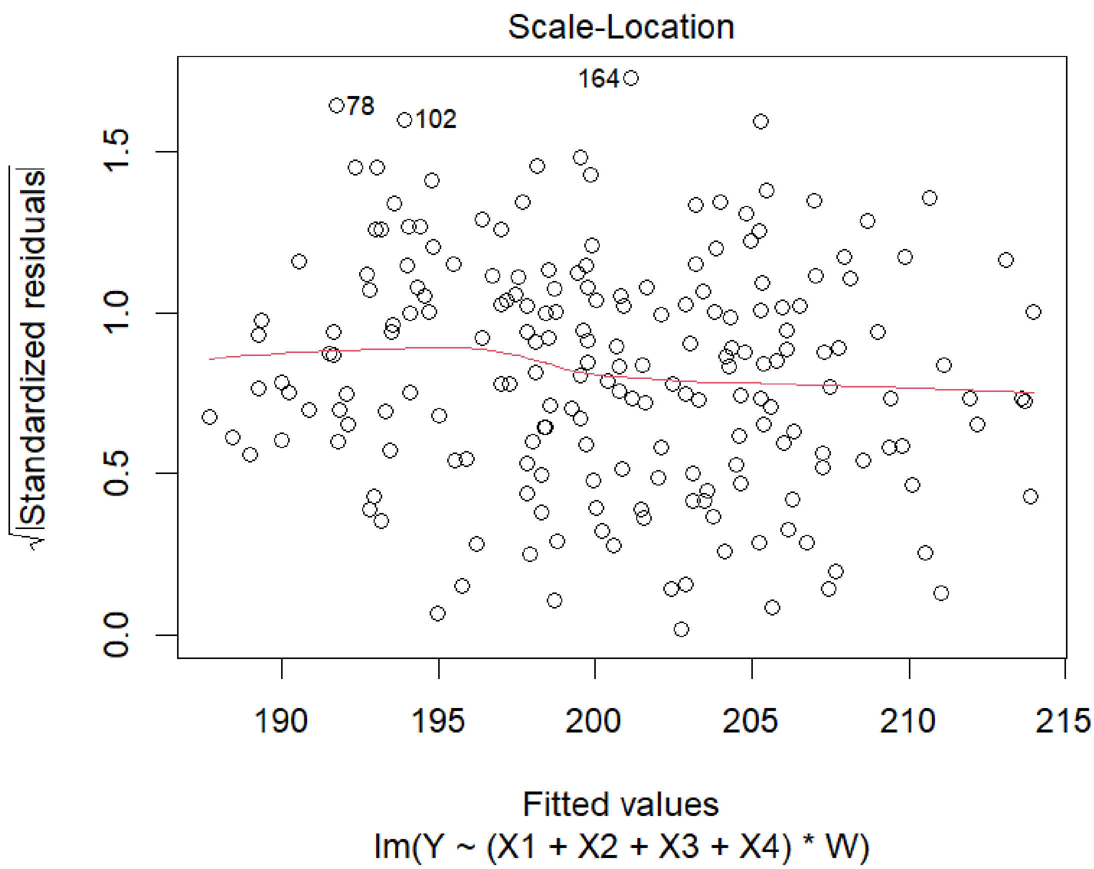
Solution

Showing some plots and a Shapiro-Wilk test to check the regression assumptions


Residuals vs Fitted
Residuals
Fitted Values
4
2
0
-2
-4
190
195
200
205
210
215




```
      Shapiro-Wilk normality test

data:  residuals(model_all)
W = 0.9907, p-value = 0.2253
```



Based on the above images all the assumptions (linearity, independence of residuals, normality of residuals and homoscedasticity of residuls) are not violated

Question f

Use the “stepwise regression” approach to examine whether you can reduce the dimension of the model

Solution

- Stepwise Regression Approach - Steps

```
Start:  AIC=266.69
Y ~ (X1 + X2 + X3 + X4) * W

       Df Sum of Sq    RSS    AIC
- X1:W  2    5.2069 658.33 264.28
- X3:W  2   10.3202 663.44 265.83
- X2:W  2   12.4535 665.58 266.47
- X4:W  2   12.9877 666.11 266.63
<none>              653.12 266.69

Step:  AIC=264.28
Y ~ X1 + X2 + X3 + X4 + W + X2:W + X3:W + X4:W

       Df Sum of Sq    RSS    AIC
- X3:W  2     8.731 667.06 262.91
<none>              658.33 264.28
- X2:W  2    20.832 679.16 266.51
+ X1:W  2     5.207 653.12 266.69
- X4:W  2    37.618 695.95 271.39
- X1    1   159.098 817.43 305.57

Step:  AIC=262.91
Y ~ X1 + X2 + X3 + X4 + W + X2:W + X4:W

       Df Sum of Sq    RSS    AIC
<none>              667.06 262.91
+ X3:W  2     8.731 658.33 264.28
- X3    1    11.587 678.65 264.36
- X2:W  2    21.134 688.20 265.15
+ X1:W  2     3.618 663.44 265.83
- X4:W  2    35.695 702.76 269.34
- X1    1   162.414 829.48 304.50
```

We can observe that the final predictors are X1 + X2 + X3 + X4 + W + X2:W + X4:W

- Summary of the final model

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + W + X2:W + X4:W, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0269 -1.2964  0.0009  1.1942  5.6151 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  30.5231     6.7564   4.518 1.10e-05 ***
X1            0.9587     0.1413   6.784 1.46e-10 ***
X2            2.2899     0.3218   7.117 2.23e-11 ***
X3            0.2439     0.1346   1.812  0.07159 .  
X4           -0.1849     0.2903  -0.637  0.52487    
WB           -7.0505    10.8716  -0.649  0.51743    
WC          -29.9678    10.0516  -2.981  0.00325 ** 
X2:WB        -0.5349     0.2384  -2.244  0.02602 *  
X2:WC        -0.4785     0.2481  -1.928  0.05531 .  
X4:WB         0.2633     0.1687   1.560  0.12036    
X4:WC         0.4966     0.1563   3.178  0.00173 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.879 on 189 degrees of freedom
Multiple R-squared:  0.9153,	Adjusted R-squared:  0.9109 
F-statistic: 204.4 on 10 and 189 DF,  p-value: < 2.2e-16
```

Question g

Using the model found in (f), provide a point estimate and a 95% confidence interval for the prediction of Y when: (X1,X2,X3,X4,W) = (120, 30, 10, 90,B)

Solution

```
       fit      lwr      upr
1 200.6604 200.1839 201.1369
```

In the above image we can see the prediction of Y and the corresponding 95% confidence interval for this

Question h

Using the cut() function, create a categorical variable (named Z) with 3 levels based on the quantiles of X4. Provide the contingency table of Z and W

Solution

Below is the related contingency table of Z and W

```
         A  B  C
Low     27 13 27
Medium  24 27 15
High    25 27 15
```

Question I

Run the parametric two-way ANOVA of Y on the categorical variables W and Z (including the interaction term). Provide the fit, examine the assumptions, and comment on the significance of the terms

Solution

Below is the two-way Anova model summary

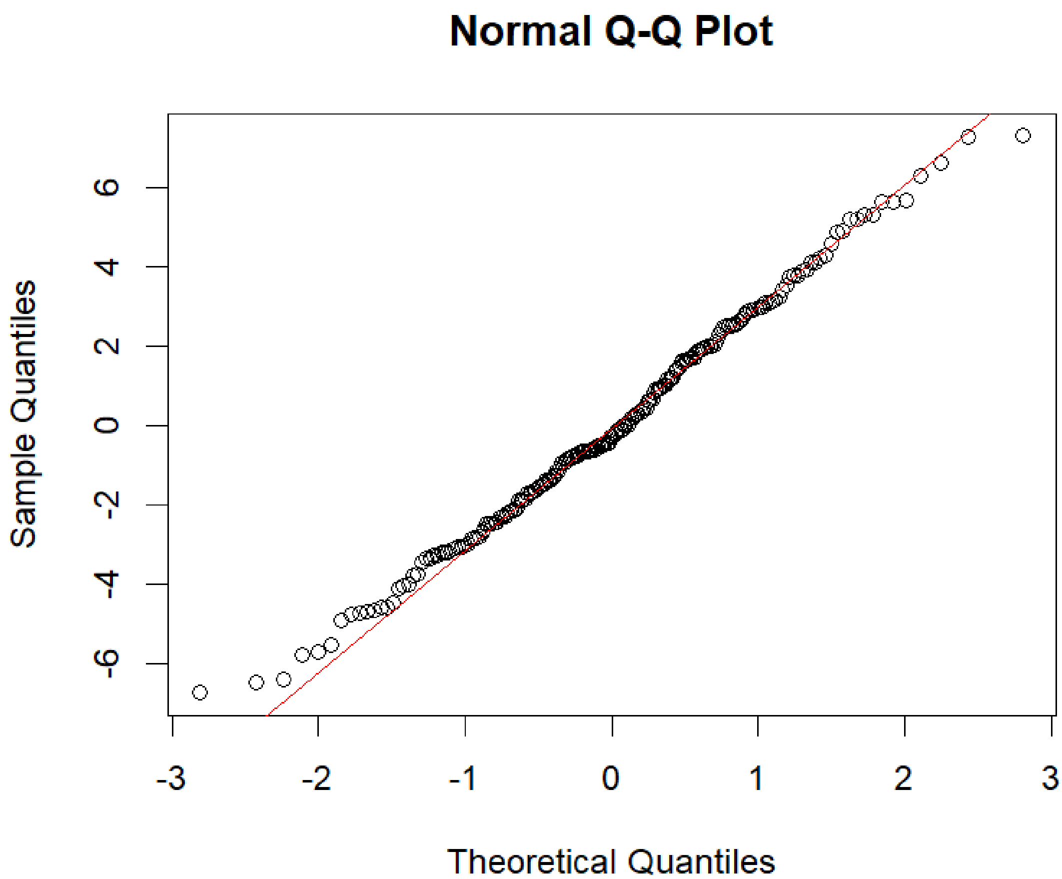
```
             Df Sum Sq Mean Sq F value   Pr(>F)    
W             2    333   166.7  19.010 2.96e-08 ***
Z             2   5839  2919.7 332.926  < 2e-16 ***
W:Z           4     32     8.0   0.912    0.458    
Residuals   191   1675     8.8                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Comments on the significance of the terms:

- The categorical variable W has a statistically significant effect on Y (p<0.001) -> W influences Y
- The categorical variable Z has a statistically significant effect on Y (p<0.001) -> Z influences Y
- The interaction effect between W and Z is not statistically significant (p=0.458>0.05) -> No interaction effect

1. **Normality Assumption**

- Q-Q plot



- Shapiro-Wilk normality test

```
        Shapiro-Wilk normality test

data:  residuals(anova_model)
W = 0.99316, p-value = 0.481
```

From above we can observe that the normality of the residuals holds since the significance level α=0.05 is less than the p-value

2. <u>Homogeneity of Variances Assumption</u>

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   8  1.3199 0.2356
      191
```

Because, the p-value is greater than 0.05, the assumption of homogeneity of variances is not violated