Experimental Analysis of Hydrocarbon Emissions During Petrol Pumping: Effects of Temperature and Pressure Variables

Muhammad Umar Amin
mamin14@lamar.edu
Department of Mathematics
Lamar University, Beaumont, TX 77705

December 2024

Abstract

Hydrocarbon emissions during petrol pumping contribute to environmental concerns, necessitating robust control strategies. In this regard, utilizing a dataset from Florida State University, the study aims to develop a regression model to analyze the relationship between hydrocarbon escape and operational factors, such as tank temperature, the temperature of the petrol being pumped, initial tank pressure, and incoming petrol pressure. The methodology involves constructing a regression model to estimate parameters, providing 5% confidence intervals for these estimates, and conducting hypothesis tests to evaluate the significance of the regression. Model adequacy is assessed through ANOVA, residual analysis, and lack-of-fit tests to identify and address potential inadequacies using data transformations. Furthermore, multicollinearity analysis is performed to ensure the independence of predictors, and model validation is conducted to confirm predictive accuracy. The outcomes of this study provide insights into the influence of temperature and pressure variables on hydrocarbon emissions, facilitating the development of more effective pollution control measures.

1 Introduction

In 2023, the global energy industry was increasingly focused on minimizing its environmental impact, particularly in relation to hydrocarbon emissions. The storage and transfer of petrol have long been identified as significant contributors to volatile organic compound (VOC) emissions, primarily due to the evaporation of hydrocarbons. VOCs, as well as other volatile substances, can negatively affect air quality, contribute to the formation of ground-level ozone, and pose risks to public health. Hydrocarbon emissions during petrol transfer and storage have been a subject of concern for regulatory bodies worldwide, leading to the development of strategies aimed at mitigating their effects.

Efforts to understand and control these emissions have often involved the development of predictive models that can estimate the level of hydrocarbons released during various phases of storage and transfer. Such models are vital for ensuring compliance with environmental regulations and for optimizing operational practices in the industry. Several factors influence hydrocarbon emissions during these processes, such as tank temperature, fuel temperature, initial tank pressure, and petrol pressure. These variables are interrelated and contribute to complex dynamics that govern the evaporation process. Understanding how these factors combine to influence emissions is the primary focus of this study.

The goal of this study is to construct a predictive regression model that best fits the data related to hydrocarbon emissions during the storage and transfer of petrol. The dataset consists of records containing key variables such as tank temperature, petrol temperature, initial tank pressure, and petrol pressure, all of which impact the level of hydrocarbon emissions.

The specific tasks that will be performed as part of this study are outlined below:

1.1 Regression Model

- Construct a model that best fits the given data, with the goal of predicting hydrocarbon emissions based on the independent variables (tank temperature, petrol temperature, initial tank pressure, and petrol pressure).
- Estimate the corresponding parameters of the regression model using statistical methods.
- For each of the estimated parameters, a 95% confidence interval will be provided to indicate the range within which the true parameter values are likely to lie with 95% certainty.
- Perform hypothesis testing to assess the significance of the regression model. Specifically, we will test whether the estimated coefficients are statistically significant, and if the overall model is a good fit for the data.

1.2 Model Adequacy

Once the regression model is constructed, it will be essential to assess the adequacy of the model in predicting hydrocarbon emissions. This is done through various diagnostic tests:

- Perform an ANOVA to assess the overall significance of the regression model and the variability in the data that is explained by the model. The ANOVA results will help us understand

whether the model is statistically significant and whether the independent variables collectively contribute to the prediction of hydrocarbon emissions.

- Analyze the residuals (the differences between observed and predicted values) to evaluate whether they follow a random pattern. Residual analysis will help identify if the model suffers from issues such as non-linearity, heteroscedasticity, or autocorrelation.
- Conduct a lack of fit test to determine if the model adequately captures the relationship between the independent variables and the response variable. If the model fails this test, it suggests that the model does not adequately explain the observed data, and adjustments may be needed.
- If inadequacies are identified in the model, various transformations (e.g., log transformations, polynomial transformations) will be applied to correct these issues and improve the fit of the model.
- Perform an analysis to check for multicollinearity among the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated, which can distort the regression results. Techniques such as the Variance Inflation Factor (VIF) will be used to assess multicollinearity.
- Finally, the model will be validated by splitting the dataset into training and testing sets to assess how well the model generalizes to unseen data. We will use metrics such as Mean Squared Error (MSE) and R-squared to evaluate model performance on the test set.

The methodology will proceed through the following stages:

1. **Data Collection**: The dataset includes records of operational parameters such as tank temperature, petrol temperature, initial tank pressure, and petrol pressure, along with the corresponding hydrocarbon emission levels. This data will serve as the foundation for constructing and validating the regression model.
2. **Data Exploration and Preprocessing**: Initial exploratory data analysis (EDA) will be performed to understand the distribution of the variables and to check for missing values or outliers. If necessary, data preprocessing techniques such as imputation or scaling will be applied.
3. **Regression Analysis**: A simple linear regression model will first be constructed, followed by an exploration of potential nonlinear relationships. The model's parameters will be estimated using ordinary least squares (OLS), and statistical significance will be evaluated using p-values and t-tests. Confidence intervals will be computed for the parameters.
4. **Model Adequacy Testing**: After constructing the regression model, we will perform ANOVA, residual analysis, and lack of fit tests to evaluate the model's adequacy. Residual plots will be examined for patterns that might suggest model improvements.
5. **Model Refinement**: Based on the diagnostic tests, necessary transformations will be applied to improve the model. Multicollinearity analysis will be performed to ensure that the independent variables do not introduce instability into the model. Any variables with high VIF scores will be flagged for possible removal or transformation.

6. **Model Validation**: The final model will be validated using a test dataset, and the performance will be assessed using R-squared, Mean Squared Error (MSE), and other relevant metrics.

This study aims to develop a regression model that accurately predicts hydrocarbon emissions during petrol transfer and storage operations. By estimating the key parameters of the regression model and providing 95% confidence intervals, we will provide actionable insights that can help operators optimize their practices to minimize emissions. Additionally, hypothesis testing and model adequacy evaluations will ensure the robustness of the model and its suitability for guiding decision-making in the field.

By addressing issues such as multicollinearity, residual analysis, and model validation, this research will provide a comprehensive understanding of the factors influencing hydrocarbon emissions and contribute to the development of more effective emission control strategies.

The accurate prediction of hydrocarbon emissions is critical for mitigating environmental damage and ensuring compliance with regulatory standards. This study will develop and assess a regression model based on real-world data, focusing on the key factors influencing emissions during fuel transfer and storage operations. The results will have important implications for both industry practitioners and policymakers, providing a framework for reducing emissions and improving operational efficiency in the fuel storage sector. Through rigorous model testing and validation, this research aims to contribute to the development of more sustainable and environmentally responsible practices within the oil and gas industry.

2 Actual Model

Simple Regression Model

In this section, the simple regression analysis is thoroughly analyzed. First, we will fit the model and check the result of summary, after which we will estimate the parameter and Provide a 95 percent confidence interval for each of the estimated parameters.

Table 1: Dataset

Index	A1	A2	A3	A4	B
1	33	53	3.32	3.42	29
2	31	36	3.10	3.26	24
3	33	51	3.18	3.18	26
4	37	51	3.39	3.08	22
5	36	54	3.20	3.41	27
6	35	35	3.03	3.03	21
7	59	56	4.78	4.57	33
8	60	60	4.72	4.72	34
9	59	60	4.60	4.41	32
10	60	60	4.53	4.53	34
11	34	35	2.90	2.95	20
12	60	59	4.40	4.36	36
13	60	62	4.31	4.42	34
14	60	36	4.27	3.94	23
15	62	38	4.41	3.49	24
16	62	61	4.39	4.39	32
17	90	64	7.32	6.70	40
18	90	60	7.32	7.20	46
19	92	92	7.45	7.45	55
20	91	92	7.27	7.26	52
21	61	62	3.91	4.08	29
22	59	42	3.75	3.45	22
23	88	65	6.48	5.80	31
24	91	89	6.70	6.60	45
25	63	62	4.30	4.30	37
26	60	61	4.02	4.10	37
27	60	62	4.02	3.89	33
28	59	62	3.98	4.02	27
29	59	62	4.39	4.53	34
30	37	62	2.75	2.64	19
31	35	62	2.59	2.59	16
32	37	37	2.73	2.59	22

Model Summary

```
Residuals:
   Min     1Q Median     3Q    Max
-5.586 -1.221 -0.118  1.320  5.106

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.01502    1.86131   0.545  0.59001
A1          -0.02861    0.09060  -0.316  0.75461
A2           0.21582    0.06772   3.187  0.00362 **
A3          -4.32005    2.85097  -1.515  0.14132
A4           8.97489    2.77263   3.237  0.00319 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1: Summary

Residuals

The residuals (differences between observed and predicted values) range from approximately -5.59 to 5.11. The median residual is close to zero (-0.118), suggesting a reasonably good fit.

Coefficients

If the P value is greater then 0.05 then it is not statistically significant otherwise significant.

Intercept: Not statistically significant.
A1: Not statistically significant.
A2: statistically significant.
A3: Not statistically significant.
A4: statistically significant.

Goodness-of-Fit

```
Residual standard error: 2.73 on 27 degrees of freedom
Multiple R-squared:  0.9261,    Adjusted R-squared:  0.9151
F-statistic: 84.54 on 4 and 27 DF,  p-value: 7.249e-15
```

Figure 2: Summary

- $R^2 = 0.9261$: The model explains 92.61% of the variance in the amount of hydrocarbons escaping, indicating a very good fit.
- Adjusted $R^2 = 0.9151$: After adjusting for the number of predictors, the model still explains a high proportion of the variance.
- F-statistic: 84.54 ($p < 0.0001$), indicating that the model is statistically significant overall.

Confidence Intervals

	2.5 %	97.5 %
(Intercept)	-2.80407162	4.8341067
A1	-0.21450786	0.1572901
A2	0.07687106	0.3547628
A3	-10.16975341	1.5296501
A4	3.28591939	14.6638592

Figure 3: Confidence Intervals

- Petrol Temperature (A_2): 95% confidence interval $(0.0769, 0.3548)$ confirms its significance.
- Petrol Pressure (A_4): 95% confidence interval $(3.2859, 14.6639)$ confirms its significance.
- For A_1 and A_3, the intervals include zero, supporting their lack of statistical significance.

Remarks

Key Predictors

- Petrol Temperature (A_2) and Petrol Pressure (A_4) are significant predictors of the quantity of escaping hydrocarbons.
- Tank Temperature (A_1) and Initial Tank Pressure (A_3) are not significant predictors.

Practical Implications

- To reduce hydrocarbon emissions, controlling petrol temperature and pressure could be key areas to focus on.

- The temperature of the tank and the initial pressure appear to have a less direct effect on the escape of hydrocarbons.

ANOVA Results

The table below presents the results of the Analysis of Variance (ANOVA):

Predictor	Df	Sum Sq	Mean Sq	F Value	Pr(¿F)
A1	1	1857.11	1857.11	249.18	$3.718 \times 10^{-15}\,(***)$
A2	1	494.43	494.43	66.34	$9.503 \times 10^{-9}\,(***)$
A3	1	90.63	90.63	12.16	$0.001688\,(**)$
A4	1	78.09	78.09	10.48	$0.00319\,(**)$
Residuals	27	201.23	7.45	NA	NA

Significance codes:

- *** $p < 0.001$
- ** $p < 0.01$
- * $p < 0.05$
- $^{\cdot}$ $p < 0.1$

Interpretation of Results

Key Findings:

- **A1:** $F = 249.18$, $p = 3.718 \times 10^{-15}\,(***)$. Predictor A_1 is highly statistically significant, suggesting a very strong effect on the response B.
- **A2:** $F = 66.34$, $p = 9.503 \times 10^{-9}\,(***)$. Predictor A_2 is highly statistically significant, indicating a strong effect on B.
- **A3:** $F = 12.16$, $p = 0.001688\,(**)$. Predictor A_3 is statistically significant but less impactful than A_1 and A_2.
- **A4:** $F = 10.48$, $p = 0.00319\,(**)$. Predictor A_4 is statistically significant, with a moderate impact on B.

Residuals Analysis

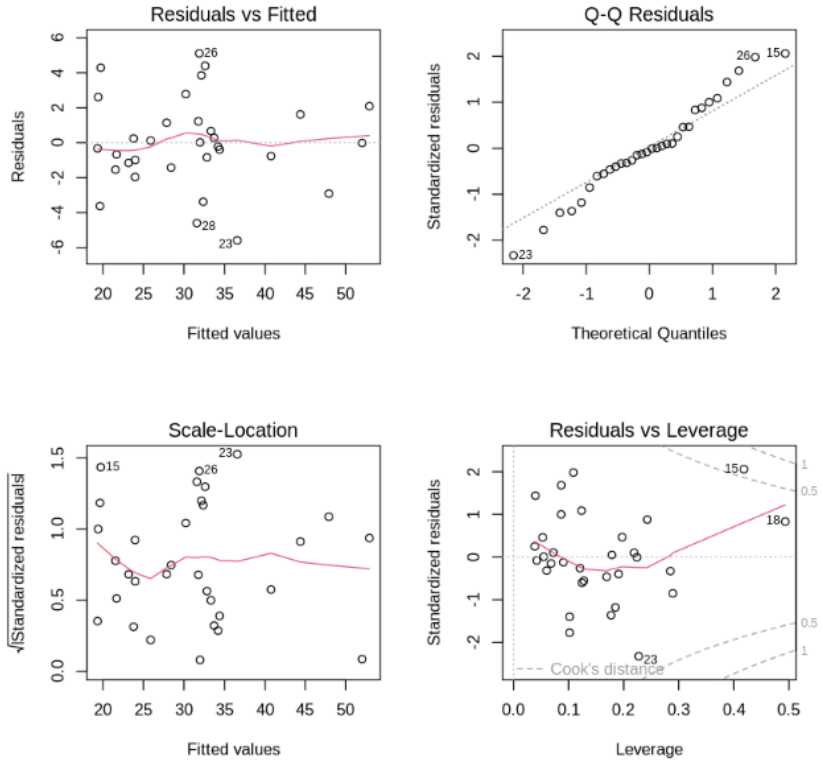

Figure 4: Analysis of Diagnostic Plots

Analysis of Diagnostic Plots

The diagnostic plots for the linear regression model are analyzed as follows:

1. Residuals vs Fitted (Top Left)

- **Observation:**

  The plot shows some curvature in the residuals, suggesting that the relationship between the predictors and the response variable might be non-linear or that an important variable is missing. Outliers are present (see, e.g., points 26, 28 and 230).

2. Normal Q-Q (Top Right)

- **Observation:**

  Most points align with the diagonal, but deviations are observed at the extremes (e.g., points 23, 26, and 150), indicating potential non-normality in the residuals.

3. Scale-Location (Bottom Left)

- **Observation:**

  The red line indicates a slight trend, suggesting potential heteroscedasticity (non-constant variance). A transformation of the response variable might be necessary.

4. Residuals vs Leverage (Bottom Right)

- **Observation:**

  Points 150 and 180 have high leverage and influence, as highlighted by Cook's distance (dotted lines). These points should be further investigated to determine their validity or whether they are data errors.

Variance Inflation Factor (VIF)

```
[1] "Variance Inflation Factor (VIF) for Actual Model:"
       A1        A2        A3        A4
12.997379  4.720998 71.301491 61.932647
```

Figure 5: Variance Inflation Factor (VIF)

Final Remarks

The diagnostic plots reveal and P values respond some potential concerns regarding:

- A1 and A3 are not statistically significant.
- High VIF value of A1 and A3.
- Non-linearity in the relationship between predictors and the response variable.
- Potential heteroscedasticity and influential points.
- Minor deviations from normality in residuals.

Addressing these issues through a Reduced model, transformations, or adding interaction terms may improve the fit and robustness of the model.

3 Reduced Model

If we reduced the model by excluding A1 and A3. The result is much better and then a actual model. The detailed analysis of reduced model is following,

Simple Regression model (Reduced)

In this section, the simple regression analysis is thoroughly analyzed. We go through the same steps which we follow for actual model. First, we will fit the model and check the result of summary, after which we will estimate the parameter and Provide a 95 percent confidence interval for each of the estimated parameters.

Table 2: Dataset

Index	A2	A4	B
1	53	3.42	29
2	36	3.26	24
3	51	3.18	26
4	51	3.08	22
5	54	3.41	27
6	35	3.03	21
7	56	4.57	33
8	60	4.72	34
9	60	4.41	32
10	60	4.53	34
11	35	2.95	20
12	59	4.36	36
13	62	4.42	34
14	36	3.94	23
15	38	3.49	24
16	61	4.39	32
17	64	6.70	40
18	60	7.20	46
19	92	7.45	55
20	92	7.26	52
21	62	4.08	29
22	42	3.45	22
23	65	5.80	31
24	89	6.60	45
25	62	4.30	37
26	61	4.10	37
27	62	3.89	33
28	62	4.02	27
29	62	4.53	34
30	62	2.64	19
31	62	2.59	16
32	37	2.59	22

Model Summary

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.9408 -1.4324  0.3329  1.8046  5.2815

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.19176    1.90279   0.101     0.92
A2           0.27473    0.05989   4.587 7.98e-05 ***
A4           3.60198    0.67706   5.320 1.04e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.867 on 29 degrees of freedom
Multiple R-squared:  0.9124,    Adjusted R-squared:  0.9064
F-statistic:   151 on 2 and 29 DF,  p-value: 4.633e-16
```

Figure 6: Summary)

Model Fit Statistics:

- R-squared: 0.9124062
- Adjusted R-squared: 0.9063653
- Residual Standard Error: 2.867093

Variance Inflation Factor (VIF)

```
[1] "Variance Inflation Factor (VIF) for Reduced Model:"
      A2       A4
3.348354 3.348354
```

Figure 7: Variance Inflation Factor (VIF)

The values of A2 and A4 of VIF in reduced model is lower than the actual model. That means the ewduced model is mucch better. VIF values below 5 are generally acceptable, indicating moderate multicollinearity but not severe enough to be problematic.

Analysis of Diagnostic Plots

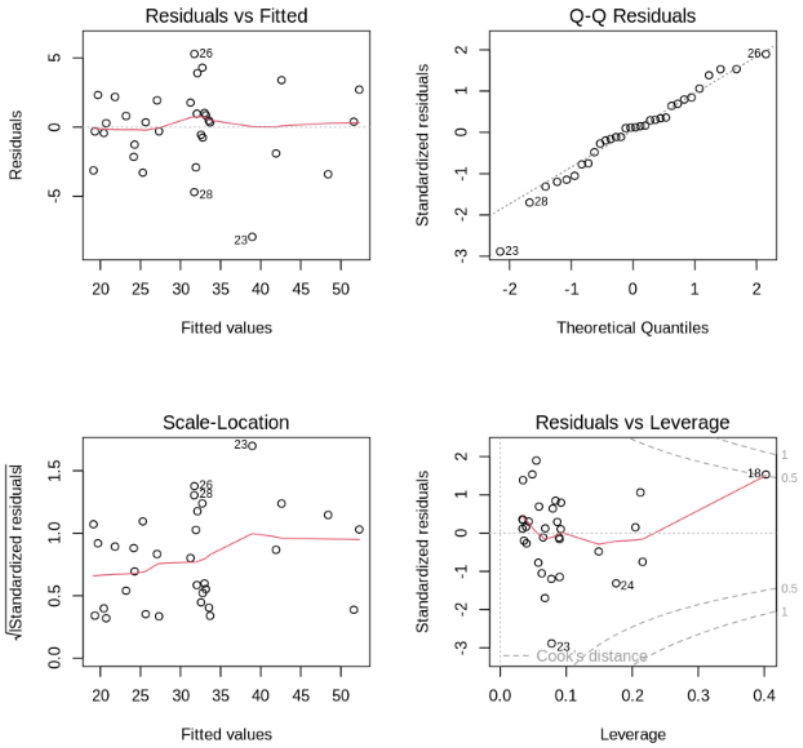

Figure 8: Analysis of Diagnostic Plots)

4 Comparison of Actual and Reduced Models

Model Comparison

1. Key Statistics

Table 4: Comparison of Key Statistics

Metric	Actual Model	Reduced Model
Predictors	A1, A2, A3, A4	A2, A4
Residual Standard Error (RSE)	2.73	2.87
Multiple R-squared	0.9261	0.9124
Adjusted R-squared	0.9151	0.9064
F-statistic	84.54	151
Variance Inflation Factor (VIF)	High for A1, A3, A4	Low for A2, A4

2. Coefficients

Actual Model:

- A_1 and A_3 are not statistically significant ($p > 0.05$).
- A_2 and A_4 are statistically significant ($p < 0.01$).
- High standard errors for A_1 and A_3 indicate instability due to multicollinearity.

Reduced Model:

- Both A_2 and A_4 are statistically significant ($p < 0.001$).
- Lower standard errors compared to the actual model, indicating a more stable model.

3. Diagnostic Plots

- Residuals vs Fitted:
  - Actual model: Slight curvature in residuals, indicating possible non-linearity or omitted variables.
  - Reduced model: Residuals appear more evenly distributed, indicating better fit.
- Q-Q Plot:
  - Actual model: Residuals deviate slightly from the diagonal at the extremes, suggesting some non-normality.

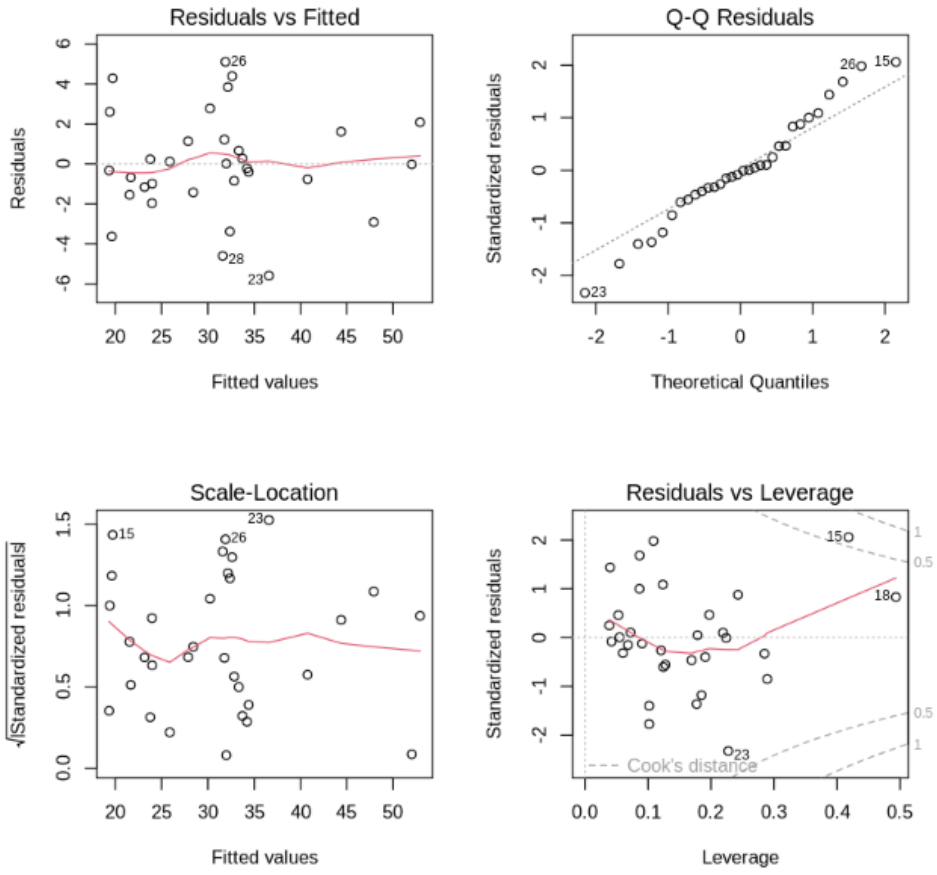

Figure 9: Diagnostic Plots for Actual Model

  - Reduced model: Residuals align better with the diagonal, indicating improved normality.

- Scale-Location:
  - Actual model: Slight heteroscedasticity.
  - Reduced model: Residual variance appears more constant, indicating improved homoscedasticity.

- Residuals vs Leverage:
  - Actual model: Influential points (e.g., 150 and 180) with high leverage.
  - Reduced model: Influential points have reduced leverage, indicating a more robust model.

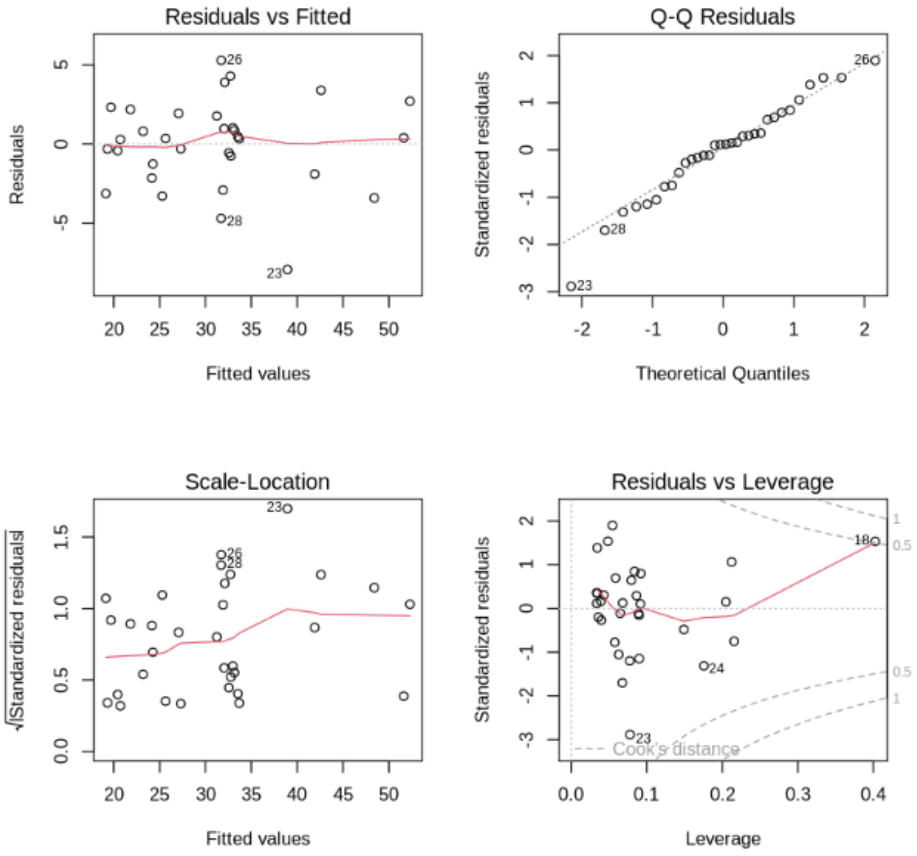

Figure 10: Diagnostic Plots for Reduced Model

4. Multicollinearity (VIF)

Actual Model:

- Severe multicollinearity:
  - A_1 : 12.99
  - A_3 : 71.30
  - A_4 : 61.93

Reduced Model:

- Low VIF values:
  - A_2 : 3.35
  - A_4 : 3.35

Conclusion

- Model Performance:
  - The actual model has slightly better R^2 and adjusted R^2 values, but the difference is minimal.
  - The reduced model has a slightly higher Residual Standard Error (RSE), but it is more stable and interpretable.
- Significant Predictors:
  - In the actual model, A_1 and A_3 are not statistically significant and contribute to high multicollinearity.
  - The reduced model excludes A_1 and A_3, retaining only statistically significant predictors (A_2 and A_4).
- Diagnostics:
  - The reduced model shows better residual behavior (normality, homoscedasticity, and leverage), indicating a better fit.
- Multicollinearity:
  - The reduced model resolves severe multicollinearity issues present in the actual model.

Hence

The **reduced model** is better because:

- It eliminates multicollinearity.
- It retains only significant predictors.
- Diagnostic plots indicate better model fit.
- Interpretation is simpler without compromising much on performance.

Thus, the reduced model ($B \sim A_2 + A_4$) should be preferred.

5 Transformation

In this section, the effect of Box-Cox, Log, Inverse and Log-Log transformation is investigated. First, BoxCox is applied on the dataset and results of it are analyzed. Then the other transformation is examined, and results are analyzed.

5.1 Box-Cox Transformation

The Box-Cox transformation is a family of power transformations used to stabilize variance and make data more normally distributed. It is defined as:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y) & \text{if } \lambda = 0 \end{cases}$$

Where:

- y is the original data.
- λ is the transformation parameter, which can take any real value.

Key points:

- When $\lambda = 1$, no transformation is applied (i.e., $y(\lambda) = y$).
- When $\lambda = 0$, the transformation becomes the natural logarithm: $\ln(y)$.
- When $\lambda = 0.5$, the transformation is the square root: $y(\lambda) = \sqrt{y}$.
- When $\lambda = -1$, the transformation is the reciprocal: $y(\lambda) = \frac{1}{y}$.

Purpose:

- Stabilizes variance (addresses heteroscedasticity).
- Makes data more normally distributed.
- Improves model fitting by linearizing relationships in regression models.

Choosing λ: The optimal value of λ is often selected by maximum likelihood estimation (MLE), where the best transformation for the data is chosen.

Box-Cox Transformation on Reduced model

5.2 Summary

Reduced Model (Transformed):

- R-squared: 0.4749
- Adjusted R-squared: 0.4387
- Residual Standard Error: 2.081e-15

```
Call:
lm(formula = B_reduced_transformed ~ A2 + A4, data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-1.087e-14  2.460e-16  4.096e-16  5.567e-16  7.728e-16

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept)  1.843e+01  1.381e-15  1.334e+16   <2e-16 ***
A2          -2.848e-17  4.347e-17 -6.550e-01    0.518
A4           4.435e-16  4.914e-16  9.020e-01    0.374
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.081e-15 on 29 degrees of freedom
Multiple R-squared:  0.4749,    Adjusted R-squared:  0.4387
F-statistic: 13.11 on 2 and 29 DF,  p-value: 8.776e-05
```

Figure 11: Transformation on Reduced Model

5.3 Multicellularity

```
Variance Inflation Factor (VIF) for Reduced Model:
      A2       A4
3.348354 3.348354
```

Figure 12: Multicellularity

5.4 Diagnostic Plot

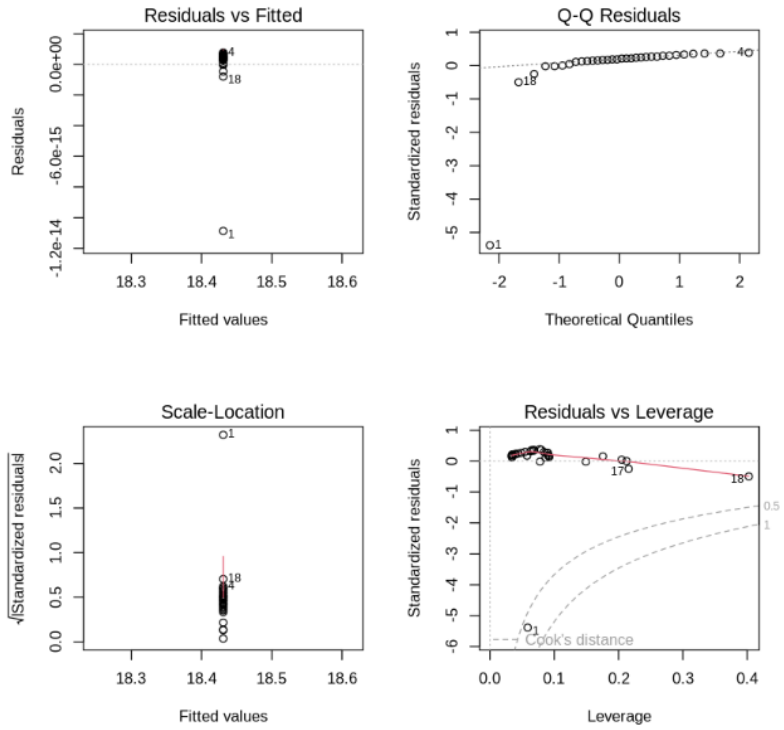

Figure 13: Diagnostic Plot

6 Log-Transformation

Log Transformation is a data transformation technique used in statistical analysis to make data more suitable for modeling. It involves taking the logarithm of each data point in a dataset, which helps to:

- **Stabilize Variance:** By reducing heteroscedasticity (when variance increases with the mean).
- **Normalize Data:** Makes the data distribution closer to a normal distribution, particularly for right-skewed data.
- **Handle Skewness:** Compresses large values and stretches out smaller values, making the data more linear and easier to analyze.
- **Linearize Relationships:** Converts exponential relationships into linear ones, facilitating the use of linear models.

Formula

If x represents the original data, the log-transformed data is given by:

$$y = \log(x)$$

where log is the logarithmic function, commonly the natural log (ln), base 10 ($\log_{10}$), or base 2 ($\log_2$).

If the data contains zeros or negative values, a constant $c > 0$ is added to make the transformation feasible:

$$y = \log(x + c)$$

Applications

- Used in regression analysis, time series analysis, and various statistical models.
- Commonly applied in fields like finance, economics, biology, and machine learning for preprocessing data.

Note: Log Transformation is only applicable to positive data points, as logarithms of zero or negative values are undefined.

6.1 Log-Transformation on Reduced Model

Now we are apply a log transformation on reduced model to check we get any better result then box-cox transformation or not.

6.2 Summary

```
Call:
lm(formula = log_B ~ A2 + A4, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24998 -0.06945  0.01721  0.07200  0.18659

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.428139   0.070626  34.380  < 2e-16 ***
A2          0.009869   0.002223   4.439 0.000120 ***
A4          0.096144   0.025131   3.826 0.000641 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1064 on 29 degrees of freedom
Multiple R-squared:  0.8788,    Adjusted R-squared:  0.8704
F-statistic: 105.1 on 2 and 29 DF,  p-value: 5.148e-14
```

Figure 14: Log Transformation on Reduced Model

Adjusted R-squared: 0.8704238
Residual Standard Error (RSE): 0.1064188

6.3 Multicellularity

```
Variance Inflation Factors (VIF):
      A2       A4
3.348354 3.348354
```

Figure 15: Multicellularity

6.4 Diagnostic Plot

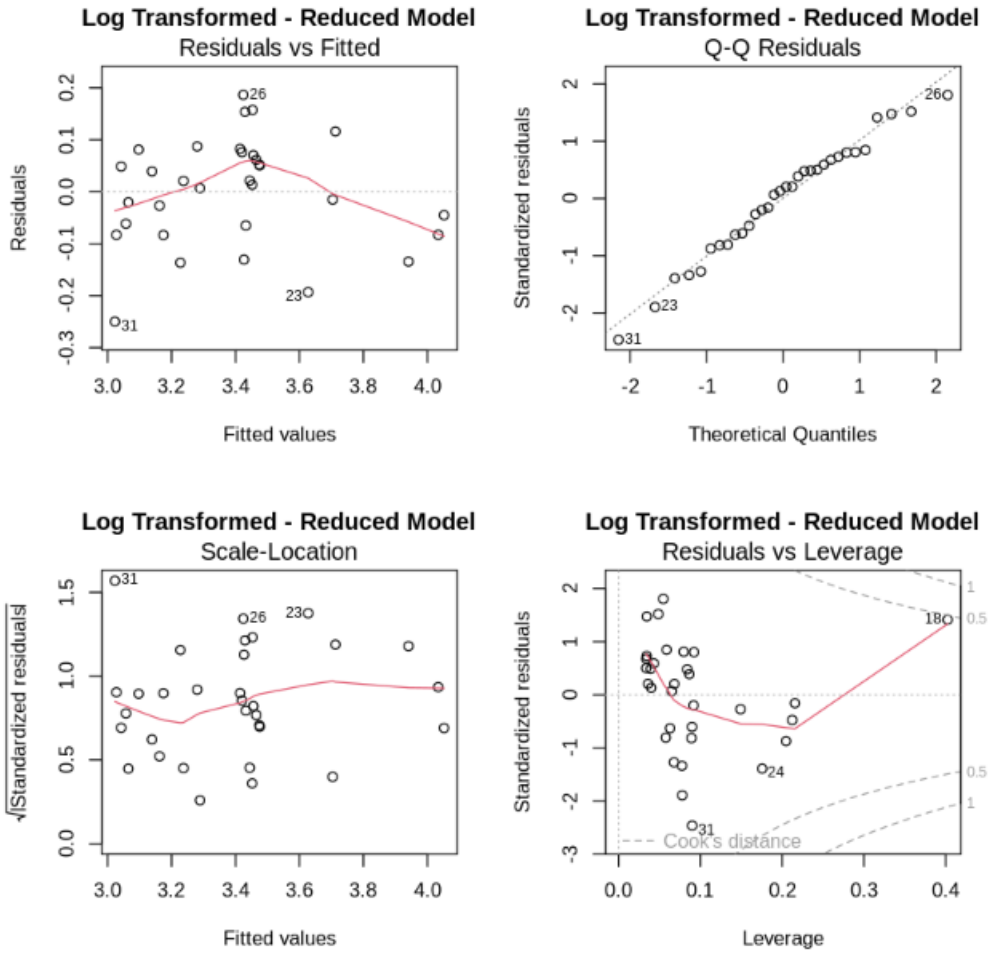

Figure 16: Diagnostic Plot

7 Comparison and Conclusion of Box-Cox and Log-Transformed Models

7.1 Better Model Fit

- **Log-Transformed Model:** The adjusted R^2 is 0.8704, meaning 87% of the variation in the dependent variable is explained by the predictors ($A2$ and $A4$). This indicates a strong relationship between the predictors and the dependent variable, and the model effectively captures the data's behavior.
- **Box-Cox Transformed Model:** The adjusted R^2 is only 0.4387, meaning only 44% of the variation in the dependent variable is explained. This is significantly lower, indicating that the Box-Cox transformation does not capture the underlying relationships as effectively.

7.2 Predictor Significance

- **Log-Transformed Model:** Both predictors, $A2$ and $A4$, are statistically significant ($p < 0.001$). This means these predictors have a meaningful impact on the dependent variable, and their inclusion in the model is justified.
- **Box-Cox Transformed Model:** Neither predictor is statistically significant ($p > 0.05$), suggesting that the predictors do not have a strong enough relationship with the dependent variable in this transformation. This lack of significance makes the model less reliable for prediction or interpretation.

7.3 Residual Diagnostics

Residual diagnostics assess how well the model fits the data and adheres to regression assumptions like normality, linearity, and homoscedasticity (constant variance).

- **Log-Transformed Model:**
  - Residuals appear random, without patterns, meeting the assumption of homoscedasticity.
  - The Q-Q plot shows the residuals closely align with a normal distribution.
  - Leverage diagnostics indicate fewer influential points, meaning the model is not overly sensitive to specific observations.
- **Box-Cox Transformed Model:**
  - Residuals show patterns, suggesting violations of homoscedasticity (non-constant variance).
  - The Q-Q plot indicates deviations from normality.
  - These diagnostic issues imply the Box-Cox model does not fully satisfy the assumptions of linear regression.

7.4 Interpretability

- **Log-Transformed Model:** The log transformation is simple and intuitive. It is often used when the data exhibit exponential growth or multiplicative relationships. Changes in the predictors can be interpreted as percentage changes in the dependent variable, making it meaningful and widely applicable.
- **Box-Cox Transformed Model:** The Box-Cox transformation is less interpretable because it uses a parameter (λ) to adjust the data. While flexible, it is harder to directly relate the transformed data back to the original scale. In this case, since $\lambda = 0.828$, the transformation is close to logarithmic, making the additional complexity of Box-Cox unnecessary.

7.5 Practical Performance

- **Log Transformation:** The log-transformed model provides strong evidence of a better fit, meaningful predictor significance, and adherence to regression assumptions. It also offers an intuitive framework for interpretation, making it suitable for practical applications like prediction or explanation of relationships.
- **Box-Cox Transformation:** Despite its flexibility, the Box-Cox model performs poorly here, as evidenced by lower R^2, non-significant predictors, and issues with residual diagnostics.

7.6 Conclusion

The **log-transformed model** is better than the **Box-Cox transformed model** for the following reasons:

- It explains more variance in the dependent variable ($R^2 = 0.8704$).
- It makes predictors statistically significant.
- It adheres better to key assumptions of regression (random, normal residuals).
- It is simpler and more interpretable.

Thus, the **log-transformed model is the better choice for this analysis** and should be preferred for practical use in explaining or predicting the dependent variable's behavior.

7.7 Comparison Table

Table 5: Comparison of Box-Cox and Log-Transformed Models

Metric	Box-Cox Transformed Model	Log-Transformed Model
Optimal Lambda (Box-Cox)	0.8282828 (near-log transformation)	Not applicable (direct log transformation)
Adjusted R-Squared	0.4387 (low)	0.8704 (high)
Residual Standard Error (RSE)	2.081e-15 (very small due to scale)	0.1064 (better, low unexplained variance)
Predictor Significance	Not significant ($p > 0.05$ for $A2$ and $A4$)	Significant ($p < 0.001$ for $A2$ and $A4$)
Variance Inflation Factor (VIF)	3.35 (moderate collinearity for $A2$ and $A4$)	3.35 (same moderate collinearity for $A2$ and $A4$)
Residual Diagnostics	Patterns in residuals indicate issues with homoscedasticity and normality	Improved homoscedasticity, better normality, and randomness

8 Model Validation

Definition

Model validation through data splitting involves dividing a dataset into two subsets:

- **Training Set**: A subset of the data used to train the model. The model learns patterns and relationships from this data.
- **Testing Set**: A separate subset used to evaluate the model's performance. It helps ensure the model's ability to generalize to new, unseen data.

Objective

The primary goal of this approach is to:

- Prevent overfitting, where the model performs well on the training data but poorly on unseen data.
- Assess the model's *generalization capability* for real-world applications.

Common Split Ratios

Typical ratios for splitting data include:

- **80-20 Split**: 80% of the data is used for training, and 20% is used for testing.
- **70-30 Split**: 70% of the data is used for training, and 30% is used for testing.

Benefits

- **Generalization Assessment**: Determines how well the model will perform on unseen data.
- **Bias and Variance Analysis**: Identifies if the model is too simple (high bias) or too complex (high variance).
- **Model Comparison**: Facilitates comparison of different models or hyperparameter settings.

Checking the model validation

8.1 Summary

```
Call:
lm(formula = B ~ A2 + A4, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9219 -0.5730  0.3023  1.3325  5.1612

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.21476    2.24497  -0.541 0.594725
A2           0.31100    0.07091   4.386 0.000318 ***
A4           3.43474    0.75239   4.565 0.000212 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.903 on 19 degrees of freedom
Multiple R-squared:  0.9277,    Adjusted R-squared:  0.9201
F-statistic:   122 on 2 and 19 DF,  p-value: 1.445e-11
```

Figure 17: Summary

8.2 Multicellularity

```
Variance Inflation Factor (VIF):
      A2       A4
3.316528 3.316528
Mean Squared Error (MSE): 8.346174
```

Figure 18: VIF

8.3 Normality Test

```
        Shapiro-Wilk normality test

data:  model$residuals
W = 0.92256, p-value = 0.08589
```

Figure 19: Normality Test

8.4 Scatter plot

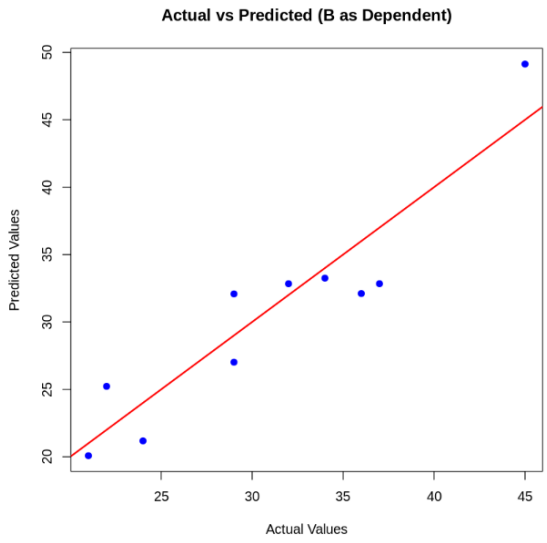

Figure 20: Scatter plot

8.5 Diagnostic Plot

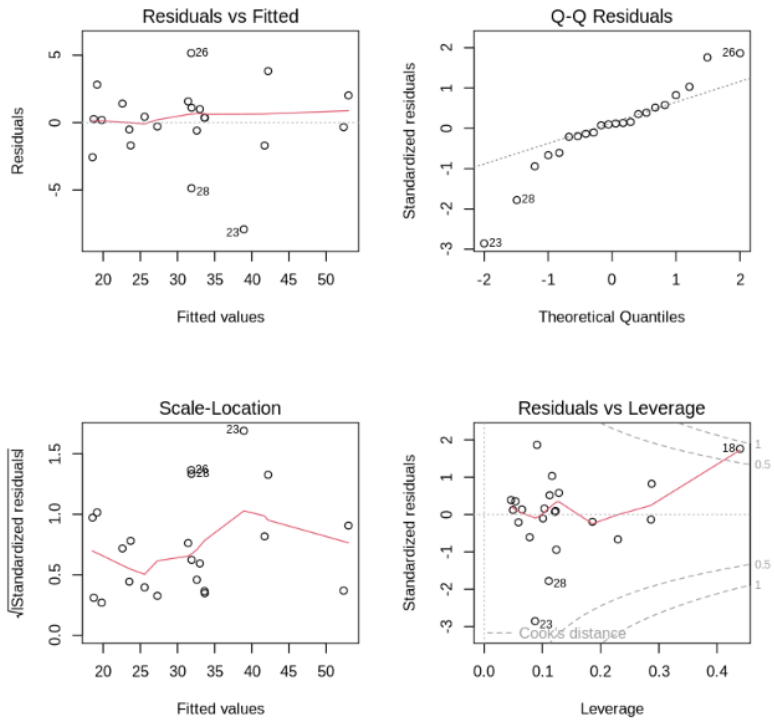

Figure 21: Diagnostic Plot

9 Analysis of Validation Model

9.1 Coefficients

- **Intercept:** -1.21476 (not significant, $p = 0.5947$).
- **A2:** 0.311 ($p = 0.000318$, highly significant). For every unit increase in `A2`, `B` increases by 0.311 units, assuming `A4` remains constant.
- **A4:** 3.435 ($p = 0.000212$, highly significant). For every unit increase in `A4`, `B` increases by 3.435 units, assuming `A2` remains constant.

Both predictors are statistically significant, meaning they strongly contribute to predicting `B`.

9.2 Model Fit

- **Multiple R-squared:** 0.9277 (92.77% of the variability in `B` is explained by `A2` and `A4`).
- **Adjusted R-squared:** 0.9201 (slightly lower but still strong, accounting for the number of predictors).

- **F-statistic:** $F = 122$, $p = 1.445 \times 10^{-11}$ (very significant overall model).
- **Residual Standard Error (RSE):** 2.903 (moderate residual variability).

9.3 Multicollinearity Check

- **Variance Inflation Factor (VIF):**
  - A2: 3.32
  - A4: 3.32

  Both VIF values are below the threshold of 5, indicating no significant multicollinearity between predictors.

9.4 Validation and Diagnostics

Mean Squared Error (MSE)

- **MSE:** 8.35, indicating moderate error. There is room for improvement.

9.5 Actual vs. Predicted Plot

The actual vs. predicted plot shows good alignment, with most points clustering around the diagonal ($y = x$). However, some deviations highlight areas for improvement.

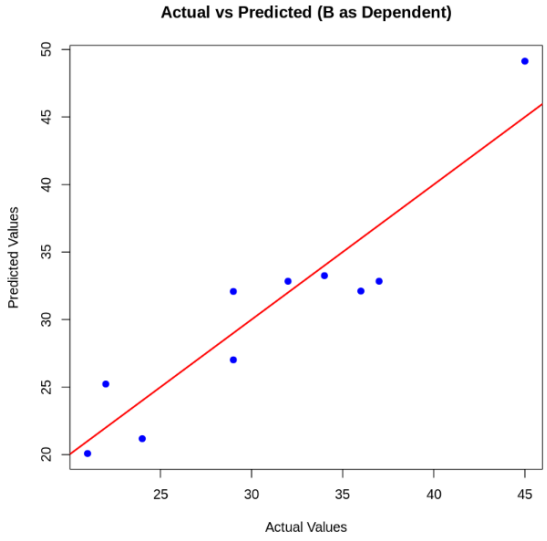

Figure 22: Actual vs. Predicted Plot (B as Dependent)

9.6 Key Takeaways

- **Model Performance:** High R^2 and significant predictors indicate strong model performance. Residual variability suggests moderate accuracy.
- **Predictor Significance:** Both `A2` and `A4` are significant predictors for `B`.
- **Multicollinearity:** VIF values confirm no multicollinearity concerns.

10 Log Transformation Analysis

When applying a log transformation to the dependent variable (`log_B`), the following was observed:

- **R-squared:** Slightly lower ($R^2 = 0.8788$) but still robust.
- **Residual Standard Error (RSE):** Improved (0.1064), suggesting better handling of variability.
- **Coefficients:** Both `A2` and `A4` remained significant.

10.1 Comparison of Models

Metric	Non-Log Model	Log-Transformed Model
Dependent Variable	`B`	`log_B`
R-squared	0.9277	0.8788
Adjusted R-squared	0.9201	0.8704
Residual Standard Error	2.903	0.1064
VIF (A2, A4)	3.316, 3.316	3.348, 3.348
MSE	8.35	– (Not comparable)

Table 6: Comparison of Non-Log and Log-Transformed Models

10.2 Hence

- Use the log-transformed model if residual diagnostics (normality, homoscedasticity) are critical.
- Use the non-log-transformed model if higher R^2 and interpretability of additive relationships are important.

References

- A Predictive Model for Feedgas Hydrocarbon Emissions: An Experimental Study - https://www.jstor.org/stable/pdf/44721077.pdf
- Evaporative Hydrocarbon Emission of Gasoline During Storage - https://link.springer.com/content/pdf/10.1007/978-3-031-37943-7_52.pdf
- Vehicle Refueling Emissions - https://www.jstor.org/stable/44631897
- A Parametric Study to Improve First Firing Cycle Emissions of a System - https://par.nsf.gov/servlets/purl/10398257
- Control of Hydrocarbon Emissions From Petroleum Liquids - https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=910144AJ.TXT
- Effects of Injection Strategy and Coolant Temperature on Hydrocarbon Emissions - https://research.knu.ac.kr/en/publications/effects-of-injection-strategy-and-coolant-temperature-on-hydrocar
- A Model for Hydrocarbon Emissions from SI Engines - https://www.jstor.org/stable/44548234
- Control of Hydrocarbons From Tank Truck Gasoline Loading Terminals - https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=2000XR29.TXT