APPLIED MASTERS EXAM SPRING 2024

Satoshi Ido

April 2024

Problem 1: Effects of Autonomous Driving Factors on Heart Rate Variability: A Crossover Design Study

1 Summary

This study aims to investigate the effects of three factors associated with autonomous driving - transparency, recommended control mode, and system reliability - on heart rate variability (HRV). Each factor is dichotomized into high and low levels, resulting in a total of eight treatment combinations. Utilizing simulated drives, subjects will participate in two unique simulations to measure HRV across various treatment combinations. The experimental design employs a crossover approach with Latin squares, ensuring balanced exposure to all treatments across subjects without necessitating 100 unique drives.

2 Introduction

The goal of this study is to explore how different combinations of transparency, recommended control mode, and system reliability impact HRV. By employing simulated drives and wearable HRV measurement devices, this research seeks to identify the most effective combinations of these factors in influencing HRV.

3 Methods

3.1 Participants

A total of 48 subjects will be recruited to participate in the experiment twice, ensuring a balanced and comprehensive examination of the treatment effects.

3.2 Procedure

Subjects will undergo a brief orientation before beginning the experiment. Each participant will experience two simulated drives separated by a two-week washout period to mitigate carryover effects.

3.3 Experimental Design

The study utilizes a crossover design with two 4x4 Latin Squares to accommodate the study of eight treatment combinations, anticipating variability between subjects and allowing for the assessment of first-order residual or carryover effects. This design ensures that each treatment combination is tested an equal number of times across all subjects but in different orders. For the sake of simplicity and clarity in presentation, alphabets (A through H) are used to denote specific treatment combinations. These alphabetic designations are further elucidated in Tables 1 and 2, showcasing how two distinct Latin Squares are constructed for periods 1 to 4 and 5 to 8, respectively.

3.3.1 Figure 1: Treatment Combinations

A: High Transparency (+), High Recommended Control Mode (+), High System Reliability (+)
B: High Transparency (+), High Recommended Control Mode (+), Low System Reliability (-)
C: High Transparency (+), Low Recommended Control Mode (-), High System Reliability (+)
D: High Transparency (+), Low Recommended Control Mode (-), Low System Reliability (-)
E: Low Transparency (-), High Recommended Control Mode (+), High System Reliability (+)
F: Low Transparency (-), High Recommended Control Mode (+), Low System Reliability (-)
G: Low Transparency (-), Low Recommended Control Mode (-), High System Reliability (+)
H: Low Transparency (-), Low Recommended Control Mode (-), Low System Reliability (-)

3.3.2 Tables 1 and 2: Latin Squares for First and Second Simulations

	Period 5	Period 6	Period 7	Period 8
1	E	F	G	H
2	F	G	H	E
3	G	H	E	F
4	H	E	F	G

Table 1: Latin Square for Treatments A, B, C, D (Periods 1 to 4)

	Period 1	Period 2	Period 3	Period 4
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

Table 2: Latin Square for Treatments E, F, G, H (Periods 5 to 8)

4 Analysis

4.1 Statistical Model

The response variable, HRV, is analyzed using the model:

$$y_{ijk} = \mu + \times P_i + \tau_j + S_k + r_{ij'} + \epsilon_{ijk}$$

where:

$$
\begin{aligned}
y_{ijk} &= \text{HRV for the } k\text{th subject in the } i\text{th period receiving the } j\text{th treatment} \\
\mu &= \text{Overall mean HRV} \\
P_i &= \text{Effect of the } i\text{th period (=block)} \\
\tau_j &= \text{Effect of the } j\text{th treatment} \\
S_k &= \text{Effect of the } k\text{th subject (=block)} \\
r_{ij'} &= \text{First-order residual effect (for } i \neq 1) \\
\epsilon_{ijk} &\sim N(0, \sigma^2) \text{ (Random error term)}
\end{aligned}
$$

4.2 ANOVA Table for Crossover Design

The ANOVA table summarizes the analysis of variance (ANOVA) for the crossover design with Latin Squares. This ANOVA table will be calculated, including the sum of squares, degrees of freedom, mean squares, and F-statistics after running the first-order residual effects model provided above.

Table 3: Analysis of Variance (ANOVA) Table

Source	DF	Sum of Squares	Mean Square	F-value	P-value
Period (P_i)	7	SS_P	$MS_P = SS_P/DF_P$	MS_P/MSE	P-value
Treatment (τ_j)	7	SS_τ	$MS_\tau = SS_\tau/DF_\tau$	MS_τ/MSE	P-value
Subject (S_k)	47	SS_S	$MS_S = SS_S/DF_S$	MS_S/MSE	P-value
First-order Residual ($r_{ij'}$)	288	SS_r	$MS_r = SS_r/DF_r$	MS_r/MSE	P-value
Error (ϵ_{ijk})	35	SSE	$MSE = SSE/DF_E$		
Total	383	SST			

From this ANOVA table, main effects will be tested. This will help to answer whether the combination of transparency, recommended control mode, and system reliability on HRV.

5 Conclusion

This research provides valuable insights into the combinations between transparency, recommended control mode, and system reliability on heart rate variability (HRV) within the context of autonomous driving. Through a methodical experimental design employing a crossover approach with Latin squares, we ensured comprehensive exposure to all treatment combinations, thereby facilitating a robust analysis of their effects on HRV. The analysis, underscored by the ANOVA table, illuminates the significant impact of these factors on physiological responses during simulated driving tasks.

6 Appendix

6.1 Degrees of Freedom Explanation

This section provides an updated explanation for calculating the degrees of freedom for each factor in the model, reflecting the structure of the crossover study with two Latin squares.

1. **Period (P_i):**
   - $DF = p - 1$ where $p = 8$ (number of periods).
   - $DF_{P_i} = 8 - 1 = 7$.
2. **Treatment (τ_j):**
   - $DF = t - 1$ where $t = 8$ (number of treatments).
   - $DF_{\tau_j} = 8 - 1 = 7$.
3. **Subject (S_k) (considered as blocks):**
   - $DF = n - 1$ where $n = 48$ (number of subjects).
   - $DF_{S_k} = 48 - 1 = 47$.
4. **First-order Residual Effect ($r_{ij'}$):**
   - Accounting for the washout period, effectively splitting the experiment into two phases with no carryover between them, and recognizing that the first period of each phase does not contribute to carryover, the calculation is as below:
   - Each subject has 6 potential carryover effects.
   - $DF_{r_{ij'}} = n \times 6$ where $n = 48$ (number of subjects).
   - $DF_{r_{ij'}} = 48 \times 6 = 288$.
5. **Error (ϵ_{ijk}):**
   - Total observations $= n \times p = 48 \times 8 = 384$.
   - Error DF = Total observations - DF_{P_i} - DF_{τ_j} - DF_{S_k} - $DF_{r_{ij'}}$.
   - $DF_{\epsilon_{ijk}} = 384 - 7 - 7 - 47 - 288 = 35$.

Problem 2: Analyzing Gas Consumption Patterns: An Analytical Study on Interval Mileage and Gallons with Key Visualizations and Statistical Modeling Insights

1 Summary

Based on the data visualization and analysis, we observe seasonal patterns of interval mileage and interval gallons over years. The paired t-test indicates a statistically significant difference between the computer-generated interval mileage and actual mileage, suggesting that the auto-generated numbers for interval mileage are often overestimated. Considering the sample size, data characteristics, and simplicity of interpretations, a multiple linear regression model is an appropriate method to analyze the interval gallons, aiming to predict future gas consumption.

2 Introduction

Our goal in this study is to identify any insightful patterns within the data. This report focuses on visualization and statistical modeling. Visualization encompasses both original and derived data to explore observed patterns in overall and at-fill-up data. The statistical model assesses the impact of these patterns on interval gallons, the volume of gas required to refill the tank during a specified period.

3 Methods

For simplicity and clarity, data labeled ‘Overall‘ use ‘overall‘ as a prefix, and ‘At Fill-up data‘ use ‘interval‘.
Visualizations were generated based on ‘Date‘, ‘Year‘, ‘Month‘, and ‘Day of Week‘. Lag data for ‘interval miles‘, ‘interval gallons‘, ‘overall miles‘, and ‘overall gallons‘ were created to capture the impact of past events on current outcomes. This includes rolling averages and exponentially weighted moving averages (EWMAs) to highlight recent trends. Graphs also feature the Exponentially Weighted Standard Deviation (EWSD), emphasizing the relevance of recent observations in estimating volatility.
The paired t-test compares ‘The computer-generated interval MPG‘ with ‘Actual interval MPG‘, testing the hypothesis that computer-generated MPG values are consistently higher.

4 Analysis

4.1 Visualization

Initially, we examine the histograms of “Interval Gallons” and “Interval of Miles.” The histogram of interval gallons exhibits a left-skewed distribution, indicating that the most common range for the amount of gas filled up is between 8 to 10 gallons. Conversely, the histogram of Interval Miles reveals a right-skewed distribution, demonstrating that interval mileage typically ranges between 400 to 700 miles per interval.

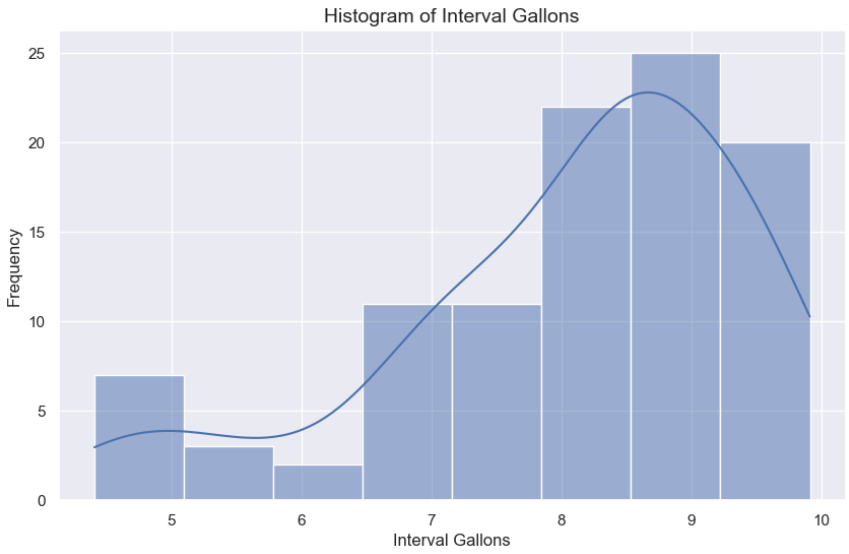

Figure 1: Histogram of Interval Gallons

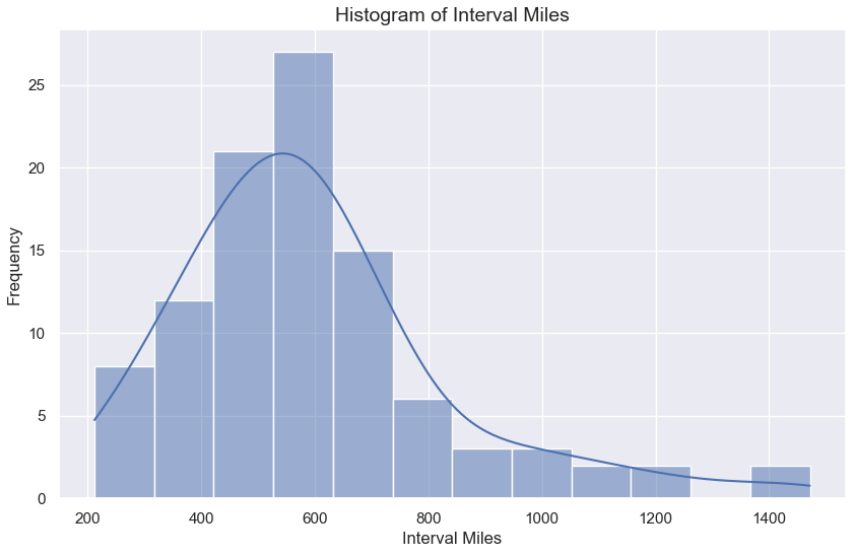

Figure 2: Histogram of Interval Mileages

Next, we analyze the average interval gallons and average interval miles by year, alongside their Exponentially Weighted Moving Averages (EWMAs) by year. In 2020, both mileage and gas consumption decreased, largely due to restrictions on human activities and the closure of campuses amidst the pandemic, reducing the necessity for travel. Conversely, in 2021, as normal activities

gradually resumed, we observed the highest mileages driven and gas consumption within the period studied. The EWSD plot reveals minor variations in interval gallons filled, indicative of predominately long-distance driving in 2021. Post-2021, both mileage and gas consumption have shown a gradual decline. Continuous observation is necessary to understand trends into 2024.

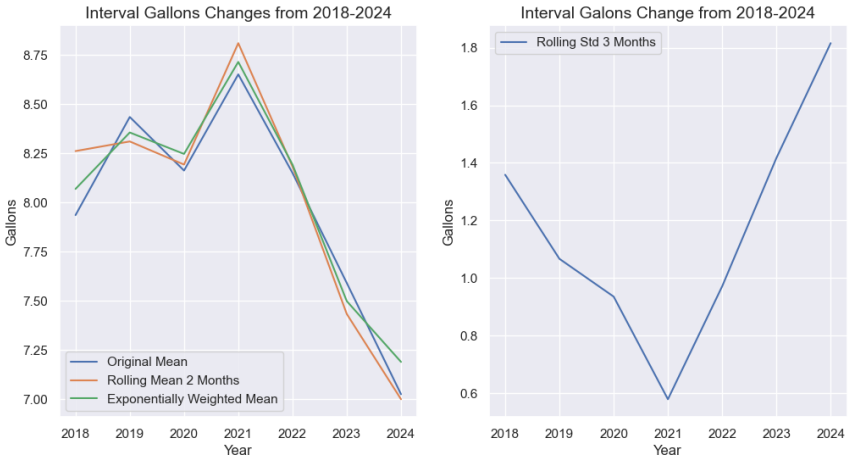

Figure 3: Interval Gallons by Year

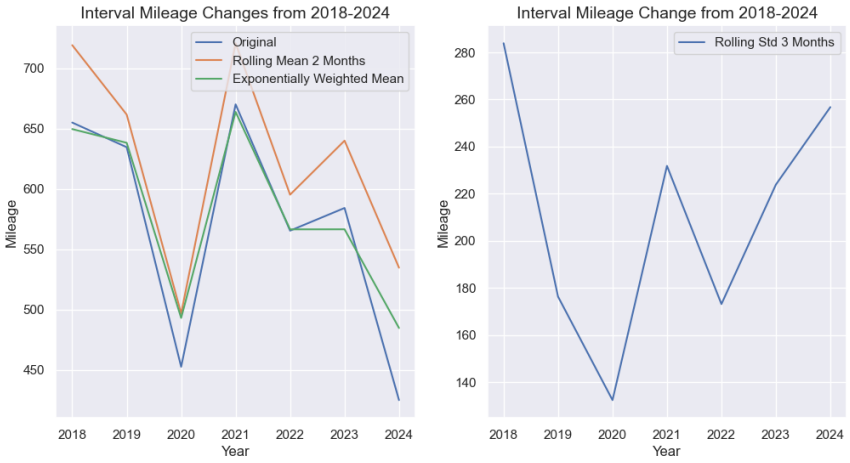

Figure 4: Interval Mileage by Year

Reviewing monthly data for mileage and gas consumption reveals distinct trends. Notably, long-distance driving and significant gas usage occur especially from May, and September through November, correlating with increased activities during warmer months. A distinctive feature between spring and fall activities is the higher EWSD in fall, suggesting a mix of short and long-distance driving.

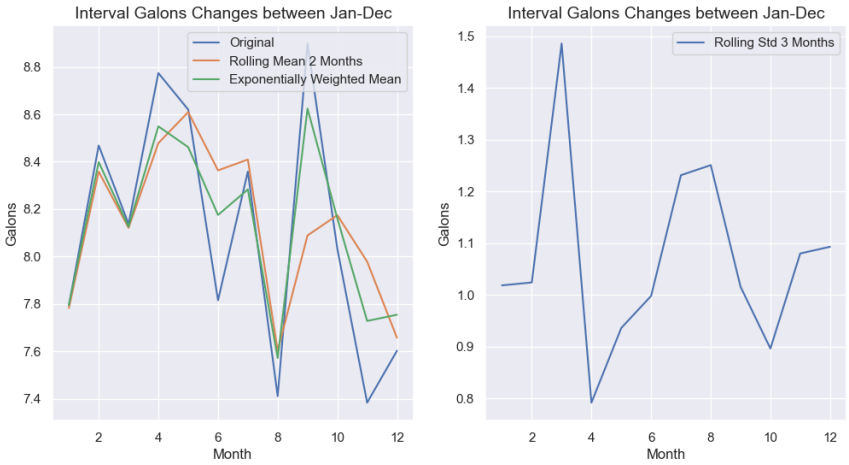

Figure 5: Interval Gallons by Month

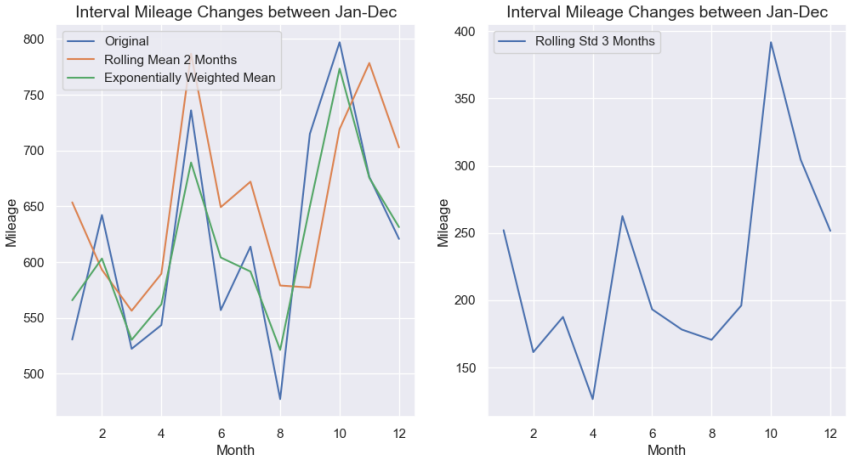

Figure 6: Interval Mileages by Month

Lastly, an examination of the average interval gallons filled up over the week reveals that refueling predominantly occurs mid-week, specifically on Wednesday or Thursday (weekday numbers are labeled as monday = 1 and sunday = 7). This pattern reflects a consistent trend in the driver's weekly behavior.

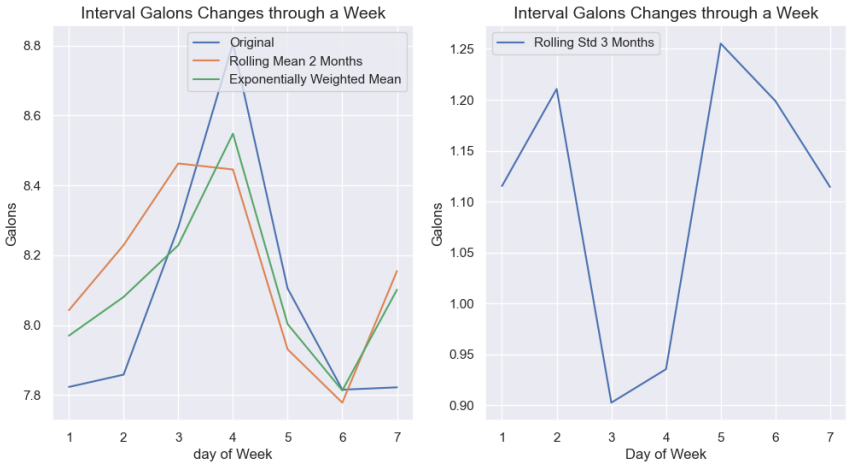

Figure 7: Interval Gallons by Day of Week

4.2 Paired T Test

To determine whether the computer-based MPG values are consistently overestimated, we conducted a paired t-test.

The hypotheses for the test are defined as follows:

- Null hypothesis (H_0): The mean difference between the computer-generated MPG and the actual MPG is zero. Formally, $H_0 : \mu_{diff} = 0$.
- Alternative hypothesis (H_1): The mean difference between the computer-generated MPG and the actual MPG is greater than zero, indicating that the computer-generated MPG values are typically higher than the actual MPG values. Formally, $H_1 : \mu_{diff} > 0$.

Upon running the t-test, we obtained the following results:

Statistic	Value
Test Statistic (t)	20.434
p-value	$< 2.2 \times 10^{-16}$
Degrees of Freedom (df)	100
Mean Difference	4.728713

Table 4: Paired t-test Results

Given the extremely low p-value and the fact that the 95% confidence interval for the mean difference does not include 0 and is entirely above 0, we reject the null hypothesis in favor of the alternative hypothesis. This indicates significant statistical evidence to conclude that the computer-generated interval MPG is, on average, 4.73 MPG greater than the actual interval MPG.

4.3 Multiple Linear Regression Analysis of Gas Consumption

We conducted a multiple linear regression analysis to examine the relationship between gas usage over time intervals (measured in gallons) and other associated factors. The formulated model is delineated as follows:

$$\hat{Y} = 51.927 + 2.303X_{\text{lag}} - 5.962\log(X_{\text{miles, lag}}) + \varepsilon \tag{1}$$

where:

- $\hat{Y}$ signifies the Box-Cox transformed interval gallons, serving as the dependent variable.
- 51.927 is the intercept of the model, representing the expected transformed interval gallons when other predictors are not considered.
- X_{lag} is the 1-period lag of interval gallons, serving as an explanatory variable to measure the influence of past gas usage on current levels.
- 2.303 quantifies the effect of a one-unit increase in X_{lag} on $\hat{Y}$, holding other factors constant.
- $\log(X_{\text{miles, lag}})$ is the natural log transformation of the 1-period lag in interval mileage data, employed to stabilize the variance and achieve a linear relationship with the response variable.
- -5.962 measures the influence of a one-unit change in $\log(X_{\text{miles, lag}})$ on $\hat{Y}$.
- ε embodies the error term, encapsulating random fluctuations not explained by the regression model.

The coefficients were estimated using the least squares method. Subsequent residual diagnostics will be executed to affirm that the model satisfies the core assumptions of linear regression, encompassing linearity, independence, equal variance (homoscedasticity), and normal distribution of the error term.

5 Conclusion

This study focus on finding patterns within vehicular gas consumption data, aiming to derive insights well rounded understanding of gas usage. Through these data visualization, we identified seasonal patterns in interval mileage and interval gallons over the years. The application of a paired t-test revealed a statistically significant overestimation in computer-generated interval mileage compared to actual measurements, indicating a systematic bias in the auto-generated mileage figures. Our exploration further led us to employ a multiple linear regression model, focusing on interval gallons as the dependent variable. This model stands out for its simplicity and interpretability, proving to be a fitting method to predict future gas consumption. By analyzing variables such as the lag of interval gallons and mileage, we could discern underlying trends that influence gas usage. The findings from this study not only highlight the discrepancies in computed vs. actual mileage but also pave the way for refining prediction models for gas consumption. Future work could benefit from integrating more nuanced data and advanced modeling techniques to enhance the accuracy and applicability of consumption predictions. As we refine our models and understanding, we inch closer to achieving more sustainable and efficient gas usage patterns, resonating with the broader goal of optimizing resource utilization in our daily lives.

6 Appendix

Normality, linearity, homoscedasticity, and independence checks, alongside VIF calculations for predictors, are essential for validating the model's assumptions and applicability. Despite a low adjusted R^2, the model's AIC of 466.95 suggests its utility in forecasting gas consumption, with caution advised due to potential overfitting or underfitting.

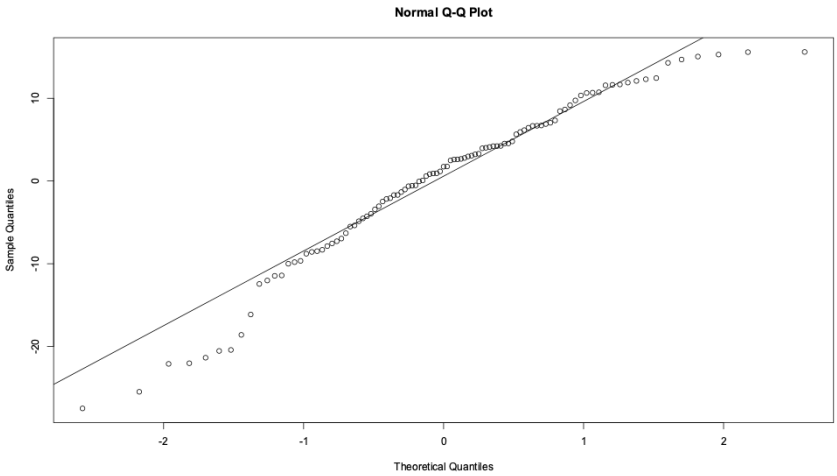

Figure 8: Normality Check - QQ Plot

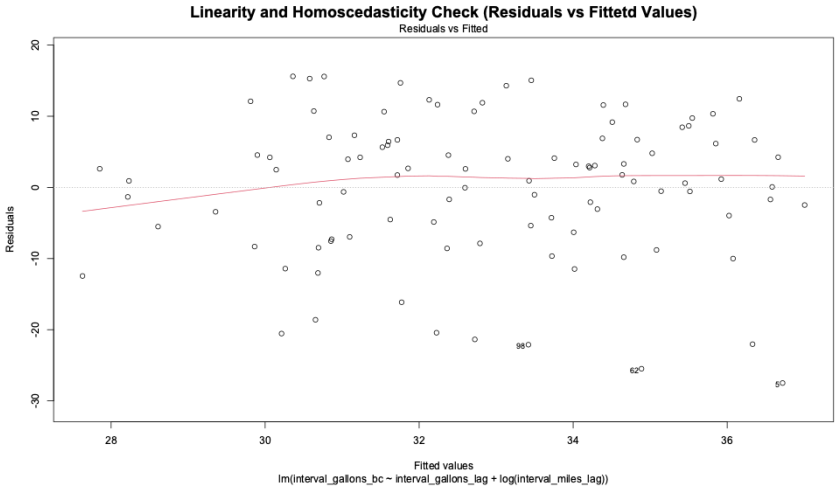

Figure 9: Linearity and Homoscedasticity Check - Residual Plot

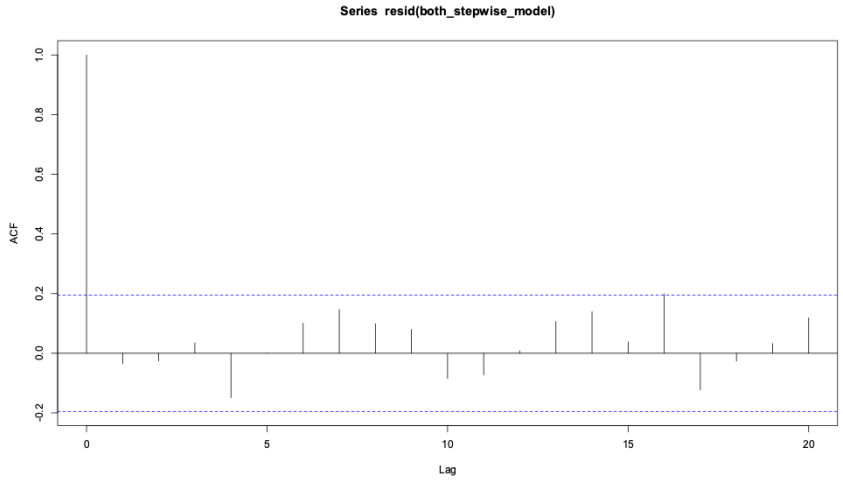

Figure 10: Independence Autocorrelation Check

The computed VIF values for the predictors are as follows:

Predictor	VIF Value
Interval Gallons Lag	1.848698
Log(Interval Miles Lag)	1.848698

Table 5: Variance Inflation Factor (VIF) for Each Predictor

To fit the model and want to interpret the expected future gallons in their original scale, it needs to be applied the inverse of the Box-Cox transformation to the predicted values.
The Box-Cox transformation is defined as:

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

The inverse transformation, which you apply to the predicted values, is:

$$y = \begin{cases} (\lambda y(\lambda) + 1)^{1/\lambda} & \text{if } \lambda \neq 0 \\ e^{y(\lambda)} & \text{if } \lambda = 0 \end{cases}$$

Where $y(\lambda)$ is the transformed value, y is the original value, and λ is the parameter used for the Box-Cox transformation.