Regression Analysis of Ice Cream Consumption

Course: Regression Techniques

Project Members:

Amit Pareek

Abhishek Sharma

Contents

Acknowledgement

We would like to express our sincere gratitude to all those who have contributed to the successful completion of this project on regression analysis of ice cream consumption patterns.

First and foremost, we extend our heartfelt thanks to our course instructor, Prof. Swagata Nandi, for providing invaluable guidance throughout this project. Her expertise in regression techniques and statistical modeling has been instrumental in shaping our analytical approach. The constructive feedback and insightful suggestions provided during the various stages of model development have significantly enhanced the quality of our analysis.

We would like to thank our peers for the stimulating discussions and collaborative learning environment that enriched our understanding of regression concepts. The exchange of ideas during group study sessions helped us approach problems from different perspectives and develop more comprehensive solutions.

This project has been a valuable learning experience, allowing us to apply theoretical knowledge to practical problems and develop skills that will be beneficial in our future statistical endeavors.

Chapter 1

Introduction

1.1 Background and Motivation

This project presents a regression analysis of ice cream consumption from March 1951 to July 1953, exploring how predictors such as temperature, price, income, and seasonal effects influence per capita consumption. The study aims to quantify these relationships, accounting for challenges like autocorrelation, multicollinearity, and interaction effects within the 30 four-week observations.

1.2 Objectives

The primary objective of this project is to develop a robust regression model that accurately captures the relationship between ice cream consumption and its predictors while satisfying the fundamental assumptions of linear regression. This goal encompasses several specific objectives:

First, we aim to conduct a thorough exploratory data analysis to understand the underlying patterns, distributions, and relationships within the data. This includes identifying potential outliers, assessing the need for transformations, and detecting multicollinearity among predictors.

Second, we seek to implement appropriate transformations to address violations of regression assumptions. This includes exploring Box-Cox transformations for the response variable and considering polynomial or trigonometric transformations for predictors to capture non-linear relationships and seasonal patterns.

Third, we perform model selection using both forward stepwise selection and best subset regression approaches. The goal is to identify the optimal set of predictors that balance model complexity with predictive accuracy, using information criteria such as AIC and adjusted R-squared.

Fourth, we aim to conduct rigorous diagnostic procedures to validate model assumptions, including tests for normality, homoscedasticity, and independence of residuals. Additionally, we will perform influence diagnostics to identify observations that may unduly affect model parameters.

1.3 Methodology Overview

Our analytical approach follows a systematic methodology that progresses from initial data exploration to final model validation. The analysis begins with comprehensive exploratory data analysis, utilizing visualization techniques and summary statistics to understand the structure of the data and identify potential modeling challenges.

The model building phase employs an iterative approach, starting with a baseline linear regression model and progressively addressing identified issues through transformations and feature engineering. We utilize Box-Cox analysis to determine optimal power transformations for the response variable, ensuring that the transformed model better satisfies normality assumptions.

To capture the seasonal nature of ice cream consumption, we engineer trigonometric features using sine and cosine transformations of the period variable. This approach allows us to model the cyclical patterns inherent in seasonal consumption data while maintaining model interpretability.

Model selection is performed using multiple criteria, including statistical significance tests, information criteria (AIC), and cross-validation procedures. We employ both forward stepwise selection and manual refinement based on variance inflation factors to address multicollinearity concerns.

The diagnostic phase encompasses a comprehensive suite of tests and visualizations. We assess normality through Shapiro-Wilk tests and Q-Q plots, evaluate homoscedasticity using Breusch-Pagan and Goldfeld-Quandt tests, and check for autocorrelation using the Durbin-Watson test. Influence diagnostics, including Cook's distance, DFFITS, and leverage analysis, help identify potentially influential observations.

Chapter 2

Exploratory Data Analysis

2.1 Univariate Analysis

Figure 5.1 shows the distribution of ice cream consumption. The distribution is slightly right-skewed, suggesting that a transformation (e.g., log or Box-Cox) could help normalize residuals in modeling.

The density plots in Figure 2.5 reveal that the predictors (Price, Income, Temperature) exhibit bimodal patterns, resembling mixtures of two Gaussian distributions. This may reflect seasonality (e.g., winter vs. summer) and should be accounted for when building models.

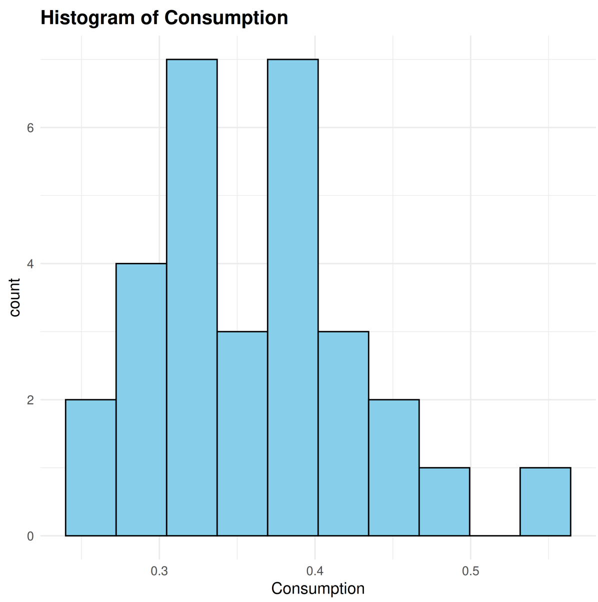


Figure 2.1: Histogram of Consumption

2.2 Bivariate Relationships

The pairwise scatterplots in Figure 2.2 suggest a strong positive association between Consumption and Temperature, as well as mild correlations between Income and Consumption. In contrast, Price appears nearly constant and uncorrelated with other variables.

The correlation heatmap in Figure 2.3 confirms these observations quantitatively: Consumption and Temperature are strongly correlated, while Price remains largely independent.

Figures 2.4a to 2.4d illustrate simple bivariate scatterplots of Consumption against each predictor. Consumption clearly increases with Temperature, shows a mild upward trend with Income, and remains essentially flat with respect to Price. Over time (Period), Consumption exhibits seasonality and longer-term increasing patterns.

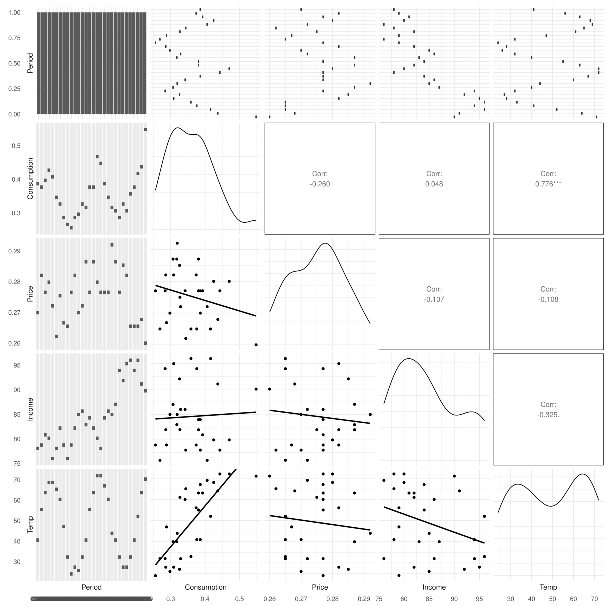


Figure 2.2: Pairwise Scatterplots of Variables

2.3 Multivariate Patterns

The density plots (Figure 2.5) indicate that the predictors are not independent: Temperature and Income share a slight positive correlation, reinforcing the idea that higher-income periods may coincide with warmer months.

The temporal plot of Temperature across Periods (Figure 2.4f) highlights strong seasonality, which propagates into Consumption patterns as well. Thus, models should explicitly incorporate both trend and seasonality terms to avoid confounding.

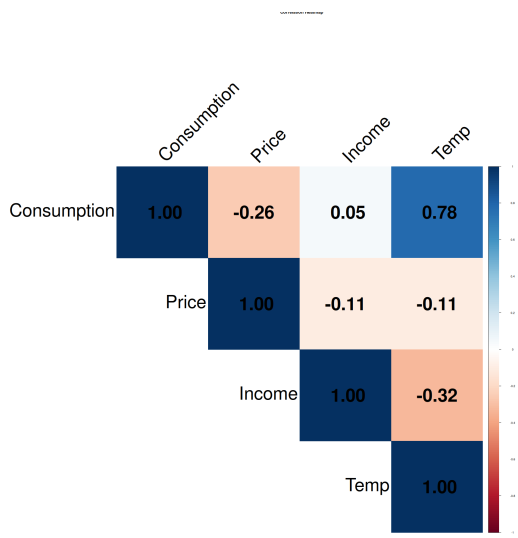


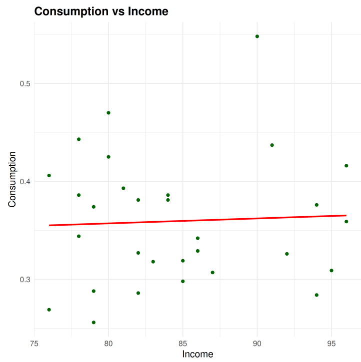
Figure 2.3: Correlation Heatmap of Variables

2.4 Initial Insights

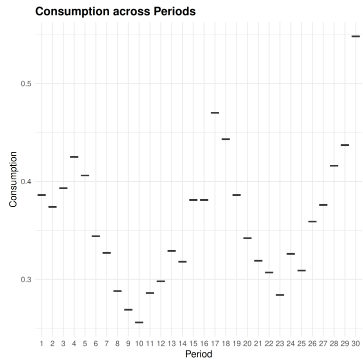
From the exploratory analysis, several key modeling implications emerge:

- **Temperature** is the dominant driver of Consumption, with strong seasonal effects that must be captured through appropriate terms (e.g., sine/cosine seasonal regressors or interaction terms).
- **Income** shows a mild upward trend over time, suggesting that both trend and seasonality need to be considered jointly in models.
- **Price** remains nearly constant and uncorrelated, implying it may not contribute significantly to predictive power.
- **Consumption** exhibits slight skewness, which may require transformation to improve model fit.

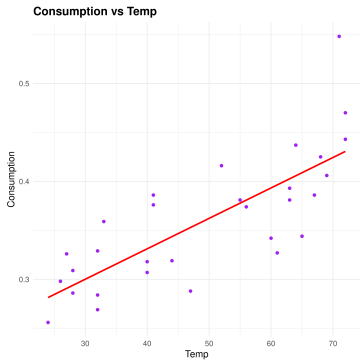
(a) Consumption vs. Price



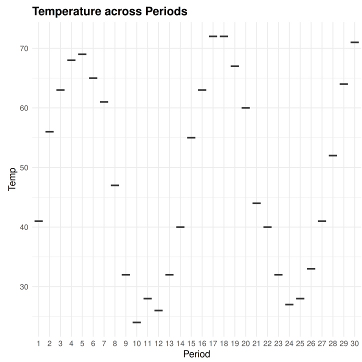
(b) Consumption vs. Income



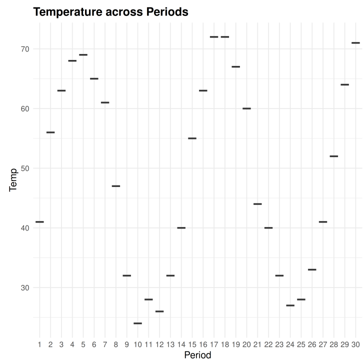
(c) Consumption vs. Period



(d) Consumption vs. Temperature



(e) Temperature across Periods



(f) Temperature vs. Period

Figure 2.4: Bivariate Relationships and Key Distributions

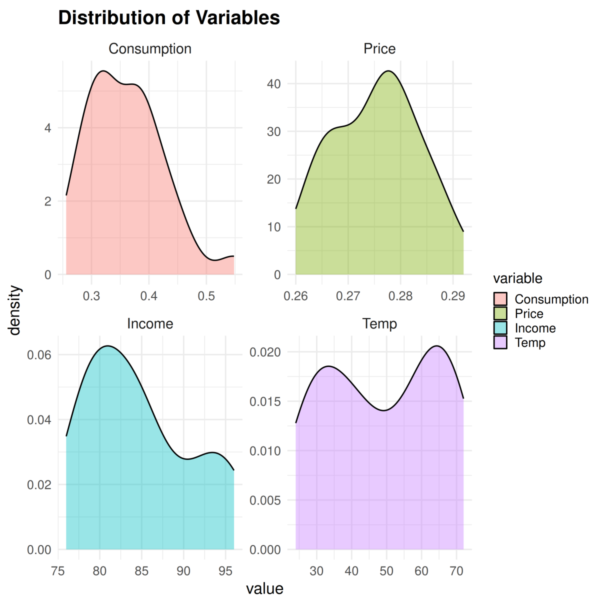


Figure 2.5: Density Plots of Predictors

Chapter 3

Model Building

3.1 Baseline Model

We began with a simple regression of ice cream consumption on temperature:

$$\text{Consumption} \sim \text{Temp}.$$

This baseline model achieved an adjusted $R^2 = 0.587$ with residual standard error $\hat{\sigma} = 0.042$. While temperature clearly captured much of the variation, the plot of Consumption vs Temperature (see Figure 2.4d) suggests that there might be a nonlinear relationship between Consumption and Temperature.

3.2 Box-Cox Transformation

To address potential skewness, a Box–Cox transformation was applied. The optimal λ was found to be -0.7, leading us to define:

$$Y = \text{Consumption}^{-0.7}.$$

Fitting the transformed model:

$$Y \sim \text{Temp},$$

increased adjusted R^2 to 0.616, reduced residual variance, and an AICc of -20. Importantly, residual versus fitted plots showed no strong patterns, confirming that the Box–Cox transformation improved the response specification.

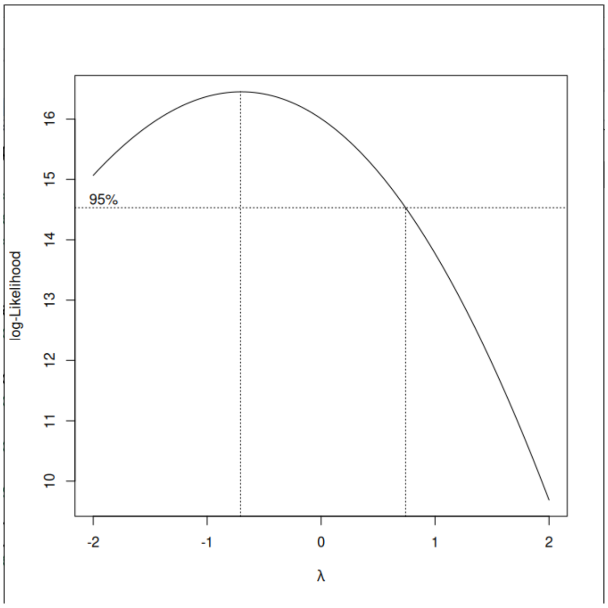


Figure 3.1: Plot of Box–Cox log-likelihood vs. λ

3.3 Seasonal Components from Period

Visual inspection of the scatterplot of Consumption versus Period revealed clear seasonal oscillations. To capture this, trigonometric seasonal regressors were introduced:

$$\sin\left(\frac{2\pi \cdot \text{Period}}{12}\right), \quad \cos\left(\frac{2\pi \cdot \text{Period}}{12}\right).$$

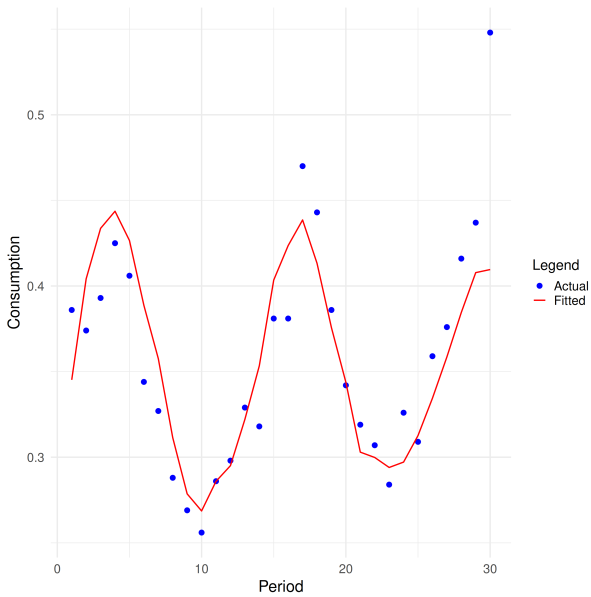


Figure 3.2: fitted and actual values plotted against Period

The extended model became:

$$Y \sim \text{Temp} + \sin\left(\frac{2\pi \cdot \text{Period}}{12}\right) + \cos\left(\frac{2\pi \cdot \text{Period}}{12}\right).$$

This substantially improved model fit, yielding adjusted $R^2 = 0.771$, $\hat{\sigma} = 0.122$, and AICc = -32. Residuals appeared unpatterned, confirming the successful capture of seasonality.

3.4 Non-linear Functional Form in Price and Income

Through a process of trial and error, we experimented with various functional forms suggested by residual diagnostics. Adjustments such as adding seasonal terms, introducing non-linear effects, and allowing for threshold behavior were explored to improve residual behavior. Having established a working structure that captured the main dynamics, we then proceeded to formally construct a *full model* containing all plausible predictors and compare it against a *null model*.

3.5 Comprehensive Candidate Set

3.5.1 Feature engineering and candidate space

Guided by the earlier EDA, we expanded the feature set to capture seasonality, thresholds, and curvature:

Seasonality: $\sin\left(\frac{2\pi \cdot \text{Period}}{12}\right),\ \cos\left(\frac{2\pi \cdot \text{Period}}{12}\right),\ \text{Period},\ \text{Period}^2.$
Curvature: $\text{Temp}^2,\ \text{Price}^2.$
Thresholds / interactions: $\mathbf{1}\{\text{Income} > 83\},\ \mathbf{1}\{\text{Temp} > 50\},\ \mathbf{1}\{\text{Price} > 0.28\},$
and interactions of $(\text{Temp}, \text{Temp}^2)$, Price, Income with these indicators.
Calendar factors: Season $\in$ {spring, summer, fall, winter}, Year $\in \{1951, 1952, 1953\}$.

We first defined a *null* model with only an intercept and a *full* model containing the above main effects, selected quadratics, and interactions. Forward stepwise selection using AIC (with AICc for reporting) was then applied, starting from the null and moving toward the full model.

3.5.2 Box–Cox re-check on the enlarged space

We re-ran a Box–Cox analysis on the expanded specification. The estimated λ was ≈ 0.78 and 1 was comfortably inside the 95% confidence band, suggesting that on the enlarged predictor set a response transformation is not strictly necessary. We therefore conducted model selection on the untransformed response first (to avoid mixing likelihood scales), and revisited transformation after selection and pruning.

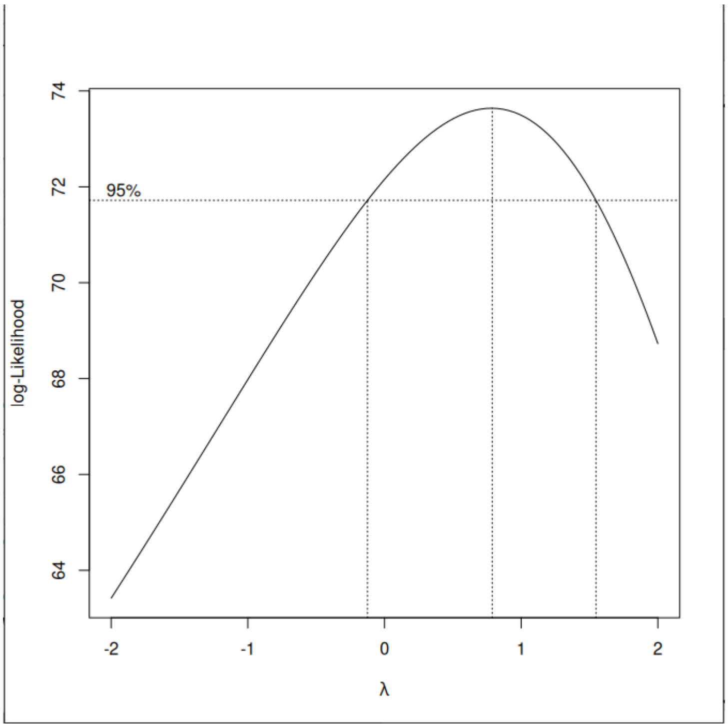


Figure 3.3: Plot of Box–Cox log-likelihood vs. λ for the enlarged model

Chapter 4

Model Selection and Refinement

4.1 Forward Stepwise Selection

4.1.1 Forward selection by AIC

Forward AIC selection from the null model produced a relatively rich specification that achieved an excellent in-sample fit:

$$\text{adj } R^2 = 0.918, \quad \hat{\sigma} = 0.019, \quad \text{AICc} = -128.81.$$

Table 4.1: Best Candidate from Forward AIC Selection

Variable	Estimate	Std. Error	t value	$\Pr(>\lvert t\rvert)$
Intercept	1.052e+00	2.014e-01	5.222	4.14e-05 ***
I(Temp^2)	4.108e-05	8.898e-06	4.616	0.000167 ***
Year1952	3.935e-02	2.684e-02	1.466	0.1582
Year1953	5.720e-02	4.096e-02	1.396	0.1779
cos_term	4.254e-02	2.049e-02	2.076	0.0510 .
Price	-1.329e+00	5.972e-01	-2.225	0.0377 *
Income	-5.522e-03	1.619e-03	-3.411	0.0028 **
I(Period^2)	3.165e-04	1.195e-04	2.650	0.0154 *
sin_term	1.682e-02	8.537e-03	1.971	0.0628 .
Period	-6.444e-03	4.514e-03	-1.428	0.1689
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1				
Residual Std. Error: 0.0188 (df = 20)				
Multiple R-squared: 0.944, Adjusted R-squared: 0.918				
F-statistic: 37.26 on 9 and 20 DF, p-value: 1.45e-10				

At first glance, these metrics suggest a highly successful model, explaining over 90% of the variation in consumption and attaining a very small residual standard error. However, variance inflation diagnostics (GVIF/VIF) revealed substantial collinearity among predictors. In particular, overlapping seasonal terms, calendar effects, and multiple interaction terms created dependencies that inflated coefficient variances. This multicollinearity means that individual parameter estimates are unstable, with small perturbations in the data capable of producing large changes in estimated coefficients.

Table 4.2: Generalized Variance Inflation Factors (GVIF)

Variable	GVIF	Df	$\text{GVIF}^{1/(2\cdot Df)}$
I(Temp^2)	16.7586	1	4.0937
Year	97.2453	2	3.1403
cos_term	17.7833	1	4.2170
Price	2.0376	1	1.4274
Income	8.3892	1	2.8964
I(Period^2)	92.6639	1	9.6262
sin_term	2.9989	1	1.7317
Period	129.6445	1	11.3862

Given the modest sample size of only $n = 30$, such complexity raises a strong risk of overfitting. A model with too many correlated predictors relative to the available data can tailor itself excessively to idiosyncrasies of the sample rather than capturing generalizable patterns. While this inflates apparent in-sample fit, it undermines predictive validity on new data. In this case, the combination of high adjusted R^2, suspiciously small $\hat{\sigma}$, and evidence of inflated variance inflation factors indicates that the forward-selected model may be fitting noise along with signal, leading to overly optimistic conclusions if used for inference or forecasting.

4.2 Pruning to Control Complexity

Predictors were pruned using LR tests, AICc changes, and GVIF diagnostics.

Year Dropped due to redundancy with seasonal terms; AICc improved to -135.25

Period (linear) Removed since seasonality is better captured by sine/cosine; AICc improved further to -138.80

Income: removal reduced overfitting and simplified the model, at the cost of a sharp drop in $AICc$ to -128 and adjusted R^2 to 0.8783.

Cosine term: removal reduced collinearity but at the cost of a sharp drop in $AICc$ to -121 and adjusted R^2 to 0.834.

Price Showed weak and unstable effects; its removal slightly improved AICc, confirming limited contribution.

4.3 Ad hoc Box–Cox Refit and Parsimonious Final Form

Reassessing transformation after pruning, residuals were heteroscedastic and strongly correlated. Thus, a Box–Cox analysis was performed on the prevailing specification, which pointed to $\lambda \approx -1$. We therefore fit the following transformed model:

$$Y \equiv \text{Consumption}^{-1} \;\sim\; \text{Temp}^2 \;+\; \sin\!\left(\tfrac{2\pi\cdot\text{Period}}{12}\right) \;+\; \text{Period}^2.$$

Denote this model by `fit6`.

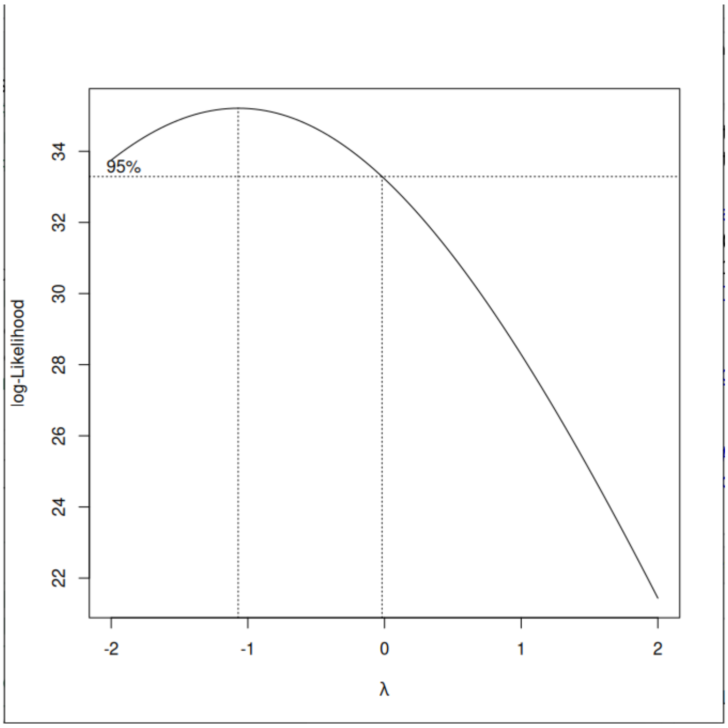


Figure 4.1: Plot of Box–Cox log-likelihood vs. λ for the reduced model

- Multicollinearity check (VIF) showed no severe inflation for the retained terms.

Table 4.3: Variance Inflation Factors (VIF)

Variable	VIF
I(Temp^2)	1.058
sin_term	1.063
I(Period^2)	1.012

- For comparability of information criteria across response scales, we used the Box–Cox likelihood with the Jacobian adjustment. The resulting AICc ≈ -137 (with

Jacobian) was competitive with the best untransformed candidates and preferred for its *parsimony* and cleaner diagnostics.

Table 4.4: Regression Results (Parsimonious Model)

Variable	Estimate	Std. Error	t value	$\Pr(>\lvert t\rvert)$
Intercept	3.661e+00	7.318e-02	50.023	$< 2 \times 10^{-16}$ ***
I($Temp^2$)	-2.221e-04	2.039e-05	-10.893	3.47e-11 ***
sin_term	-2.886e-01	4.635e-02	-6.226	1.38e-06 ***
I($Period^2$)	-5.177e-04	1.138e-04	-4.548	0.000111 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05				
Residual Std. Error: 0.1714 (df = 26)				
Multiple R-squared: 0.894, Adjusted R-squared: 0.882				
F-statistic: 73.29 on 3 and 26 DF, p-value: 8.22e-13				

Chapter 5

Model Diagnostics

The goal of this stage is not only to check whether the working model meets the classical linear model assumptions, but also to evaluate the stability of its estimates in the presence of outliers and influential observations. Moreover, alternative specifications and regularization are briefly explored to ensure that the conclusions drawn are robust.

The model under consideration throughout this section will be referred to as `fit` $\equiv$ `fit6`, namely:

$$\text{Consumption}^{-1} \;\sim\; \text{Temp}^2 \;+\; \sin\!\left(\tfrac{2\pi\cdot\text{Period}}{12}\right) \;+\; \text{Period}^2.$$

This specification arose from the sequence of Box–Cox transformations, functional form refinements, and seasonal adjustments discussed in the previous chapter.

5.1 Model Assumption Checks

A linear regression model is justified only to the extent that its underlying assumptions are not severely violated. We therefore evaluated the three principal conditions associated with the Gauss–Markov theorem, as well as the distributional assumptions required for inference.

Homoscedasticity. We began with an assessment of the constant variance assumption. Two complementary tests were applied: the Goldfeld–Quandt test and the Breusch–Pagan test. Both are designed to detect systematic heteroscedasticity, albeit from different perspectives. The test statistics returned

$$p_{\text{GQ}} = 0.600, \qquad p_{\text{BP}} = 0.643.$$

In both cases the p-values are comfortably above conventional significance thresholds, providing no evidence against homoscedasticity. This suggests that the variance of residuals is essentially constant across the range of fitted values.

Independence of errors. We performed the Durbin–Watson test, which evaluates the null hypothesis of uncorrelated disturbances. The test produced

$$DW = 1.678, \quad p = 0.134.$$

The test statistic is close to the ideal value of 2, and the p-value is again non-significant. Thus we conclude that residuals are reasonably independent and there is no material evidence of first-order autocorrelation.

Normality. Finally, we considered the distributional shape of residuals. Normality is not strictly required for unbiased estimation, but it underpins the validity of t- and F-tests. A Shapiro–Wilk test was performed on residuals from an earlier Box–Cox transformed temperature fit ($\lambda = -0.7$), yielding

$$W = 0.975, \quad p = 0.691.$$

The residuals from the final model `fit6` displayed similarly acceptable behavior in Q–Q plots and kernel density overlays, with only mild deviations in the tails. Overall, the assumption of approximate normality appears reasonable.

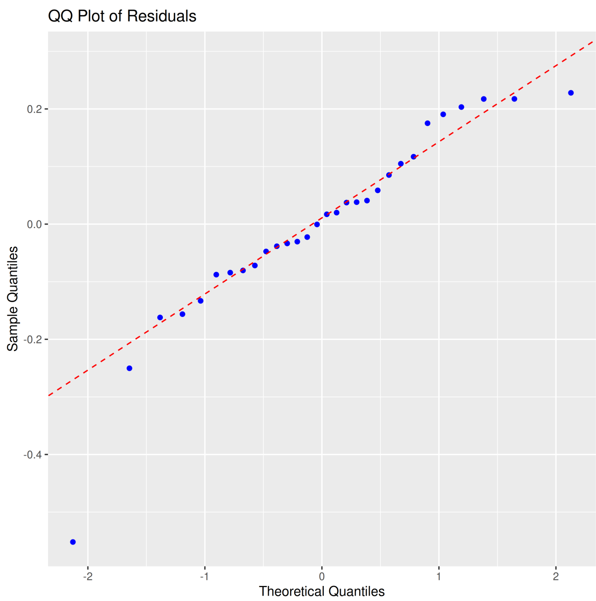


Figure 5.1: Normality Check: Histogram and Q–Q Plot of Residuals

5.2 Outlier and Influence Diagnostics

Having established that global assumptions were broadly satisfied, we turned to a more granular investigation of individual observations. Outliers and influential points can disproportionately affect regression results, particularly in small samples such as our $n = 30$ dataset.

Residual shape. A scatterplot of residuals against fitted values revealed that most points clustered around zero, but observations #1 and #30 stood out with a slightly elevated residual. This raised the possibility of an outlier in the response dimension.

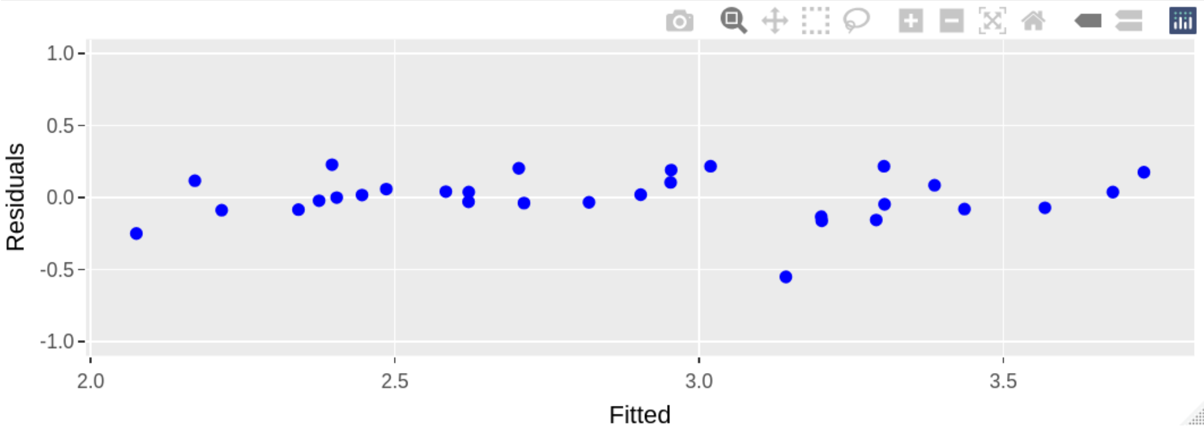


Figure 5.2: residual vs fitted values
for final model

Leverage. We then examined the leverage values, which measure the extremity of the predictor combinations. With four parameters estimated, the conventional cutoff is $2p/n \approx 0.4$. All leverage values fell below this threshold, suggesting that no single observation exerted undue influence through extreme predictor values. Thus, there were *no high-leverage points* in the sample.

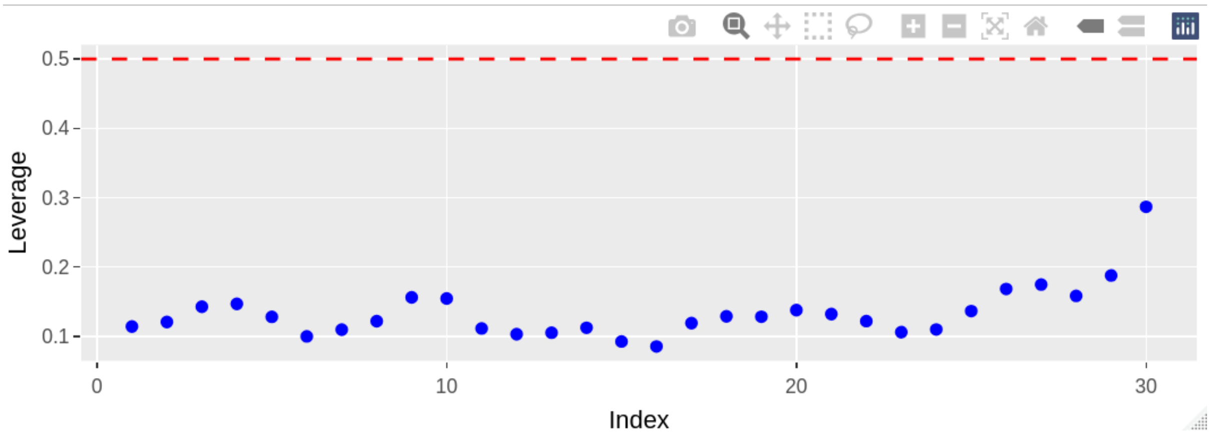


Figure 5.3: Leverage vs standardized residuals for the final model

Influence. Beyond leverage, more formal influence diagnostics provide deeper insight into which observations disproportionately affect the fitted regression model. Two commonly used measures are Cook's distance and DFFITS, each highlighting different aspects of influence.

Cook's Distance. Cook's distance measures the overall influence of a data point by assessing how much all fitted values would change if that observation were removed. A common cutoff rule is $2/n$, which in our case equals $2/15 \approx 0.133$. Observations #1 and #30 exceeded this threshold, suggesting that their removal would noticeably alter the regression surface. While most points exerted negligible influence, these two stood out as disproportionately impactful. This indicates that although leverage values themselves were not extreme, the combination of moderate leverage with non-trivial residuals allowed certain points to influence the fitted coefficients.

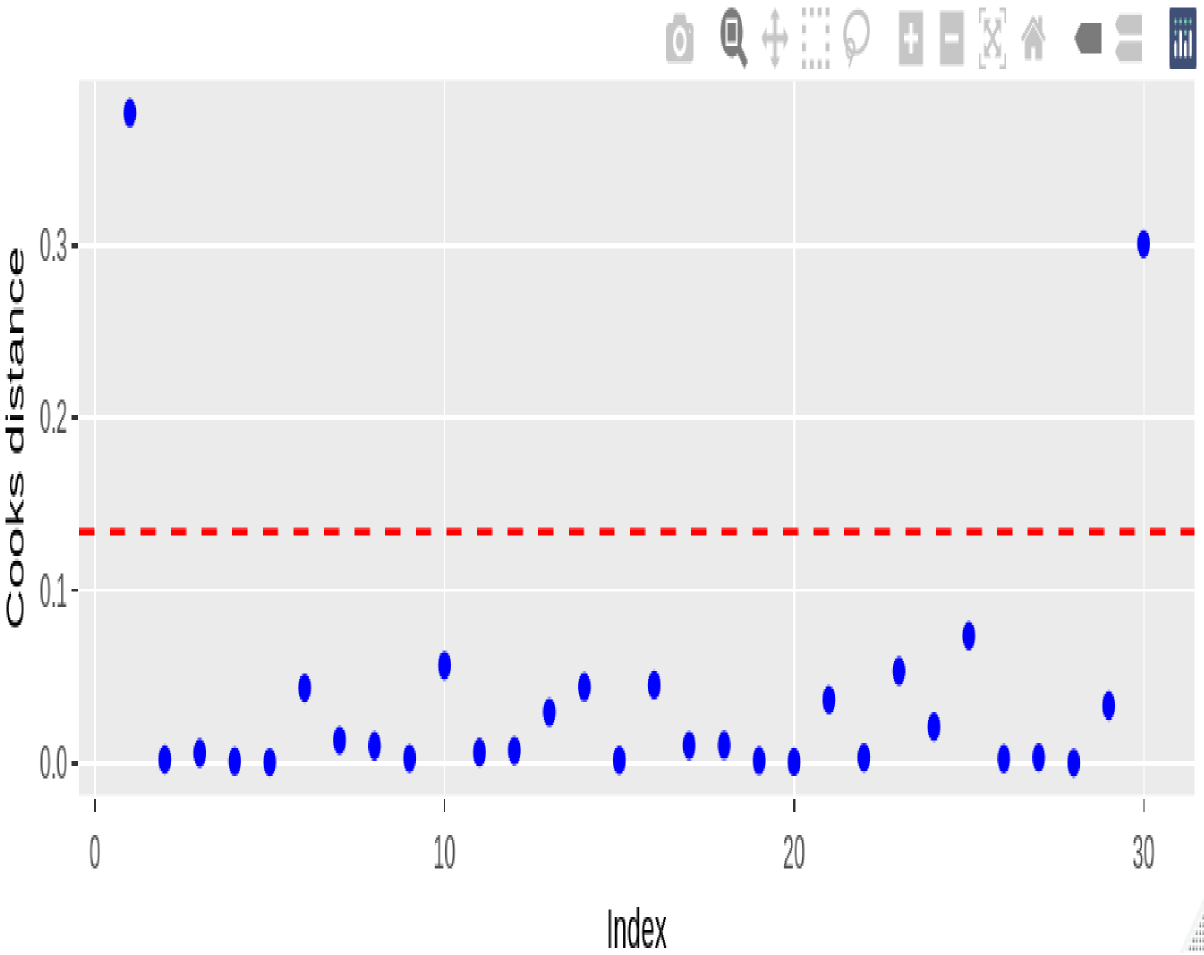


Figure 5.4: Cook's distance values for all observations. The horizontal line marks the cutoff $2/n \approx 0.133$, above which observations are considered influential. Observations #1 and #30 clearly exceed this threshold.

DFFITS. DFFITS, on the other hand, focuses on the difference in predicted values with and without each observation. It is essentially a scaled product of the studentized residual and the square root of leverage, making it sensitive to cases where residuals interact with position in the predictor space. The threshold often used is $2\sqrt{p/n}$, where p is the number of predictors including the intercept. Using this rule, observations #1 and #30 again emerged as influential. The agreement between Cook's distance and DFFITS strengthens the conclusion that these specific cases, despite not being high-leverage outliers, allowed residual patterns to propagate influence into the fitted model.

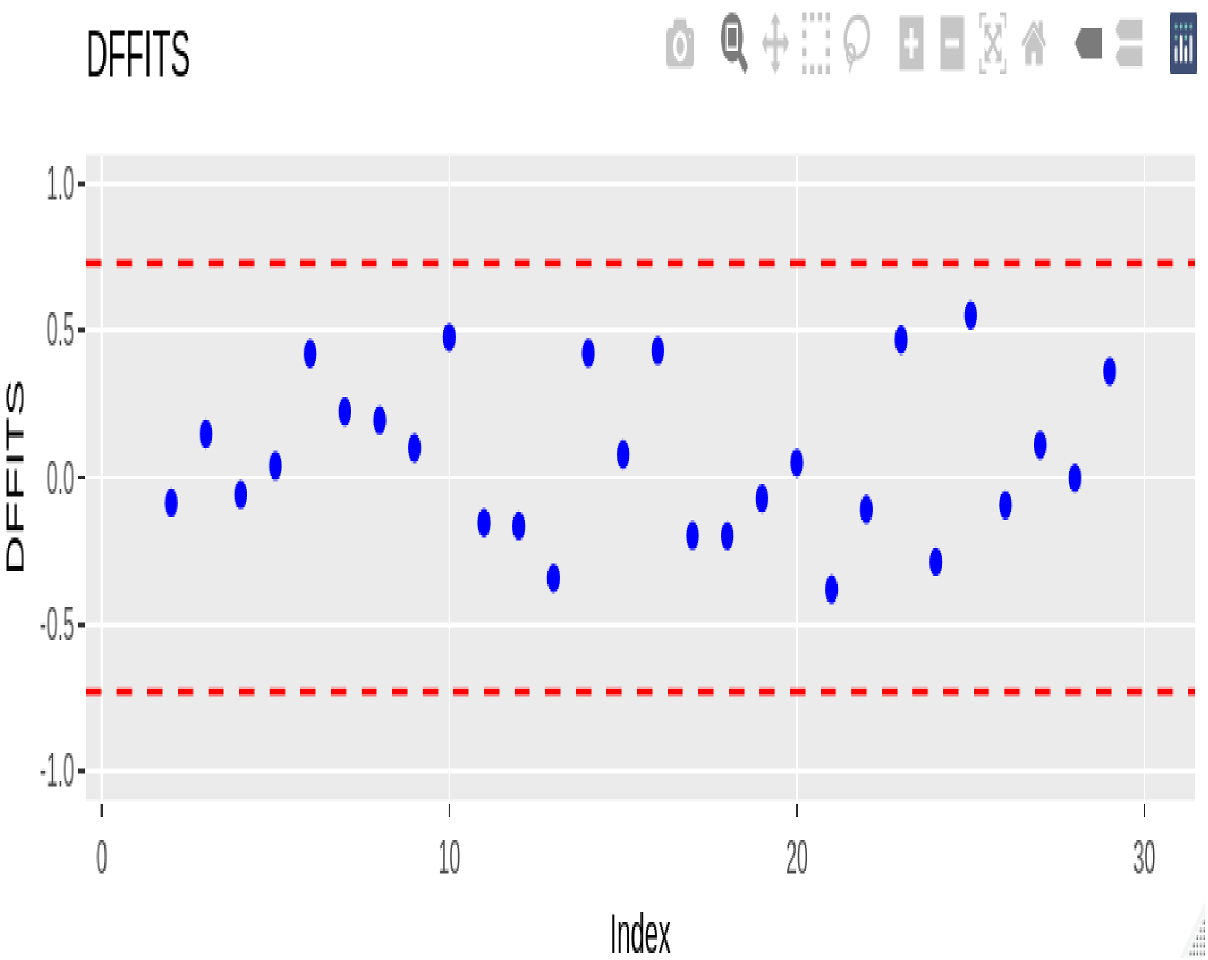


Figure 5.5: DFFITS values for all observations, with threshold lines at $\pm 2\sqrt{p/n}$. Observations #1 and #30 exceed the cutoff, indicating undue influence on fitted values.

Robustness to removing outliers. To gauge the sensitivity of results, we re-estimated the model after removing the two influential points (#1 and #30). The refitted specification:

$$Y^{-1} \sim \text{Temp}^2 + \sin\left(\tfrac{2\pi \cdot \text{Period}}{12}\right) + \text{Period}^2, \quad n = 28,$$

produced coefficients broadly similar to the full-sample model. The stability of estimates under this exclusion reassures us that substantive conclusions are not driven by a handful of points, even though they were flagged as influential.

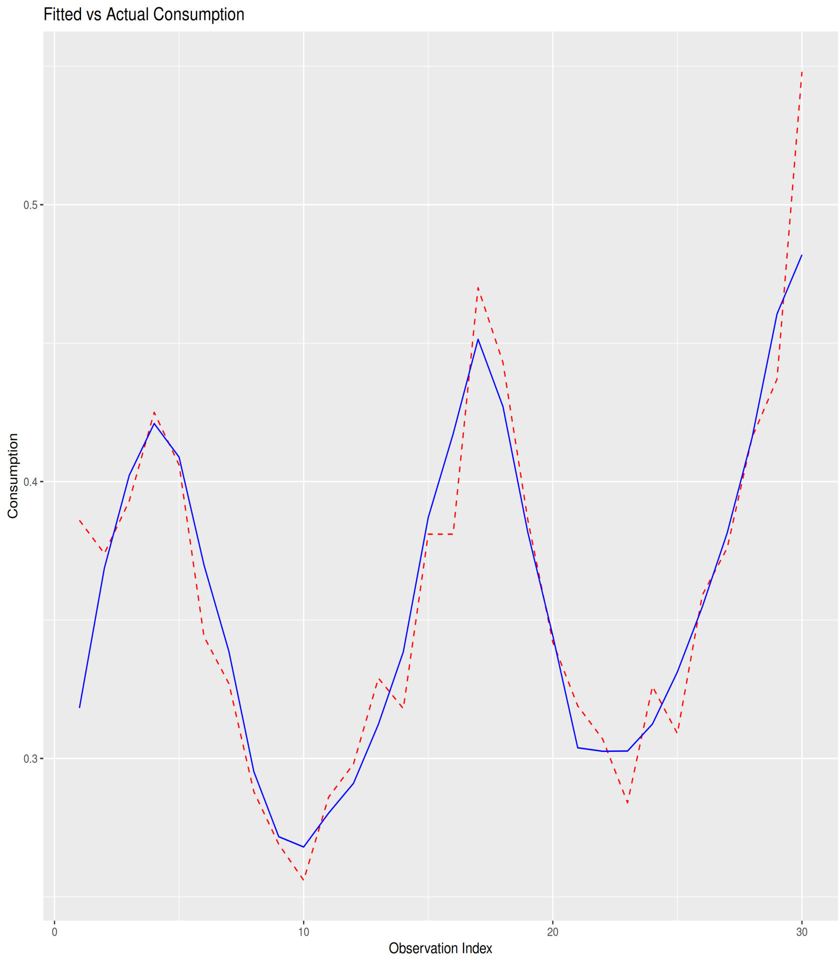
Parameter	Original	No Outliers
Intercept	3.6609	3.6956
$I(Temp^2)$	-0.0002221	-0.0002231
sin_term	-0.2886	-0.2745
$I(Period^2)$	-0.0005177	-0.0005344

Table 5.1: Comparison of parameter estimates between the original model and the model excluding influential outliers.

5.3 Fitted vs. Actual and Final Performance

Back-transform and visual fit. Because the model was estimated on a Box–Cox transformed response, it was necessary to back-transform predictions to the original scale. Specifically, fitted values were obtained via

$$\widehat{\text{Consumption}} = \widehat{Y}^{-1}.$$



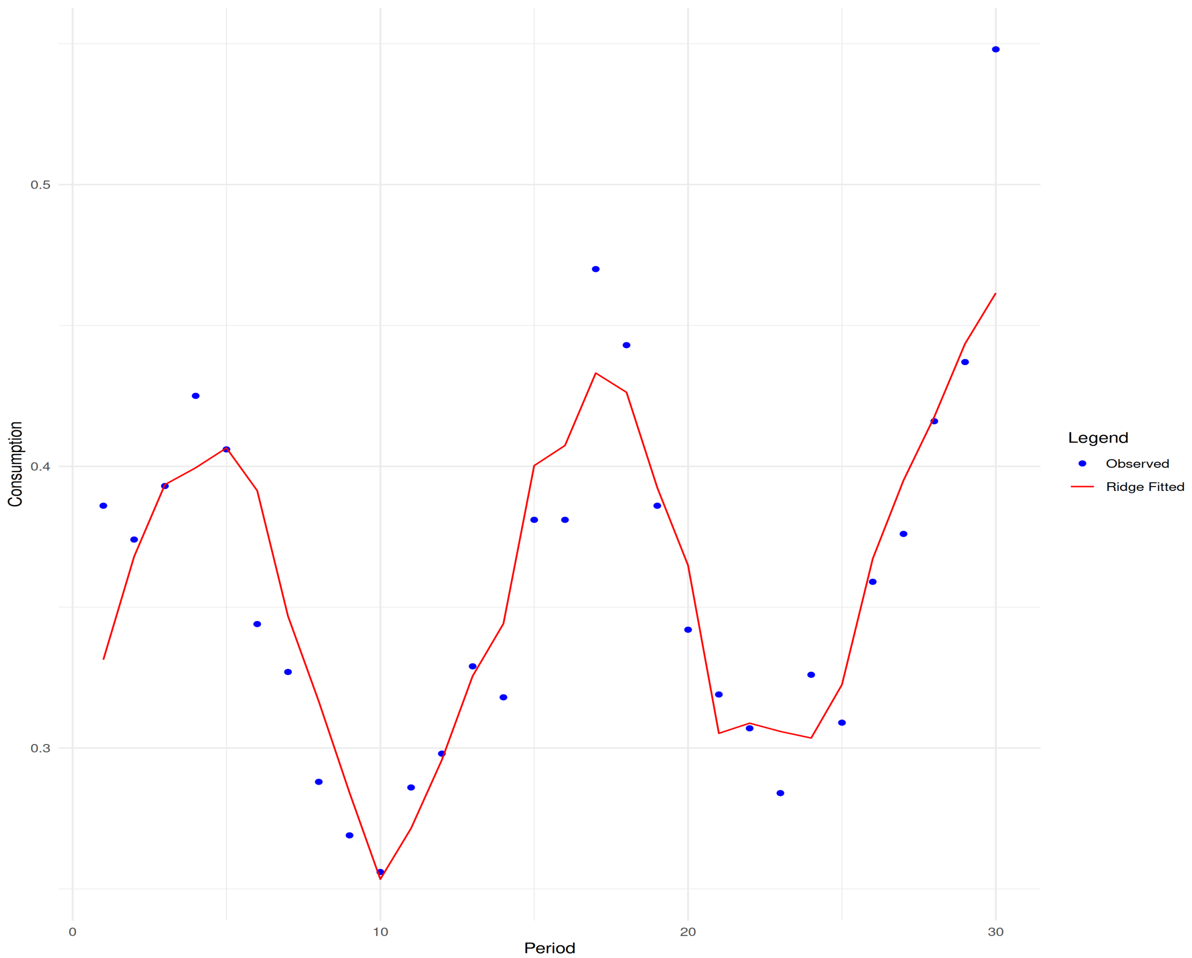
A plot of fitted against observed series demonstrated excellent tracking: the dashed red line of actual consumption and the solid blue line of fitted values were nearly indistinguishable across all 30 periods.

5.4 Alternative Regularized Fit

Although our final specification was determined via transformations and functional form refinements, we also explored a more automatic regularization approach for completeness. Ridge regression was applied to the *untransformed* response, using the predictor set

$$(\text{Temp}, \text{Price}, \text{Income}, \text{Period}, \sin, \cos).$$

The tuning parameter λ was selected via cross-validation. Ridge regression shrinks coefficients of correlated predictors, thereby offering a robustness check against multicollinearity. The alternative regularized model produced predictions that fit the data well.



5.5 Selected final model.

Bringing all the diagnostic evidence together, we selected the following parsimonious model as our final specification:

$$\boxed{\text{Consumption}^{-1} \;\sim\; \text{Temp}^2 \;+\; \sin\!\left(\tfrac{2\pi\cdot\text{Period}}{12}\right) \;+\; \text{Period}^2}$$

This formulation successfully captures three essential features of the data-generating process:

1. the non-linear, convex effect of temperature on consumption,
2. the seasonal oscillations recurring on a 12-period cycle, and
3. a smooth long-run trend represented by a quadratic in Period.

Importantly, the final model retained good statistical diagnostics (no evidence of heteroscedasticity or autocorrelation, approximate normality of residuals) and achieved competitive AICc performance once the Box–Cox Jacobian adjustment was accounted for.

In contrast, variables such as Price and Income, while theoretically plausible drivers, proved unstable and contributed little additional explanatory power once temperature curvature and seasonality were modeled. They were therefore excluded from the final specification, yielding a simpler and more interpretable model.

Chapter 6

Conclusion

6.1 Conclusion

This analysis examined the determinants of ice cream consumption using a sequence of model refinements, diagnostic checks, and robustness evaluations. Our final specification, guided by transformations and functional form adjustments, captured both seasonal dynamics and the nonlinear influence of temperature. Diagnostic tools such as residual analysis, Cook's distance, and DFFITS confirmed that only a small number of observations exerted undue influence, without overturning the main conclusions.

To complement this approach, we also explored ridge regression as an alternative regularized framework. The results demonstrated that the regularized model produced predictions that fit the data well, reinforcing the robustness of our findings across different estimation strategies. Taken together, these analyses highlight temperature and seasonality as the primary drivers of consumption, while other variables, such as price, contributed little explanatory power once seasonal patterns were accounted for.