Regression Analysis of Ice Cream Consumption

Course Project: *Regression Techniques*

Amit Pareek

Abhishek Sharma

Advisor: **Dr. Swagata Nandi**

Background and Motivation

Background and Motivation

Ice cream consumption is influenced by both **climatic** and **economic** conditions. Higher temperatures increase demand, while income and price reflect purchasing power and affordability. Understanding these relationships helps interpret market behavior and support better production planning.

Challenges:

- Small sample size ($n = 30$) limits model complexity.
- Strong seasonality and trend must be modeled explicitly.
- Possible multicollinearity (e.g., between Temperature and Income).

Scope of the Presentation

Scope of the Presentation

This presentation provides a structured overview of the regression analysis performed on the ice cream dataset. The content follows a logical sequence from exploration to interpretation:

1. **Exploratory Data Analysis (EDA):** Examine data structure, trends, and interrelationships among variables.
2. **Model Building:** Develop regression models, including polynomial and trigonometric terms for seasonality.
3. **Model Selection:** Compare candidate models using metrics like Adjusted R^2, AICc, and multicollinearity diagnostics (VIF).
4. **Model Diagnostics:** Check residual assumptions — normality, independence, and constant variance.
5. **Interpretation and Conclusion:** Identify key predictors and interpret how climatic and economic factors affect consumption.

Exploratory Data Analysis (EDA)

Data Description

Dataset Overview:

The dataset consists of 30 observations, each representing ice cream consumption for a four-week period from March 1951 to July 1953. It includes both economic and climatic variables that influence consumer behavior.

Variables:

- **Consumption:** Per-capita ice cream consumption (dependent variable).
- **Temperature:** Mean air temperature during the period.
- **Price:** Average price of ice cream per unit.
- **Income:** Per-capita income level.
- **Period:** Sequential index (1–30) denoting each four-week interval.

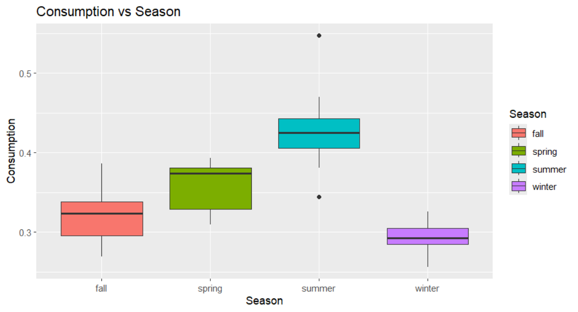
Dataset Overview

```
> head(ice)   # first 6 rows
  Period Consumption Price Income Temp Year Season
1      1       0.386 0.270     78   41 1951 spring
2      2       0.374 0.282     79   56 1951 spring
3      3       0.393 0.277     81   63 1951 spring
4      4       0.425 0.280     80   68 1951 summer
5      5       0.406 0.272     76   69 1951 summer
6      6       0.344 0.262     78   65 1951 summer
```

We have added two categorical variables: **Season** and **Year** to better capture how ice-cream consumption varies across different time periods and seasonal cycles.

```
> summary(ice)  # summary statistics
     Period       Consumption          Price            Income           Temp          Year        Season 
 Min.   : 1.00   Min.   :0.2560   Min.   :0.2600   Min.   :76.00   Min.   :24.00   1951:10   fall  :6  
 1st Qu.: 8.25   1st Qu.:0.3113   1st Qu.:0.2685   1st Qu.:79.25   1st Qu.:32.25   1952:13   spring:9  
 Median :15.50   Median :0.3515   Median :0.2770   Median :83.50   Median :49.50   1953: 7   summer:9  
 Mean   :15.50   Mean   :0.3594   Mean   :0.2753   Mean   :84.60   Mean   :49.10             winter:6  
 3rd Qu.:22.75   3rd Qu.:0.3912   3rd Qu.:0.2815   3rd Qu.:89.25   3rd Qu.:63.75                       
 Max.   :30.00   Max.   :0.5480   Max.   :0.2920   Max.   :96.00   Max.   :72.00                       
```

EDA: Seasonal Variation in Consumption

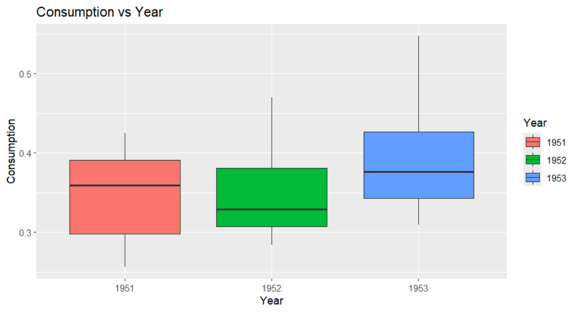


Consumption vs Season

The boxplot shows how Consumption varies across Seasons:

- Summer has the highest median consumption — peak ice-cream demand in warm months.
- A few outliers in Summer indicate exceptionally high consumption in certain periods.
- Winter shows the lowest consumption, consistent with lower temperatures.
- Spring and Fall show moderate consumption, lying between Summer and Winter.
- Overall, this seasonal pattern aligns with the earlier Temperature–Consumption relationship.

EDA: Yearly Variation in Consumption

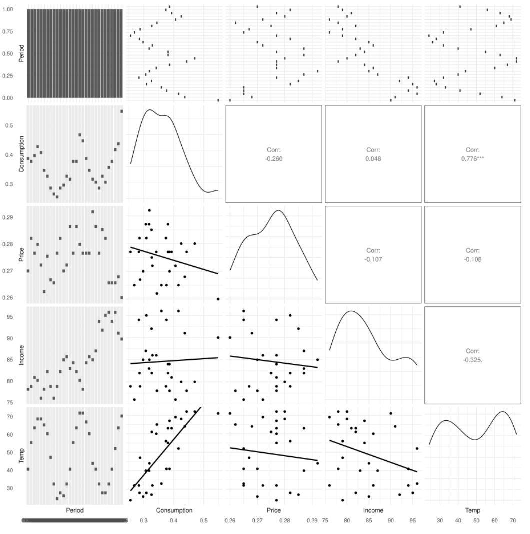


Consumption vs Year (1951–1953)

The boxplot compares Consumption levels across the years 1951–1953:

- 1953 shows the highest median consumption and greater variability, indicating an overall rise in ice-cream demand.
- 1951 and 1952 have relatively lower and more consistent consumption levels.
- This suggests a gradual increase in average consumption over time.

EDA: Bivariate Relationship



- The scatterplot matrix shows relationships among all key variables.
- Consumption is highly correlated with Temperature ($r = 0.78$) — higher temperatures lead to greater ice-cream consumption.
- Price and Income have only weak correlations with Consumption, indicating limited direct influence.
- The Period variable shows a clear seasonal pattern, with repeating fluctuations in both Consumption and Temperature.

EDA: Correlation Analysis

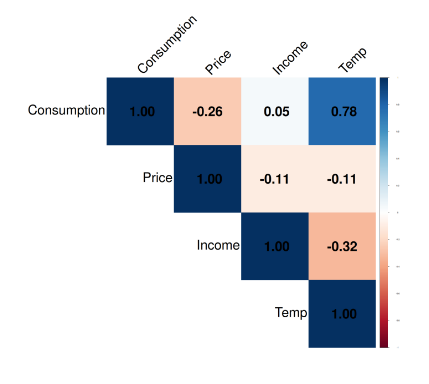
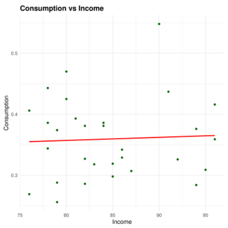


Figure 1: Correlation Heatmap of Variables

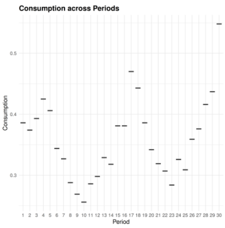
- The correlation heatmap confirms the earlier scatterplot findings quantitatively.
- Consumption and Temperature show a strong positive correlation (r = 0.78), indicating higher sales during warmer periods.
- Price has a mild negative correlation with Consumption (r = -0.26), while Income shows almost no direct influence (r = 0.05).
- Temperature and Income are moderately negatively correlated (r = -0.32).
- Overall, Temperature emerges as the dominant factor driving ice-cream consumption.

EDA: Bivariate Scatterplots

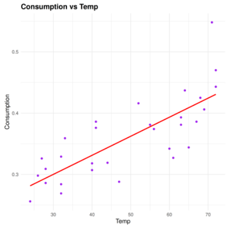
(a) Consumption vs. Price



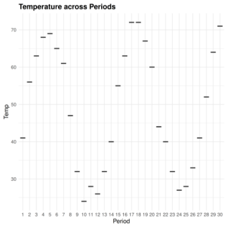
(b) Consumption vs. Income



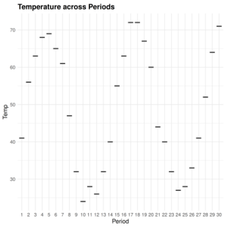
(c) Consumption vs. Period



(d) Consumption vs. Temperature

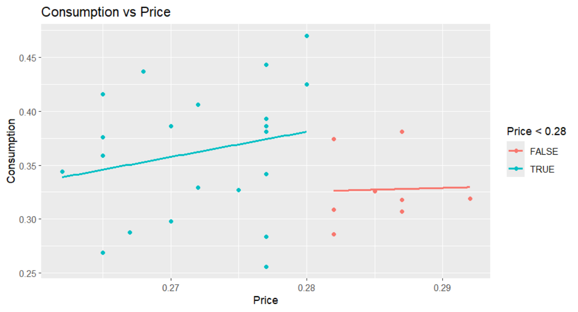


(e) Temperature across Periods



(f) Temperature vs. Period

EDA: Consumption vs Price

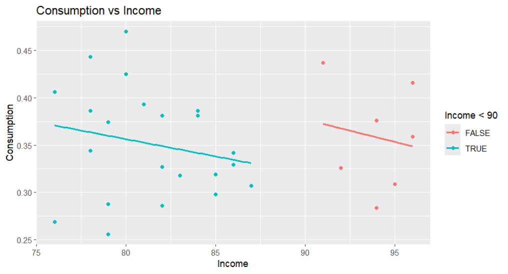


The plot shows the relationship between Consumption and Price, with the dataset divided at Price $= 0.28$.

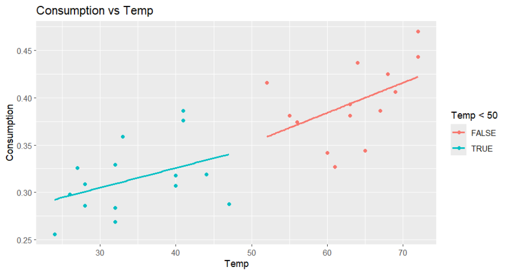
For Price < 0.28, Consumption shows a slightly increasing trend, indicating that lower prices are generally associated with higher consumption.

For Price > 0.28, the trend becomes almost flat, suggesting that beyond this point, further price increases do not significantly affect consumption. .

EDA: Consumption vs Income and Temperature



Consumption vs Income



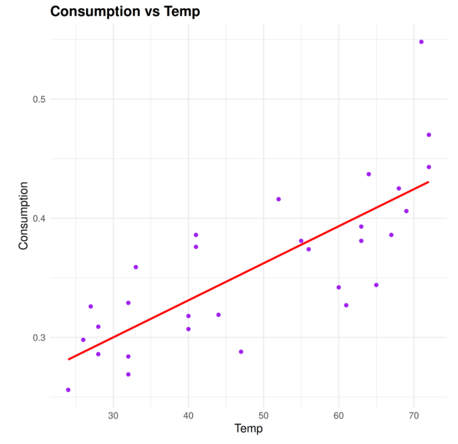
Consumption vs Temperature

EDA: 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** shows a nonlinear pattern — Consumption increases up to a certain Price level and then becomes almost flat.
- **Consumption** exhibits slight skewness, which may require transformation to improve model fit.

Baseline Model



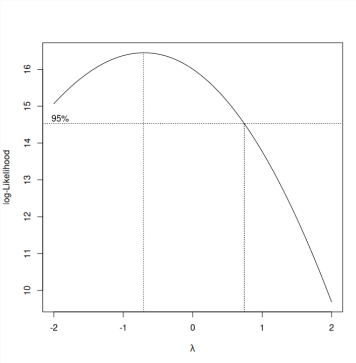
Consumption vs Temperature

- A simple linear regression of Ice Cream Consumption on Temperature:

$$\text{Consumption} \sim \text{Temperature}$$

- Achieved an Adjusted $R^2 = 0.587$ and Residual Standard Error $\hat{\sigma} = 0.042$.
- **Interpretation:** Temperature alone explains a large portion of the variation in consumption.
- However, the scatterplot indicates a possible nonlinear relationship, suggesting that a simple linear model may not fully capture the trend.

Box–Cox Transformation



Box–Cox

Transformation Plot

- To address potential skewness in the response variable, a Box–Cox transformation was applied.
- The optimal transformation parameter was found to be $\lambda = -0.7$, leading to the transformed variable:

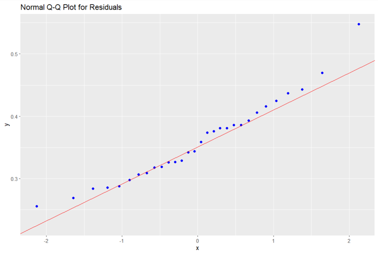
$$Y = \text{Consumption}^{-0.7}$$

- The transformed model was fitted as:

$$Y \sim \text{Temperature}$$

- The transformation resulted in:
  - Increased adjusted $R^2 = 0.616$
  - Reduced residual variance
  - AICc = -20.65
- Residual versus fitted plots showed no strong patterns, confirming that the transformation improved the response specification.

Model Diagnostics: Normality Check



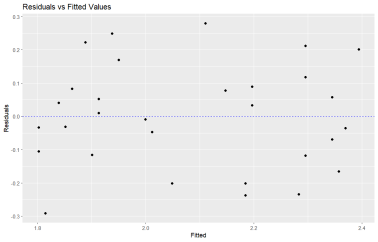
Normal Q–Q Plot of Residuals

- To assess the normality assumption, we applied the Shapiro–Wilk test to the residuals of the fitted model.

$$W = 0.97528, \quad p\text{-value} = 0.6908$$

- Since the p-value is much greater than 0.05, we fail to reject the null hypothesis of normality.
- The Q–Q plot also shows that most residuals lie close to the reference line, with no systematic deviation at the tails.
- These results indicate that the residuals are approximately normally distributed, satisfying the normality assumption of linear regression.

Model Diagnostics: Homoscedasticity Check



Residuals vs Fitted Values Plot

- To verify the assumption of constant variance, two tests were applied:
  - Goldfeld–Quandt test: $GQ = 0.9658$, $p = 0.5245$
  - Breusch–Pagan test: $BP = 0.0512$, $p = 0.8209$
- Both p-values are much greater than 0.05, indicating no significant heteroscedasticity.
- The residuals–fitted plot shows points randomly scattered around zero, suggesting that the variance of residuals remains approximately constant.
- Therefore, the homoscedasticity assumption holds true for this model.

Model Diagnostics: Independence of Errors

- To test whether residuals are independent, we performed the Durbin–Watson test.

$$\text{Autocorrelation} = 0.6207, \quad \text{D–W Statistic} = 0.5576, \quad p\text{-value} = 0$$

- The Durbin–Watson statistic is substantially below 2, indicating strong positive autocorrelation among residuals.
- Since the p-value is 0, we reject the null hypothesis of no autocorrelation.
- This suggests that the residuals are not independent, violating one of the key assumptions of the classical linear regression model.
- The presence of autocorrelation implies that time-dependent effects remain unexplained, and the model may need to incorporate time-series components or lagged variables.

Modeling Seasonality with Trigonometric Terms

Motivation

Consumption displayed clear **cyclical fluctuations** across months, suggesting a seasonal pattern. To capture this periodicity, sine and cosine terms were introduced as seasonal regressors:

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

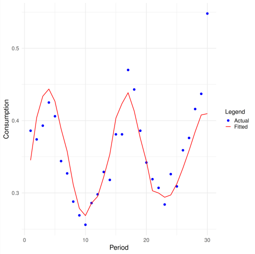
Extended Model

$$Y = \beta_0 + \beta_1 \text{ Temp} + \beta_2 \sin\left(\frac{2\pi \cdot \text{Period}}{12}\right) + \beta_3 \cos\left(\frac{2\pi \cdot \text{Period}}{12}\right) + \varepsilon$$

Key Results

- Adjusted $R^2 = 0.77$
- Residual SD $(\hat{\sigma}) = 0.12$
- AICc $= -32$

Including trigonometric terms significantly improved model fit, confirming seasonal structure.



Model Refinement: Transformations

Box-Cox Transformation for Response Variable

Motivation:

Initial models showed residual patterns suggesting a transformation of the response variable could improve fit.

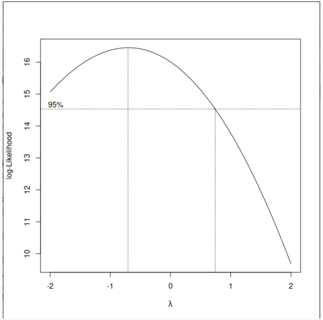
Box-Cox Analysis:

Applied Box-Cox procedure to find optimal power transformation:

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

Results:

- Optimal $\lambda = -0.7$ initially
- Final simplified model: $\lambda = -1$
- Transformed response: Consumption^{-1}



Box-Cox likelihood profile showing $\lambda = -0.7$ as optimal

Model Improvement:

- Adjusted R^2: $0.771 \rightarrow 0.816$
- Better residual behavior
- Normalized distribution

Model Selection and Optimization

Initial Model Complexity

Started with Full Model Including:

- **Polynomial Terms:** Temp^2, Period^2, Price^2
- **Interaction Terms:** Temp $\times$ Income, Temp $\times$ Price, etc.
- **Seasonal Components:** $\sin(2\pi \cdot \text{Period}/12)$, $\cos(2\pi \cdot \text{Period}/12)$
- **Categorical Variables:** Year (1951-1953), Season
- **Indicator Interactions:** high_income, high_temp, high_price

Total: 15+ predictors

Problem: Overfitting with only 30 observations

Forward Stepwise Selection

Table 1: Best Candidate Model Summary

Variable	Estimate	Std. Error	t value	$Pr(>\|t\|)$
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

Table 2: Generalized Variance Inflation Factors

Variable	GVIF	Df	$GVIF^{1/(2 \cdot Df)}$
$I(Temp^2)$	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

Interpretation of Stepwise Selection Results

Key Observations

- The stepwise procedure retained both **nonlinear (quadratic)** and **seasonal** terms:
  - Temp^2 — confirms a nonlinear temperature–consumption relationship.
  - sin and cos terms — capture strong periodic seasonality.
- **Income** and **Price** are significant, showing economic influence beyond weather factors.
- Year dummies (1952, 1953) remain weak — suggesting minor year-to-year shifts once trend and seasonality are modeled.
- Adjusted $R^2 = 0.918$ indicates an excellent in-sample fit, but several predictors exhibit high GVIF values.

Implications

- Possible **multicollinearity** among time-based and polynomial terms.
- Complex models risk **overfitting** with only 30 observations.
- Simplification is needed for robust inference.
- GVIF values indicate correlation among time-related predictor

Addressing Multicollinearity and Model Simplification

Background

Forward stepwise selection produced an overfitted model:

- High adjusted $R^2 = 0.918$ but inflated GVIFs (> 9) for several predictors.
- Signs of redundancy among trend, seasonal, and polynomial terms.

Simplified Model Retained:

$$Y = \beta_0 + \beta_1 \,\text{Temp} + \beta_2 \,\text{Temp}^2 + \beta_3 \,\sin\left(\frac{2\pi \cdot \text{Period}}{12}\right) + \varepsilon$$

Pruning Strategy

Predictors were systematically dropped based on likelihood-ratio tests, AICc improvement, and GVIF diagnostics:

- **Removed:** Year dummies (redundant with seasonal terms) $\rightarrow$ AICc improved to –135.25
- **Removed:** Linear *Period* (captured by sine/cosine) $\rightarrow$ AICc improved to –138.80
- **Removed:** *Income*, *Price*, and *Cosine term* — reduced collinearity, slightly lowered adjusted R^2 but enhanced stability.

Ad hoc Box–Cox Refit and Parsimonious Final Model

Motivation

After pruning, residuals still exhibited heteroscedasticity and autocorrelation. A Box-Cox analysis on the reduced specification suggested a power transformation with $\lambda \approx -1$.

Transformed Model

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

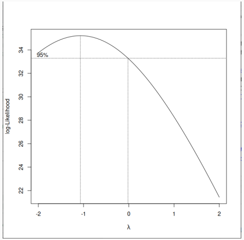
Interpretation

- Temp^2: strong curvature—consumption rises then flattens at high temperatures
- $\sin(2\pi\,\text{Period}/12)$: captures dominant seasonal oscillation
- Period^2: mild long-term trend effect

Final Fit: $F(3, 26) = 73.3$, $p < 10^{-12}$ **Multiple $\mathbf{R^2}$ = 0.894, Adjusted $\mathbf{R^2}$ = 0.882**

Key Diagnostics

- No severe multicollinearity (VIF $\approx$ 1.0 - 1.1 for all predictors)
- Adjusted $R^2 = 0.882$, Residual SD $= 0.1714$
- AICc (with Jacobian) $\approx$ -137 — competitive and parsimonious



Box-Cox log-likelihood profile for the reduced model

Final Model

Coefficient Estimates

Parameter	Estimate	p-value
Intercept	3.77	< 0.001
Temp^2	-3.02×10^{-5}	< 0.001
sin_term	-4.01×10^{-2}	< 0.001
Period^2	5.84×10^{-5}	0.002

Variance Inflation Factors (VIF)

Variable	VIF
$\text{I}(\text{Temp}^2)$	1.058
sin_term	1.063
$\text{I}(\text{Period}^2)$	1.012

Regression Results (Parsimonious Model)

Variable	Estimate	Std. Error	t value	$\Pr(>\lvert t\rvert)$
Intercept	3.661e+00	7.318e-02	50.023	$< 2 \times 10^{-16}$ ***
$\text{I}(\text{Temp}^2)$	-2.221e-04	2.039e-05	-10.893	3.47e-11 ***
sin_term	-2.886e-01	4.635e-02	-6.226	1.38e-06 ***
$\text{I}(\text{Period}^2)$	-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

Model Diagnostics

Diagnostic 1: Checking Functional Form

Purpose

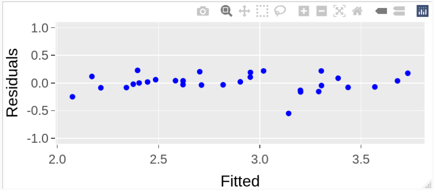
To assess whether residuals show systematic patterns indicating nonlinearity or omitted variable.

Interpretation

- Random scatter around zero - model is well specified.
- No significant Curvature suggesting missing nonlinear term or transformation.

Observation for this model:

Residuals are approximately centered, with no clear curvature, suggesting good functional form after Box-Cox refit.



Residuals vs. Fitted values plot for the final model

Diagnostic 2: Homoscedasticity (Constant Variance)

1. Breusch–Pagan (BP) Test

- Based on regressing squared residuals on model predictors:

$$e_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \cdots + \alpha_k X_{ki} + u_i$$

- Test statistic:

$$\text{BP} = nR^2_{\text{aux}} \sim \chi^2_k$$

- Decision rule: Reject H_0 : constant variance if $p < 0.05$.
- For this model: $\text{BP} = 1.5504,\ p = 0.6707 \Rightarrow$ fail to reject H_0.

2. Goldfeld–Quandt (GQ) Test

- Sort data by an explanatory variable (e.g., Temp), omit central observations.
- Compute ratio of residual variances between two groups:

$$GQ = \frac{s_1^2}{s_2^2} \sim F_{(n_1, n_2)}$$

- For this model: $GQ = 0.87,\ p = 0.587 \Rightarrow$ no evidence of heteroscedasticity.

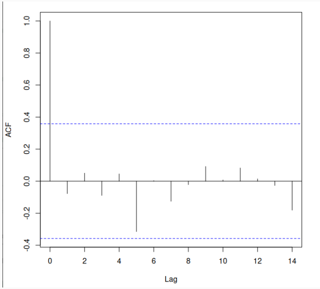
Diagnostic 3: Autocorrelation of Residuals

Durbin–Watson (DW) Test

- Tests for first-order serial correlation:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

- $DW \approx 2$ indicates no autocorrelation. $DW < 2 \rightarrow$ positive correlation; $DW > 2 \rightarrow$ negative correlation.
- For this model: $DW = 1.67$, $p = 0.134 \Rightarrow$ no evidence of serial correlation.



Diagnostic 4: Normality of Residuals

Purpose

To assess whether residuals follow a normal distribution. Normality is important for valid inference and confidence intervals in small samples.

1. Shapiro–Wilk Test

- Tests whether residuals deviate from normality:

$$W = \frac{\left(\sum_{i=1}^{n} a_i e_{(i)}\right)^2}{\sum_{i=1}^{n} (e_i - \bar{e})^2}$$

  where $e_{(i)}$ are ordered residuals and a_i are constants derived from expected normal order statistics.
- Null hypothesis: residuals are normally distributed.
- For this model: $W = 0.9752,\ p = 0.6908 \Rightarrow$ fail to reject H_0.

2. Normal Q–Q Plot

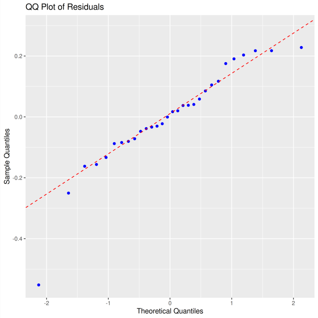
- Plots ordered standardized residuals against theoretical quantiles:

$$\text{If normal: points lie approximately on the } y = x \text{ line.}$$

- Visual inspection: residuals closely follow the reference line with minor tail deviations.

Conclusion:

Both the Shapiro–Wilk test and the Q–Q plot indicate that residuals are approximately normally distributed — no major normality violations detected.



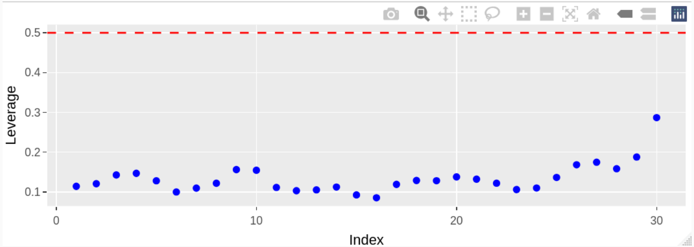
Diagnostic 5: Leverage

Definition

The leverage of observation i is the i-th diagonal element of the projection (hat) matrix:

$$h_{ii} = x_i'(X'X)^{-1}x_i$$

Observations with $h_{ii} > 2\bar{h}$ or $h_{ii} > 3\bar{h}$ are often flagged as high-leverage points.



Leverage values for fitted model

Conclusion:

The design matrix X is well-conditioned — no high-leverage observations detected.

Diagnostic 6: Influence (Cook's Distance)

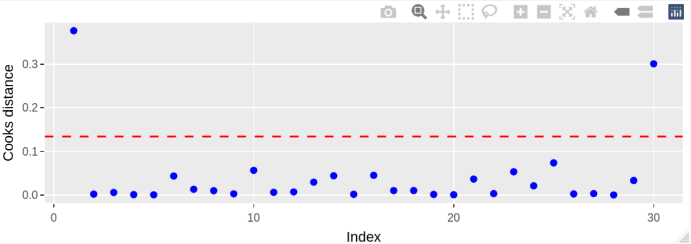
Purpose

To assess whether any individual observation unduly influences the fitted regression model. Cook's distance combines information about both leverage and residual size.

Definition

For observation i:

$$D_i = \frac{(e_i^2)}{p\,\hat{\sigma}^2} \cdot \frac{h_{ii}}{(1 - h_{ii})^2}$$



Cook's distance values for all observations

Interpretation

Observations with $D_i > 4/n$ are often considered influential.

Conclusion:

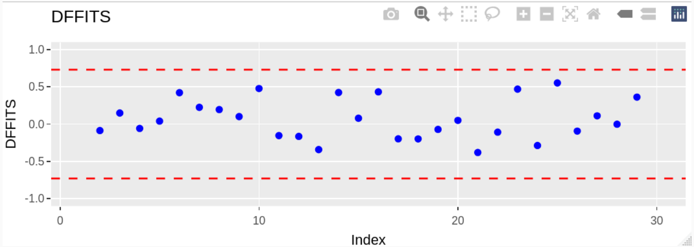
Observations 1 and 30 are flagged as influential.

Diagnostic 7: Influence (DFFITS)

Definition

DFFITS measures how much the fitted value for observation i changes when that observation is omitted from model estimation:

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$$



DFFITS values across observations

Rule of Thumb

$|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}} \Rightarrow$ potentially influential.

Conclusion:

All observations remain within acceptable limits. Their influence on fitted values is small and mostly consistent with Cook's distance results.

Robustness Check: Excluding Influential Points

Purpose

To test the stability of the estimated coefficients after excluding the two influential observations (#1 and #30).

Refitted Specification

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

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

Comparison of parameter estimates before and after outlier removal

Findings

- Coefficients are nearly identical to those from the full model.
- Signs, magnitudes, and statistical significance remain unchanged.
- The model's substantive conclusions are robust to exclusion of outliers.

Back-Transformation and Visual Fit

Purpose

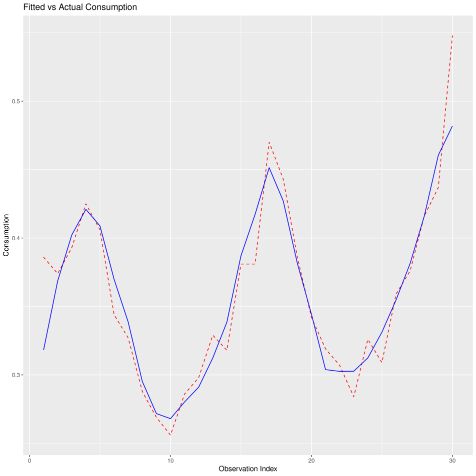
Since the model was estimated on a Box–Cox transformed response ($Y = \text{Consumption}^{-1}$), predictions were back-transformed to the original scale for interpretability.

Back-Transformation

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

Interpretation

- Fitted values accurately reproduce the observed consumption series.
- Actual (red dashed) and fitted (blue solid) lines closely align across all 30 periods.
- Indicates excellent in-sample predictive performance.



Observed (red dashed) vs. fitted (blue solid) consumption series

Closing Remarks

Summary of Findings

- The Box–Cox transformation ($\lambda \approx -1$) effectively stabilized variance and improved model diagnostics.
- Final model:

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

  achieved strong fit ($R^2_{\text{adj}} = 0.882$) with parsimonious structure.
- All diagnostic checks confirmed assumptions of linearity, homoscedasticity, normality, and independence.
- Influence diagnostics (Cook's D, DFFITS) revealed two mild outliers, but excluding them did not materially alter coefficients.
- Back-transformed fitted values tracked observed consumption extremely well.