To analyze the residuals and test the assumptions for a linear regression model, you should follow these steps:

### Step 1: Analyze Residuals

Residuals are the differences between actual and predicted values. Plotting and analyzing residuals helps verify if the assumptions of linear regression are met.

1. \*\*Residual Plot\*\*:

- Plot residuals against the predicted values to check for patterns. Ideally, residuals should be randomly scattered around zero, indicating that the model's assumptions are satisfied.

```python

residuals\_test = y\_test - y\_pred

residuals\_val = y\_val - y\_pred\_val

# Plotting residuals for the test set

plt.figure(figsize=(12, 6))

plt.scatter(y\_pred, residuals\_test, alpha=0.6)

plt.axhline(0, color='red', linestyle='--')

plt.xlabel('Predicted Values')

plt.ylabel('Residuals')

plt.title('Residuals vs Predicted Values (Test Set)')

plt.show()

```

2. \*\*Histogram of Residuals\*\*:

- Check if residuals are normally distributed by plotting a histogram. The distribution should resemble a normal curve.

```python

plt.hist(residuals\_test, bins=30, alpha=0.7)

plt.xlabel('Residuals')

plt.ylabel('Frequency')

plt.title('Histogram of Residuals (Test Set)')

plt.show()

```

### Step 2: Test Assumptions

To ensure that the basic assumptions of linear regression are met, you need to evaluate linearity, homoscedasticity (variance uniformity), and independence.

1. \*\*Linearity Assumption\*\*:

- Check if the relationship between the independent and dependent variables is linear. If the residuals plot shows a clear pattern (e.g., curvature), this indicates a potential violation of the linearity assumption.

- You can also plot the actual vs. predicted values to confirm linearity:

```python

plt.scatter(y\_test, y\_pred, alpha=0.6)

plt.plot([y\_test.min(), y\_test.max()], [y\_test.min(), y\_test.max()], 'r--')

plt.xlabel('Actual Values')

plt.ylabel('Predicted Values')

plt.title('Actual vs Predicted Values (Test Set)')

plt.show()

```

2. \*\*Homoscedasticity (Variance Uniformity)\*\*:

- Check if the variance of the residuals is constant across all levels of predicted values. This can be assessed using the residual plot:

- If residuals spread equally around zero across the plot, the variance is uniform (homoscedasticity). If residuals form a funnel shape (wider as values increase), heteroscedasticity is present.

- Use the residuals plot from Step 1 for this check.

3. \*\*Independence of Residuals\*\*:

- Check if residuals are independent of each other, especially for time-series data. This can be evaluated using the Durbin-Watson test.

- The Durbin-Watson statistic should be around 2 for independent residuals (values close to 0 or 4 indicate autocorrelation).

```python

from statsmodels.stats.stattools import durbin\_watson

dw\_test = durbin\_watson(residuals\_test)

print('Durbin-Watson statistic:', dw\_test)

```

### Step 3: Summary of Assumptions

- \*\*Linearity\*\*: Check the scatter plot of residuals vs. predicted values for patterns.

- \*\*Homoscedasticity\*\*: Confirm no clear patterns or funnels in the residuals plot.

- \*\*Independence\*\*: Ensure the Durbin-Watson statistic is close to 2.

### Interpreting Results:

- If the assumptions are violated (e.g., non-random residuals, heteroscedasticity), consider transforming variables, adding features, or using a different model (e.g., polynomial regression or robust regression).

- If residuals are normally distributed and spread randomly, the linear model is likely appropriate for the data.

IV. METHODS

1. Linear Regression(LR)

Linear regression [1] is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.

**Types of Linear Regression** [1]

There are two main types of linear regression:

* Simple Linear Regression

This is the simplest form of linear regression, because it has only one independent variable and one dependent variable. The formula for simple linear regression is:

y = β0 +β1X

where:

Y is the dependent variable

X is the independent variable

β0 is the intercept

β1 is the slope

* Multiple Linear Regression

It used to model relationships between a dependent variable and two or more independent variables. It extends simple linear regression by considering multiple predicted factors to estimate the outcome. Its formula is:

y = β0 + β1X + β2X + … + βnXn

Where:

Y is the dependent variable

X1, X2, …, Xn are the independent variables

β0 is the intercept

β1, β2, …, βn are the slopes

However, we simply introduce Simple Linear Regression within the scopes of this study.

V. Test assumptions:

Example:

A graph showing a line graph

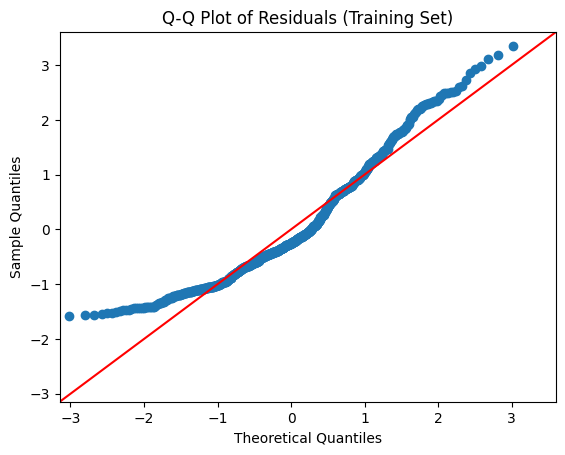
Description automatically generated with medium confidence

**1. Kiểm tra Tính Tuyến Tính (Linearity)**

* **Giả định**: Nếu mô hình hồi quy tuyến tính là phù hợp, các phần dư nên phân bố ngẫu nhiên xung quanh đường trục 0 và không có mô hình hoặc xu hướng rõ ràng nào trong các residuals.
* **Quan sát**: Trong biểu đồ này, có sự biến động lớn của residuals khi giá trị dự đoán thay đổi. Đặc biệt, có một xu hướng rõ ràng, khi residuals tăng và giảm thành dạng sóng (wave) quanh trục 0. Điều này cho thấy có thể tồn tại mối quan hệ phi tuyến tính trong dữ liệu mà mô hình tuyến tính không nắm bắt được.
* **Kết luận**: Giả định tuyến tính có thể không được thỏa mãn ở đây. Mô hình tuyến tính có thể không phù hợp để nắm bắt hết các xu hướng trong dữ liệu.

**2. Kiểm tra Tính Đồng Nhất của Phương Sai (Variance Uniformity / Homoscedasticity)**

* **Giả định**: Phương sai của residuals phải đồng đều trên toàn bộ dải giá trị dự đoán. Điều này có nghĩa là residuals không nên có xu hướng phân tán nhiều hơn hoặc ít hơn ở bất kỳ vùng giá trị nào của giá trị dự đoán.
* **Quan sát**: Biểu đồ residual cho thấy phương sai của các phần dư không đồng đều. Ở các giá trị dự đoán thấp (khoảng 1400–1500), các residuals có xu hướng phân tán ít hơn so với các giá trị dự đoán trung bình và cao hơn. Đặc biệt, ở mức dự đoán cao hơn (gần 1700–1800), các residuals có xu hướng phân tán rộng hơn nhiều so với những vùng khác, cho thấy phương sai không đồng nhất.
* **Kết luận**: Giả định đồng nhất của phương sai (homoscedasticity) không được thỏa mãn. Phương sai của residuals thay đổi theo giá trị dự đoán, điều này có thể ảnh hưởng đến độ tin cậy của mô hình hồi quy tuyến tính.
  1. **Normality**



* Trong biểu đồ này, các điểm hầu như nằm trên đường chéo, đặc biệt là ở phần trung tâm của biểu đồ (từ khoảng -1 đến +2). Điều này cho thấy rằng phần dư gần như tuân theo phân phối chuẩn trong vùng này.
* Tuy nhiên, có một số điểm ở hai đầu của biểu đồ (các phần tử cực đại và cực tiểu) bị lệch khỏi đường chéo. Đây là dấu hiệu của các phần dư nằm ở hai đầu có độ lệch so với phân phối chuẩn, có thể do tồn tại **outliers** hoặc dữ liệu có đuôi dài hơn phân phối chuẩn.
  1. **Independence**

**Durbin-Watson statistic (Training Set): 0.025981063715549414**

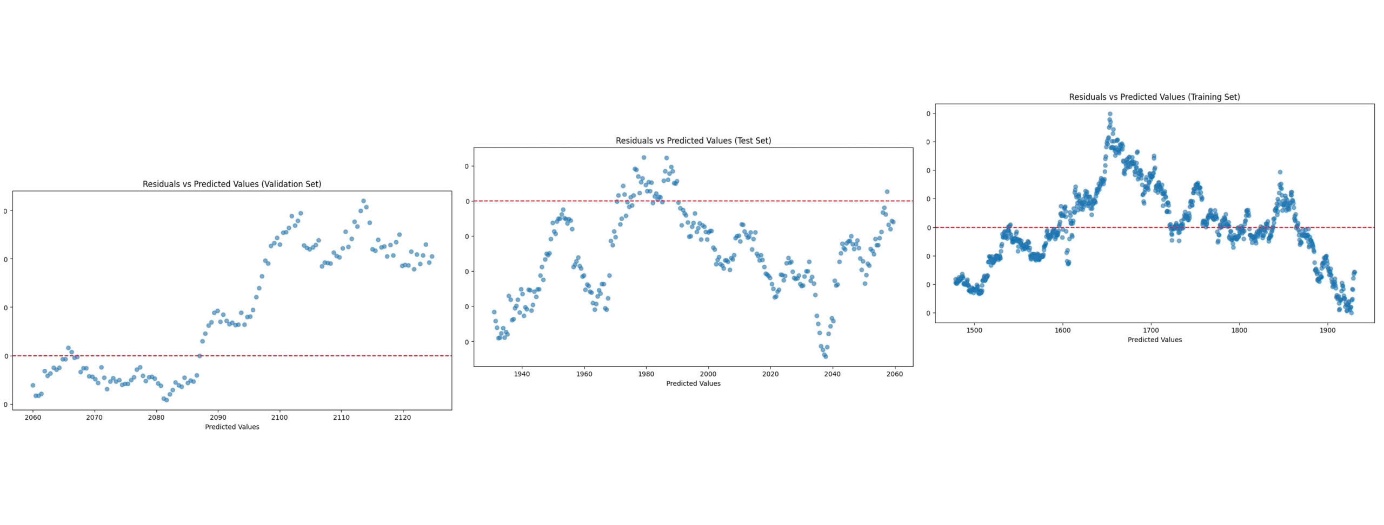
Thống kê Durbin-Watson (DW) là một chỉ số dùng để kiểm tra tính độc lập của các residuals (phần dư) trong mô hình hồi quy, đặc biệt là để phát hiện sự tự tương quan (autocorrelation) giữa các residuals. Giá trị của thống kê Durbin-Watson dao động từ 0 đến 4, với các ý nghĩa như sau:

* **Khoảng gần 2**: Phần dư độc lập, không có tự tương quan. Đây là giá trị lý tưởng cho một mô hình hồi quy tuyến tính.
* **Gần 0**: Cho thấy có sự tương quan dương mạnh giữa các residuals (autocorrelation dương). Điều này nghĩa là các residuals có xu hướng phụ thuộc vào nhau, tức là nếu một residual dương, có khả năng cao rằng residual tiếp theo cũng dương.
* **Gần 4**: Cho thấy có sự tương quan âm mạnh giữa các residuals (autocorrelation âm).

**Phân tích kết quả của bạn**

Với **Durbin-Watson statistic là 0.02598**, đây là một giá trị rất gần 0, cho thấy có **sự tự tương quan dương mạnh** giữa các residuals. Điều này có nghĩa là các residuals không độc lập, mà có xu hướng phụ thuộc vào nhau. Cụ thể hơn:

1. **Sự phụ thuộc giữa các residuals**: Các residuals có thể có một xu hướng dương liên tiếp, nghĩa là nếu một residual dương thì các residual tiếp theo cũng có xu hướng dương. Tương tự, nếu một residual âm, các residual tiếp theo cũng có xu hướng âm.
2. **Hậu quả đối với mô hình**: Khi các residuals không độc lập, các giả định về độc lập trong hồi quy bị vi phạm. Điều này có thể làm giảm hiệu quả của mô hình và khiến các kiểm định về ý nghĩa thống kê trở nên không đáng tin cậy. Sự tự tương quan thường xuất hiện khi dữ liệu có xu hướng thời gian hoặc chuỗi (time series) mà mô hình hồi quy đơn giản không thể nắm bắt hết được.

****

**Check Assumptions:**

* 1. **Linearity:**

**Assessment Using Residual Plots:**

* If the residuals are randomly scattered around the horizontal line (zero line), this indicates that the linearity assumption is likely met.
* If there are patterns in the residuals, such as systematic curves or trends, it suggests a violation of linearity.

**Analysis of the Residual Plots:**

In the three residual plots (validation, test, and training sets), there appear to be patterns where residuals systematically deviate from the zero line, particularly in regions of higher and lower predicted values. This suggests that the relationship between the predictors and the response variable may not be perfectly linear, indicating a potential violation of the linearity assumption.

* 1. **Variance Uniformity**

**Variance of Residuals:**

In each of the residual plots (validation, test, and training sets), the spread of residuals does not seem uniform. For instance:

In the validation set, the variance appears to increase as the predicted values increase.

In the test set, there are sections where the spread of residuals widens and narrows inconsistently.

In the training set, the residuals also exhibit non-uniform variance, with higher spread in certain regions of predicted values.

**Possible Trends:**

Patterns of increasing or decreasing variance in residuals across the predicted values indicate heteroscedasticity, suggesting that the assumption of variance uniformity is violated.

**A graph of a graph

Description automatically generated with medium confidence**

* 1. **Normality**

**Analysis of Q-Q plots for each dataset:**

Training set: The residuals closely follow the theoretical line, indicating that the residuals approximately follow a normal distribution. However, there are slight deviations at the extreme ends (lower and upper quantiles), suggesting potential outliers.

Test set: Similar to the training set, the residuals mostly align with the theoretical line, with minor deviations at the tails. This indicates the residuals are approximately normally distributed.

Validation set: The Q-Q plot shows significant deviations from the theoretical line, particularly at the lower and upper ends. This indicates that the residuals in this dataset deviate from a normal distribution, especially at the extremes.

**Conclusion on normality assumption:**

Training and test sets: The residuals generally satisfy the normality assumption, with only minor deviations at the tails.

Validation set: The residuals violate the normality assumption, especially in the extreme quantiles, suggesting that the residuals in this set are not normally distributed.

* 1. **Independence**

A black screen with white text

Description automatically generated

**Understanding Durbin-Watson Statistic:**

* The **Durbin-Watson statistic** ranges from 0 to 4:
  + A value of **2** indicates no autocorrelation (perfect independence).
  + A value **less than 2** suggests **positive autocorrelation**.
  + A value **greater than 2** suggests **negative autocorrelation**.
  + Values close to **0** indicate **strong positive autocorrelation**, which violates the independence assumption.

**Analysis of the Values:**

1. **Training Set**:
   * Durbin-Watson statistic: 0.01750.0175
   * This value is extremely close to 00, indicating **strong positive autocorrelation** in the residuals of the training set. Independence is clearly violated.
2. **Test Set**:
   * Durbin-Watson statistic: 0.02770.0277
   * Similar to the training set, this value is still very close to 00, suggesting **strong positive autocorrelation** and a violation of independence in the test set.
3. **Validation Set**:
   * Durbin-Watson statistic: 0.01620.0162
   * This is also very close to 00, indicating **strong positive autocorrelation** and a clear violation of independence in the validation set.

**Overall Assessment:**

The model assumptions are significantly violated in multiple areas, particularly in linearity, variance uniformity, and independence. While the normality assumption holds reasonably well for the training and test sets, the validation set shows deviations. These issues highlight the need for adjustments in the model, including potential non-linear transformations, addressing autocorrelation, and revising the error structure to improve overall model validity.

**Recommended Next Steps:**

Reassess the model's appropriateness, particularly for addressing non-linearity and heteroscedasticity.

Consider alternative models that can better handle autocorrelation, such as time-series models for temporal data or generalized least squares (GLS).

Apply transformations (e.g., logarithmic or polynomial) to stabilize variance and better capture the relationship between predictors and the response variable.

Validate any updated models against new datasets to ensure that assumptions are satisfied.