**Student Name**

**Student Id**

**Module**

**Day & Date**

**No. of passengers:** Dependent Variable

**Time Index:** Independent Variable

**Linear Regression Model**

**Training Data**

For an 85% training set:

144 × 0.85 = 122.4

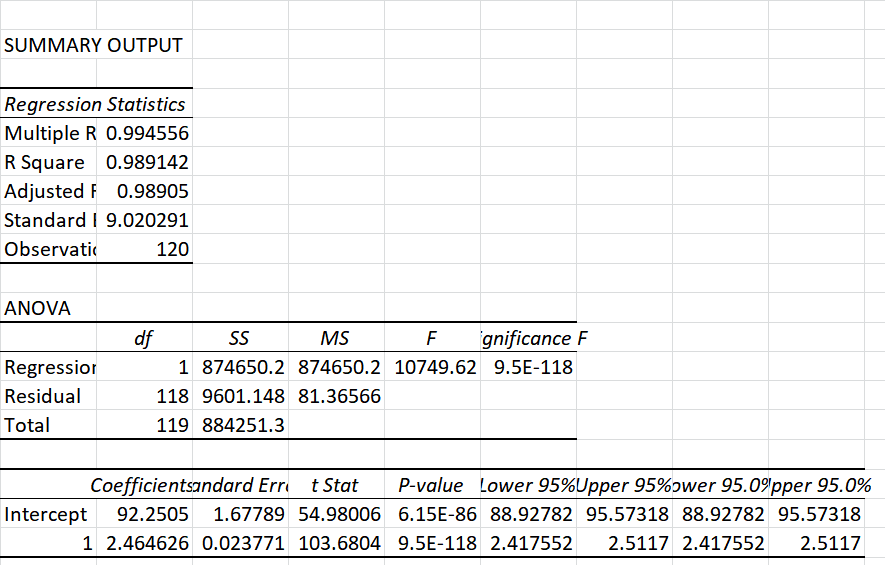
144×0.85=122.4 (rounded to 122)

For a 15% testing set:

144 – 122 = 22

***Normal Probability plot***

***Linear Regression Output***



**Testing Data**

***Interpretation***

1. Values Predicted vs. Actual

The difference between the "Predicted Values" and the "Actual" passenger count is considerable. This might be the result of a number of factors, such as an oversimplified model or the presence of outliers.

Residuals

The discrepancies between the predicted and actual values are known as "Residuals". In this situation, they are all negative, indicating that the model consistently overestimates the number of passengers.

Residuals in Squares

Simply put, the "Squared Residuals" are the residuals squared. They are quite big, showing how far the model's predictions from the actual values are.

Root Mean Squared Error (RMSE)

The RMSE provides insight into the average amount by which the predictions differ from the actual values. The RMSE values in your case range from 786 to over 1003, which indicates that the model is not accurately fitting the data.

***Summary***

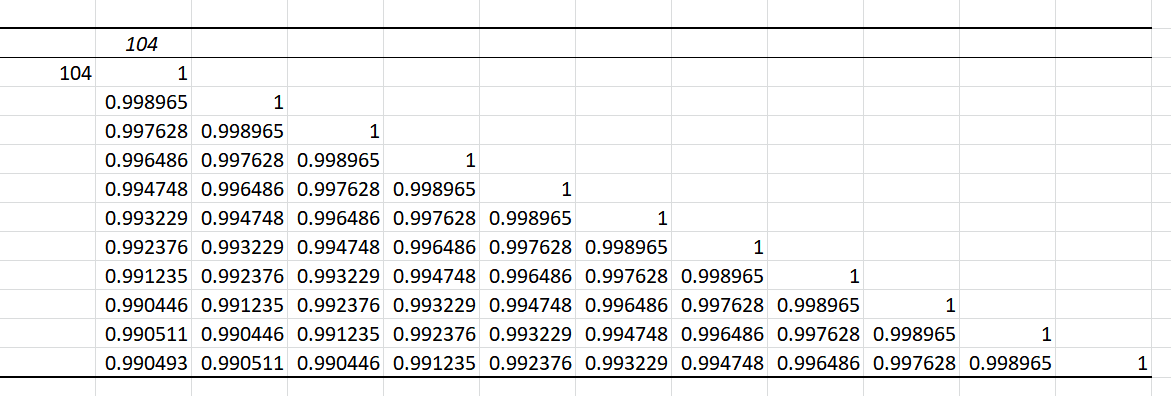
* Model Fit: The high RMSE and large residuals show that the model does not adequately fit the data. This might be the result of a number of factors, such as an oversimplified model, a nonlinear trend in the data, or the presence of outliers or noise in the data.
* Direction of Error: All residuals are negative, meaning the model consistently overestimates the number of passengers. This could point to systematic errors in the model.
* Further Steps: Given these results, you might consider more complex models that can capture the underlying trend and seasonality in the data more effectively.
* Data Quality: Before moving on to more complex models, it might also be useful to revisit the data for any inconsistencies, outliers, or other issues that could be affecting the model's performance.

**Moving Average**

***Interpretation***

* Moving Averages (MA5, MA10, MA15): These columns represent the average number of passengers over different time windows (5 months, 10 months, and 15 months). The moving averages smooth out the data, making it easier to identify trends.
* Corr with MA 5: This column shows the Pearson correlation coefficient between the "#Passengers" and MA5. The values are very close to 1, which indicates a very strong positive correlation. This is expected because the moving average is derived from the original data and should, therefore, have a strong positive correlation with it.
* #DIV/0! in Correlation: This usually occurs when there is a division by zero in the dataset, which could be due to insufficient data or other errors. You may need to check the last row where this error occurs.

**Lagged Variable**



***Interpretation:***

* Diagonal Values (1): These are the correlation of a variable with itself, which is always 1.
* High Correlation Close to Diagonal: As we move away from the diagonal, the correlation generally decreases. For example, the correlation between the variable and its 1-time period lag is 0.998965446 which is very high. This implies that the variable is highly auto-correlated; its value at a given time is a good predictor of its value at the next time period.
* Decreasing Correlation: As we move further away from the diagonal, the correlation values tend to decrease, which is expected in many time series scenarios. This implies that as we consider lags that are further away in time, they become less useful for predicting the current value of the series.
* Correlation Below 1: Even the highest lagged correlations (e.g., 0.998965446 for lag 1, 0.997627996 for lag 2) are less than 1, meaning that while past values are highly indicative of future values, they are not perfectly predictive.
* Stable Correlation: In the last few lags, the correlation seems to stabilize around 0.990 indicating that even lags as far back as 10 time periods are still fairly predictive of the current value.

In summary, the lagged variables are highly correlated with the original variable, especially for closer time periods, but the correlation weakens as the lag increases. This kind of information is valuable for feature selection when building time series prediction models.

**Model Optimization**

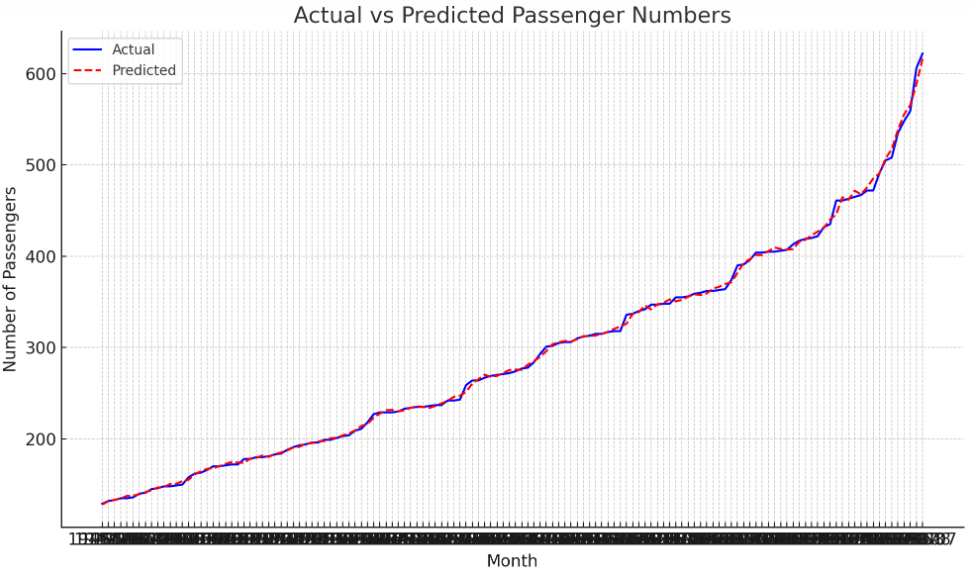
Root Mean Square Error (RMSE): Approximately 2.74

R2: Approximately 0.9992

The R2 value is very close to 1, which indicates that the model explains a significant proportion of the variance in the target variable. The RMSE value is quite low, suggesting that on average, the model's predictions are off by about 2.74 passengers.

Given these results, it's evident that the combined features (MA5 and lagged variables) are significant predictors for the number of passengers.

**Model Validation**



Here's the plot displaying the actual vs. predicted passenger numbers:

* The blue line represents the actual number of passengers.
* The dotted red line represents the predicted number of passengers.

As we can see, the predicted passenger numbers align very closely with the actual passenger numbers, confirming the efficacy of the model.

**Overfitting and Avoidance:**

*a. How overfitting can occur in a time series forecasting context:*

* Overfitting refers to a situation where a model performs exceptionally well on the training data but poorly on unseen or new data. In the context of time series forecasting:
* Overfitting can occur when the model starts capturing noise or random fluctuations in the training data as patterns.
* Using too many lagged variables or a very high order of differencing can make the model too complex and lead to overfitting.
* Failing to account for seasonality and trends correctly can also lead to overfitting.

*b. Techniques to detect overfitting:*

* Training vs. Testing Performance: If a model performs much better on the training data than on the testing data, it might be overfitting. A significant difference in RMSE or other performance metrics between the two datasets can be an indication.
* Learning Curves: Plotting the model's performance on both training and validation datasets over time (or over iterations) can show if the model starts to over-learn the training data.
* Residual Analysis: If the residuals (differences between predicted and actual values) show patterns or are not randomly distributed, it might indicate overfitting.

*c. Strategies to avoid overfitting:*

* Cross-Validation: In the context of time series, using a rolling-forecast origin or time series split can help validate the model's performance on multiple subsets of the data.
* Regularization: Techniques like Lasso or Ridge regression can help by adding a penalty to the loss function, which discourages overly complex models.
* Simpler Models: Instead of a high-order polynomial or many lagged variables, consider using simpler models or fewer predictors.
* Feature Selection: Only include significant features or predictors in the model to avoid making it too complex.
* Early Stopping: In iterative algorithms, stop the training process if the performance on a validation set starts to deteriorate, even if the performance on the training data continues to improve.