linearity between independent variables and log-odds

The assumption of **linearity between independent variables and log-odds** is essential in logistic regression, meaning that each predictor's effect should be linearly related to the log-odds (logarithmic transformation of the odds) of the outcome variable. This does not require the relationship between the predictor and the raw outcome to be linear but only with the log-odds of the outcome. Here's a deeper look at this assumption and the testing methods:

1. Understanding the Assumption

- In logistic regression, we model the probability of an event (e.g., customer segment membership) by applying a logistic function to a linear combination of predictors.
- The linearity assumption implies that as each predictor changes, the log-odds of the outcome should change linearly.
- Violations of this assumption can lead to poor model fit and biased coefficients, affecting the accuracy of predictions.

2. Box-Tidwell Test for Linearity

- The **Box-Tidwell test** is commonly used to check this linearity assumption specifically for continuous predictors. It works by creating an interaction term between each predictor and its log-transformation and then testing if the interaction is significant. If the interaction term is significant, it indicates non-linearity.

3. Alternative: Plotting Predictor vs. Log-Odds or Predicted Probabilities

- **Visualization**: If you only have a few continuous predictors, plotting each predictor against the predicted log-odds or probabilities can be an intuitive way to assess linearity. A linear relationship in these plots suggests the assumption holds; if curvilinear, transformations like polynomial terms or log-transformations on the predictors can be considered.

Script for Testing Linearity with Box-Tidwell

Let's set up a Python script that performs the Box-Tidwell test for each continuous predictor in a dataset, checks for significance, and provides recommendations for transformation if necessary.

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
```

```
from statsmodels.formula.api import logit
from scipy.stats import chi2

# Assume 'data' is your DataFrame and 'target' is the binary outcome variable (e.g., 'Segment')
# Replace 'continuous_predictors' with the list of continuous predictors in your dataset

def box_tidwell_test(data, target, continuous_predictors):
    # Adding a small constant to avoid issues with log(0)
    data = data.copy()
    data[continuous_predictors] = data[continuous_predictors].apply(lambda x: x + 1e-9)

    results = {}

    for predictor in continuous_predictors:
        # Create a transformed interaction term for the Box-Tidwell test
        data['interaction'] = data[predictor] * np.log(data[predictor])

        # Define the model formula with interaction term
        formula = f"{target} ~ {predictor} + interaction"

        # Fit the logistic regression model
        model = logit(formula, data=data).fit(disp=False)

        # Get the p-value for the interaction term
        p_value = model.pvalues['interaction']

        # Interpret the result
        if p_value < 0.05:
            results[predictor] = "Non-linearity detected. Consider transforming this predictor."
        else:
            results[predictor] = "Linearity holds."

    return results

# Example usage:
continuous_predictors = ['your_continuous_predictor1', 'your_continuous_predictor2']  # Replace with actual names
result = box_tidwell_test(data, 'Target_Segment', continuous_predictors)
print(result)
```

Explanation of the Script

- **Interaction Term**: For each continuous predictor, we create an interaction term by multiplying it with its log-transformation.
- **Logistic Regression with Interaction**: We fit a logistic regression model that includes both the predictor and its interaction term.

- **Interpretation**: If the p-value for the interaction term is significant (typically $p < 0.05$), it suggests a non-linear relationship, and we may need to transform the predictor (e.g., using polynomials, log-transformations) to achieve linearity with the log-odds.

Next Steps if Non-Linearity is Found

1. **Transform Predictors**: Apply polynomial terms (e.g., square or cubic terms) or logarithmic transformations to the affected predictor(s).
2. **Re-Test**: After transforming, rerun the Box-Tidwell test to confirm that the transformation achieves linearity.

This approach provides a systematic way to ensure the model's assumptions are met before proceeding with logistic regression.