Data Transformation with Box-Cox for Regression

Heteroscedasticity

Heteroscedasticity refers to a situation in regression analysis where the **variance of errors (residuals) is not constant** across all levels of the independent variable(s). In simple terms, the spread of the residuals **increases or decreases** as the value of the predictor variable changes.

Key Points:

✅ **Homoscedasticity** (Ideal Case): The variance of residuals is **constant** across all levels of the predictor variable.

🚫 **Heteroscedasticity** (Problematic Case): The variance of residuals **changes** across different values of the predictor variable, forming patterns like a funnel or cone shape in a residual plot.

Why is Heteroscedasticity a Problem?

1. **Violates the Assumption of Linear Regression**
    - One key assumption in regression is that residuals (errors) should have **constant variance**.
    - Heteroscedasticity **violates this assumption**, leading to unreliable statistical inferences.
2. **Affects Confidence Intervals and Hypothesis Testing**
    - Standard errors of coefficients may be **biased**, leading to incorrect **p-values** and **confidence intervals**.
    - This increases the chance of making **wrong conclusions** from your model.
3. **Reduces Model Efficiency**
    - OLS (Ordinary Least Squares) estimates remain **unbiased**, but they are no longer the **best (minimum variance)** estimators.

- Predictions become **less reliable**, especially for extreme values.

How to Detect Heteroscedasticity?

1. **Residual Plot (Scatter Plot of Residuals vs. Predicted Values)**
   - If you see a **funnel shape** (wider spread at larger values), heteroscedasticity is present.
   - Homoscedastic residuals should look **randomly scattered**.
2. **Breusch-Pagan Test & White Test**
   - These are statistical tests to check for heteroscedasticity.
   - A **low p-value** (< 0.05) suggests **heteroscedasticity is present**.
3. **Goldfeld-Quandt Test**
   - Compares the variance of residuals in different sub-samples of data.

How to Fix Heteroscedasticity?

✔️ **Log Transformation**: Taking the log of the dependent variable can stabilize variance.

✔️ **Weighted Least Squares (WLS)**: Assigns different weights to observations based on variance.

✔️ **Robust Standard Errors**: Adjusts standard errors to account for heteroscedasticity.

Example

Consider a dataset where we predict **house prices** based on **size (sq ft)**.

- If heteroscedasticity is present, the variability in house prices will **increase** as the house size increases.
- This means larger houses have **more unpredictable** prices, leading to increasing residual variance.

Box-Cox Transformation

The **Box-Cox transformation** is a **power transformation** used to stabilize variance and make data more normally distributed. It is especially useful when dealing with **heteroscedasticity** or **non-normally distributed** data.

Mathematical Formula

The Box-Cox transformation is defined as:

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

where:

- Y is the original data.
- λ is the transformation parameter.
- $\ln(Y)$ (log transformation) is used when λ=0.

Why Use Box-Cox Transformation?

✅ **Fixes Skewness**: Helps make data more **normally distributed**, improving statistical tests.

✅ **Reduces Heteroscedasticity**: Stabilizes variance in regression models.

✅ **Improves Linearity**: Makes relationships more **linear**, benefiting linear regression.

✅ **Enhances Model Accuracy**: Helps models meet assumptions for better predictions.

Choosing the Best λ Value

- The optimal λ is usually found **automatically** by maximizing the **log-likelihood** function.
- Common values of λ:
  - λ = 1 → No transformation (original data).
  - λ = 0 → Log transformation ln(Y).
  - λ = 0.5 → Square root transformation.
  - λ = -1 → Reciprocal transformation $\frac{1}{Y}$.

Example Use Case

Scenario: Suppose we are predicting house prices, but the data is highly skewed. Using a Box-Cox transformation can make it **normally distributed**, leading to better regression results.
