Diagnostic Plots

Ask: Can you define heteroskedasticity transformation of scale location as same as non stationarity and would need to include a difference to introduce stationarity same as for linear models...?

Heteroskedasticity transformation in the scale-location context is related to addressing non-constant variance in residuals, but it is not necessarily the same as non-stationarity. However, both heteroskedasticity and non-stationarity deal with variability issues in data, and sometimes similar transformations can help in handling both.

Key Differences:

1. Heteroskedasticity refers to the presence of non-constant variance in residuals of a model. This is common in regression models, where variance changes over the range of predictors. A common fix includes:

Log Transformation or Box-Cox Transformation to stabilize variance.

Weighted Least Squares (WLS) regression to account for non-constant variance.

Generalized Least Squares (GLS) to model variance explicitly.

2. Non-Stationarity in time series occurs when statistical properties like mean and variance change over time. This is different from heteroskedasticity in the sense that:

Non-stationary data often requires differencing to achieve stationarity.

A time series can be non-stationary due to trends or seasonality rather than just variance changes.

The Augmented Dickey-Fuller (ADF) test or KPSS test is commonly used to check for stationarity.

When They Overlap:

In ARCH/GARCH models (used in econometrics), heteroskedasticity in residuals is explicitly modeled, which can sometimes relate to stationarity concerns.

Some financial time series exhibit both non-stationarity (trend) and heteroskedasticity (volatility clustering), requiring both differencing and variance stabilizing transformations.

To Answer Your Question:

Heteroskedasticity transformation of scale-location is not exactly the same as non-stationarity.

In time series models, introducing stationarity often requires differencing (removing trends).

In regression models, heteroskedasticity often requires variance stabilizing transformations (e.g., log, Box-Cox, WLS).

If a linear model exhibits non-stationarity, differencing can help in stationarity, whereas heteroskedasticity might still require transformation.

Diagnostic Plots

To diagnose linear models, we check for linearity, normality of residuals, constant variance(homoscedasticity), and lack of autocorrelation by examining residual plots, histograms and statistical tests.

1. Checking Assumptions:

Linearity:

Plot the dependent variable against each independent variable to see if a straight line or a cloud of points roughly symmetrical around a straight line exists. [https://nirpyresearch.com/diagnostic-plots-for-linear-regression/#:~:text=The%20simplest%20diagnostic%20operation%20is,as%20PLS)%20of%20our%20data.

https://cran.r-project.org/web/packages/AFR/vignettes/Diagnostic-tests.html#:~:text=Diagnostic%2Dtests.R-,Goldfeld%2DQuandt%20Test,the%20errors%20are%20not%20constant.

https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/how-to-conduct-linear-regression/#:~:text=Linear%20Regression%20Analysis%20consists%20of,and%20usefulness%20of%20the%20model.

https://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod2/6/index.html#:~:text=This%20can%20be%20checked%20by,scatterplot%20with%20no%20discernible%20pattern!

https://www.linkedin.com/advice/0/what-process-checking-linearity-linear-regression-skills-statistics-wiukf#:~:text=To%20check%20linearity%2C%20you%20can,symmetrical%20around%20a%20straight%20line.

]

Normality of Residuals:
Examine a histogram or a Q-Q plot of the residuals to see if they follow a normal distribution. [https://nirpyresearch.com/diagnostic-plots-for-linear-regression/#:~:text=The%20simplest%20diagnostic%20operation%20is,as%20PLS)%20of%20our%20data.

https://math.libretexts.org/Workbench/Numerical_Methods_with_Applications_(Kaw)/6%3A_Regression/6.05%3A_Adequacy_of_Linear_Regression_Models

]

Homoscedasticity (Constant Variance):
Plot the residuals against the predicted values or the independent variables. The distribution of residuals should not vary appreciably between different parts of the x-axis scale. A scatterplot with no discernible pattern is desired.

[https://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod2/6/index.html#:~:text=This%20can%20be%20checked%20by,scatterplot%20with%20no%20discernible%20pattern!

]

Autocorrelation:
Check for any pattern in the residuals over time (if your data is time-series).
https://nirpyresearch.com/diagnostic-plots-for-linear-regression/#:~:text=The%20simplest%20diagnostic%20operation%20is,as%20PLS)%20of%20our%20data.

https://cran.r-project.org/web/packages/AFR/vignettes/Diagnostic-tests.html#:~:text=Diagnostic%2Dtests.R-,Goldfeld%2DQuandt%20Test,the%20errors%20are%20not%20constant.

Outliers:
Identify and investigate any data points that deviate significantly from the general trend.

https://math.libretexts.org/Workbench/Numerical_Methods_with_Applications_(Kaw)/6%3A_Regression/6.05%3A_Adequacy_of_Linear_Regression_Models

In linear modeling, diagnostic plots help assess whether the
2. Diagnostic Tools:

Residual Plots:
Plotting the residuals (the difference between the observed and predicted values) against the predicted values or the independent variables is a crucial step.
Purpose: To visually assess if the model assumptions are met.
Interpretation: Look for patterns or trends in the residual plot, which might indicate violations of the assumptions.

Histograms and Q-Q Plots:
These are used to visually assess the distribution of the residuals and check for normality.
Statistical Tests:
Breusch-Pagan Test: Tests for heteroscedasticity (non-constant variance).
Durbin-Watson Test: Tests for autocorrelation in the residuals.
Shapiro-Wilk Test: Tests for normality of residuals.

Coefficient of Determination (R-squared):
Measures the proportion of variance in the dependent variable that is explained by the independent variables.

Standard Error of the Estimate:
Measures the average distance between the observed values and the predicted values.
Cook's Distance:
Measures the influence of individual observations on the model's predictions.

3. Addressing Issues:
Non-Linearity: Consider transforming the variables or using a non-linear model.
Non-Normality: Consider transforming the data or using a non-parametric model.
Heteroscedasticity: Consider using weighted least squares or a different model.
Autocorrelation: Consider using a time-series model.
Outliers: Investigate the outliers and consider whether they should be removed or transformed.model assumptions hold.

The four common diagnostic plots are:

- Residuals vs Fitted
- Normal Q-Q
- Scale-Location
- Residuals vs Leverage: provide insights into issues like heteroskedasticity, non-normality, and influential points. Here’s what each means:

1. **Residuals vs Fitted**

Purpose: Checks for non-linearity and heteroskedasticity.

Residuals (errors) should be randomly scattered around zero.

A clear pattern (e.g., a curve) suggests the model doesn’t capture all relationships (non-linearity).

A fan-shaped pattern suggests heteroskedasticity (non-constant variance).

Ideal case: Residuals are randomly distributed without patterns.

2. **Normal Q-Q (Quantile-Quantile Plot)**

Purpose: Checks if residuals follow a normal distribution.

Plots theoretical quantiles vs sample quantiles.

If residuals are normally distributed, they align along the diagonal line.

Deviations (especially at the tails) suggest skewness or heavy tails (non-normality).

Ideal case: Points should mostly follow the straight diagonal line.

3. **Scale-Location (or Spread-Location Plot)**

Purpose: Detects heteroskedasticity (non-constant variance).

Y-axis: Square root of standardized residuals.

X-axis: Fitted values.

If points fan out (wider spread at higher fitted values), it indicates increasing variance → heteroskedasticity.

Ideal case: Residuals are evenly spread across the range of fitted values.

4. **Residuals vs Leverage (Cook’s Distance Plot)**

Purpose: Identifies influential points that may disproportionately impact the model.

Leverage measures how extreme an observation's predictors are.

Residuals measure how far observations deviate from the model’s predictions.

Points with high leverage and large residuals can unduly influence the regression (outliers).

Ideal case: No points with extreme leverage or high Cook’s Distance.

Summary of What Each Plot Helps With

Plot Type	Checks For	Indication of Issue
Residuals vs Fitted	Non-normality, heteroskedasticity	Curved pattern: Model misfits Fan shape : Heteroskedasticity
Normal Q-Q	Normality of residuals	Deviation from diagonal: Skewness or heavy tail
Scale Location	Heteroskedasticity	Fan shape : increasing variance
Residuals vs Leverage	Influence of points	High leverage + large residuals: influential outlier

The implementations for generating Residuals vs Fitted, Normal Q-Q, Scale-Location, and Residuals vs Leverage plots in Python and R, separately.

Python Implementation (Using statsmodels & matplotlib)

```
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.graphics.gofplots import qqplot

# Generate sample data
np.random.seed(42)
X = np.random.rand(100) * 10
y = 2 + 3 * X + np.random.randn(100) * 2  # Adding noise

# Fit a linear model
X = sm.add_constant(X)  # Add intercept
```

```
model = sm.OLS(y, X).fit()
fitted_vals = model.fittedvalues
residuals = model.resid
standardized_residuals = residuals / np.std(residuals)

# 1. Residuals vs Fitted
plt.figure(figsize=(6, 4))
sns.residplot(x=fitted_vals, y=residuals, lowess=True, line_kws={'color': 'red'})
plt.axhline(y=0, linestyle="--", color="black")
plt.xlabel("Fitted values")
plt.ylabel("Residuals")
plt.title("Residuals vs Fitted")
plt.show()

# 2. Normal Q-Q Plot
qqplot(residuals, line='45')
plt.title("Normal Q-Q Plot")
plt.show()

# 3. Scale-Location Plot
plt.figure(figsize=(6, 4))
plt.scatter(fitted_vals, np.sqrt(np.abs(standardized_residuals)), alpha=0.6)
sns.regplot(fitted_vals, np.sqrt(np.abs(standardized_residuals)), lowess=True, scatter=False,
line_kws={'color': 'red'})
plt.xlabel("Fitted values")
plt.ylabel("√|Standardized Residuals|")
plt.title("Scale-Location Plot")
plt.show()

# 4. Residuals vs Leverage (Cook’s Distance)
influence = model.get_influence()
leverage = influence.hat_matrix_diag
cooks_d = influence.cooks_distance[0]

plt.figure(figsize=(6, 4))
plt.scatter(leverage, residuals, alpha=0.6)
plt.xlabel("Leverage")
plt.ylabel("Residuals")
plt.title("Residuals vs Leverage")
plt.axhline(y=0, linestyle="--", color="black")
plt.show()
```

R Implementation

```
# Load required libraries
library(ggplot2)
library(car)

# Generate sample data
set.seed(42)
X <- runif(100, min=0, max=10)
y <- 2 + 3 * X + rnorm(100, mean=0, sd=2)  # Adding noise

# Fit a linear model
model <- lm(y ~ X)
residuals <- resid(model)
fitted_vals <- fitted(model)
std_residuals <- residuals / sd(residuals)

# 1. Residuals vs Fitted Plot
ggplot(data.frame(Fitted=fitted_vals, Residuals=residuals), aes(x=Fitted, y=Residuals)) +
  geom_point(alpha=0.6) +
  geom_smooth(method="loess", color="red", se=FALSE) +
  geom_hline(yintercept=0, linetype="dashed", color="black") +
  labs(title="Residuals vs Fitted", x="Fitted Values", y="Residuals")

# 2. Normal Q-Q Plot
qqPlot(model, main="Normal Q-Q Plot")

# 3. Scale-Location Plot
ggplot(data.frame(Fitted=fitted_vals, Scale=sqrt(abs(std_residuals))), aes(x=Fitted, y=Scale)) +
  geom_point(alpha=0.6) +
  geom_smooth(method="loess", color="red", se=FALSE) +
  labs(title="Scale-Location Plot", x="Fitted Values", y="√|Standardized Residuals|")

# 4. Residuals vs Leverage Plot
influencePlot(model, main="Residuals vs Leverage")
```

Explanation of Each Plot

1. Residuals vs Fitted

Should show random scatter. A pattern suggests non-linearity or heteroskedasticity.

2. Normal Q-Q Plot

Should be close to a straight line. Deviations indicate non-normal residuals.

3. Scale-Location Plot

Helps check heteroskedasticity. A fan-shaped pattern suggests increasing variance.

4. Residuals vs Leverage (Cook’s Distance)

Detects influential points. High-leverage points with large residuals may distort the model.