Antonio Battaglia

Homoscedasticity and Heteroscedasticity

Summary:
This essay explores the concepts of homoscedasticity and heteroscedasticity in statistical modeling, focusing on linear regression. Homoscedasticity is crucial for accurate modeling, while deviations, known as heteroscedasticity, can lead to biased estimates. Correction methods include data transformation and model specification adjustment. Various tests, like the Breusch–Pagan and Koenker–Bassett tests, assess homoscedasticity. The importance of addressing heteroscedasticity is emphasized for accurate statistical modeling. The text provides a comprehensive overview, covering challenges, correction methods, and testing procedures in statistical modeling, resembling an abstract.

Introduction:

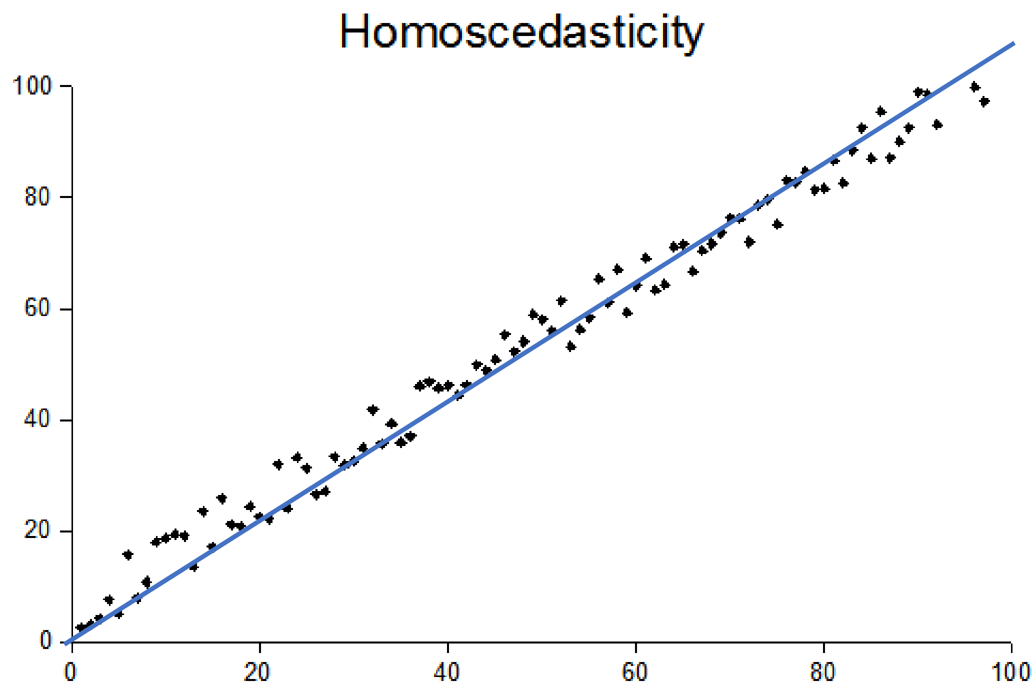
In the realm of statistics, a sequence or vector of random variables is deemed homoscedastic if all constituent random variables exhibit identical finite variance. It is also a condition referred to as homogeneity of variance. The corollary concept is heteroscedasticity, denoting the presence of variance heterogeneity. The spellings "homoskedasticity" and "heteroskedasticity" are interchangeably used. Assuming homoscedasticity when heteroscedasticity is the reality leads to unbiased yet inefficient point estimates and biased estimates of standard errors. This misjudgment may also result in an overestimation of the goodness of fit, quantified by the Pearson coefficient.

The existence of heteroscedasticity emerges as a paramount concern in regression analysis and analysis of variance, rendering statistical tests of significance invalid under the assumption that modeling errors uniformly share the same variance. While the ordinary least squares estimator remains unbiased in the presence of heteroscedasticity, its efficiency diminishes, and inferences grounded in the assumption of homoskedasticity become deceptive. Historically, generalized least squares (GLS) was a common recourse in such instances. Presently, the standard approach in econometrics favors the incorporation of heteroskedasticity-consistent standard errors over GLS, as the latter can exhibit pronounced bias in small samples when the actual skedastic function is unknown.

The pertinence of heteroscedasticity revolves around the expectations of the second moment of errors, categorizing its presence as a misspecification of the second order.

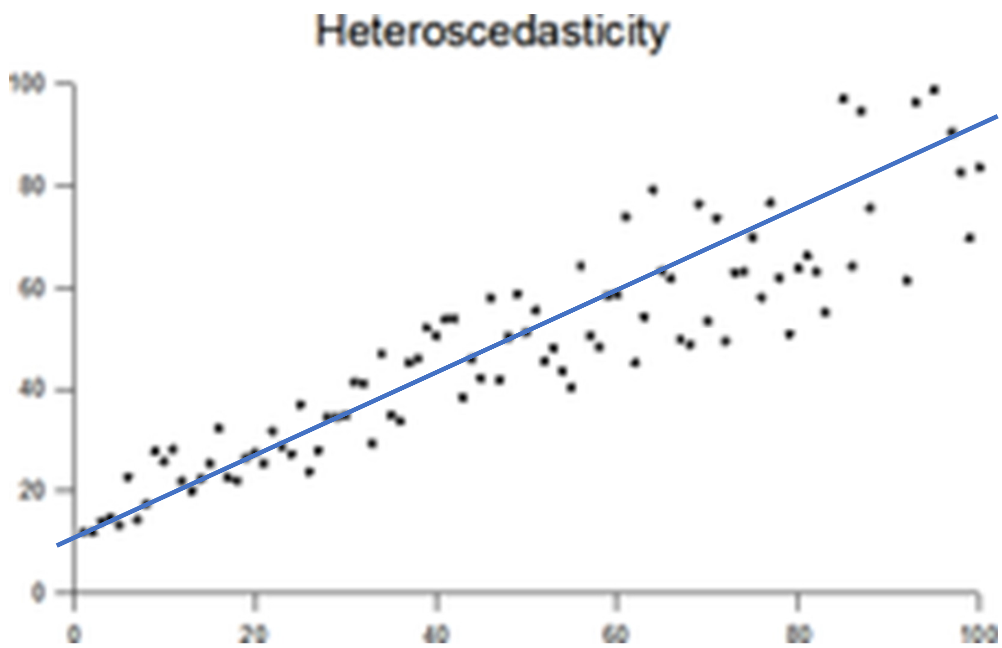
In statistic, we refer to “homoscedastic” or “homogeneity of variance”, when a sequence of random variables has the same finite variance. The assumption of homoscedasticity, denoting uniform variance, plays a pivotal role in the accuracy of linear regression models. It signifies that the error term, representing the random variability in the relationship between independent and dependent variables, remains constant across all levels of the independent variables.

The assessment of homoscedasticity often relies on visual aids, particularly scatter plots. These plots help depict patterns in the variance of residuals, differentiating between homoscedastic and heteroscedastic data.



In a scatter plot, when points align well along the central line or the trendline, it suggests a strong linear relationship between the variables being plotted. The central line, often a regression line or line of best fit, represents the general trend or pattern in the data. Points clustering closely around this line indicate that the variables are correlated and that changes in one variable correspond to predictable changes in the other.

The concept complementary to homoscedasticity is termed heteroscedasticity, denoting the presence of heterogeneity in variance. Heteroscedasticity poses a significant challenge in regression analysis and analysis of variance, as it undermines the validity of statistical tests of significance that presuppose uniform variance among modeling errors.



DEFINITION:

Consider the linear regression model expressed as:

$$y_i = x_i \beta_i + \varepsilon_i, \; i = 1, \ldots, N,$$

Where y_i is the dependent random variable, the x_i is the deterministic variable, β_i is the coefficient, and ε_i is a random disturbance term with a mean of zero. The disturbances are characterized as homoscedastic if the variance of ε_i remains constant. Conversely, if the variance of ε_i varies with i or the value of x_i , the disturbances are deemed heteroscedastic. An illustrative example of heteroscedasticity is when σi2=xiσ2 indicating a variance proportional to the value of x.

In a broader context, the heteroscedasticity of disturbances is discerned by examining the variance-covariance matrix across i .

If the diagonal elements of this matrix are non-constant, the disturbances are heteroscedastic. Three matrices (A, B, and C) are presented, each representing different scenarios. Matrix A reflects homoscedastic disturbances, where OLS is the best linear unbiased estimator. Matrix B exhibits heteroscedasticity with a time-varying variance steadily increasing across time, while Matrix C displays heteroscedasticity with variance contingent on the value of x. Matrix D, despite having non-zero off-diagonal covariances, is homoscedastic due to constant diagonal variances. However, OLS is inefficient in this case due to serial correlation.
The matrix are expressed as:

$$A = \sigma^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad B = \sigma^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \qquad C = \sigma^2 \begin{bmatrix} x_1 & 0 & 0 \\ 0 & x_2 & 0 \\ 0 & 0 & x_3 \end{bmatrix} \qquad D = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

Examples:
Heteroscedasticity frequently arises when exist a substantial disparity in the magnitudes of observations.

Consider the scenario of income versus expenditure on meals, which serves as a classic illustration of heteroscedasticity. Individuals with higher incomes often display a wider range of variability in their food expenditures. This variability arises from the fact that wealthier individuals may, at times, opt for less expensive food choices and, at other times, indulge in more costly dining. In contrast, individuals with lower incomes tend to consistently opt for more economical food options, resulting in a more uniform level of expenditure.

Another instance where heteroscedasticity manifests is in the measurement of rocket distance during a launch. Initially, measurements may be precise, recorded to the nearest centimeter per second. However, as time progresses, and the rocket covers greater distances, atmospheric conditions, inherent inaccuracies, and other factors can contribute to reduced measurement precision. After five minutes, the accuracy of distance measurements may degrade to only 100 meters. This degradation in measurement precision exemplifies heteroscedasticity, emphasizing that variability in measurement error increases with the distance traveled during the rocket launch.

Correlation:
One of the fundamental assumptions in the classical linear regression model is the absence of heteroscedasticity. Deviation from this assumption undermines the applicability of the Gauss–Markov theorem, thereby nullifying the status of Ordinary Least Squares (OLS) estimators as the Best Linear Unbiased Estimators (BLUE). In the presence of heteroscedasticity, the variance of OLS estimators is not necessarily the lowest among all unbiased estimators. Although heteroscedasticity itself does not induce bias in ordinary least squares coefficient estimates, it can lead to biased estimates of the variance (and, consequently, standard errors) of the coefficients, potentially deviating from the true population variance. Consequently, while regression analysis with heteroscedastic data yields an unbiased estimate of the relationship between predictor variables and outcomes, the integrity of standard errors and, consequently, inferences derived from data analysis may be compromised. Biased standard errors can result in incorrect inferences from hypothesis tests, leading to potential errors such as a Type II error.

Despite these challenges, under certain assumptions, the OLS estimator demonstrates a normal asymptotic distribution when appropriately normalized and centered, even in the presence of heteroscedasticity. This justification supports the use of normal or chi-square distributions in hypothesis testing, depending on the test statistic calculation. White's proposal in 1980 for a consistent estimator of the variance-covariance matrix under heteroscedasticity further legitimizes hypothesis testing using OLS estimators and White's variance-covariance estimator.

Heteroscedasticity is not exclusive to linear regression; it is a significant practical concern in ANOVA problems as well. While the F-test remains applicable in some scenarios, addressing heteroscedasticity is advised only when its impact is substantial. Cautionary voices in the field emphasize that heteroscedasticity should not be a sole reason to discard an otherwise sound model.

In the context of non-linear models such as Logit and Probit models, the consequences of heteroscedasticity are more severe. Maximum Likelihood Estimates (MLE) of parameters in these models tend to be biased and inconsistent unless the likelihood function is appropriately modified to account for the specific form of heteroscedasticity. However, in binary choice models like Logit or Probit, heteroscedasticity mainly results in a positive scaling effect on the asymptotic mean of the misspecified MLE, leaving predictions intact. Despite advancements, caution is warranted, as the virtue of a robust covariance matrix in the presence of heteroscedasticity remains uncertain, and hypotheses testing with inconsistent estimators remains a nuanced endeavor.

Correction:

Various strategies can be employed to address the issue of heteroscedasticity in statistical modeling. These corrections include:

Data Transformation:
Applying a stabilizing transformation to the data, such as logarithmic transformation, can mitigate heteroscedasticity. This is particularly beneficial for non-logarithmic series that exhibit increasing variability over time due to exponential growth. Although the variability in percentage terms may still be stable, transforming the data can enhance model performance.

Model Specification Adjustment:
Altering the model specification by incorporating different independent variables (X variables) or non-linear transformations of existing X variables can be effective in mitigating heteroscedasticity. This approach allows for a more flexible representation of the underlying relationships within the data.

Weighted Least Squares Estimation:
Employing a Weighted Least Squares (WLS) estimation method involves applying OLS to transformed or weighted values of X and Y. The weights vary across observations, typically reflecting changing error variances. One variant involves weights directly related to the magnitude of the dependent variable, known as least squares percentage regression.

Heteroscedasticity-Consistent Standard Errors (HCSE):
Heteroscedasticity-consistent standard errors, while still exhibiting bias, offer an improvement over OLS estimates. HCSE provides a

consistent estimator of standard errors in the presence of heteroscedasticity without altering the coefficient values. This method is advantageous because it corrects for heteroscedasticity when present, while reverting to conventional standard errors estimated by OLS when the data is homoscedastic.

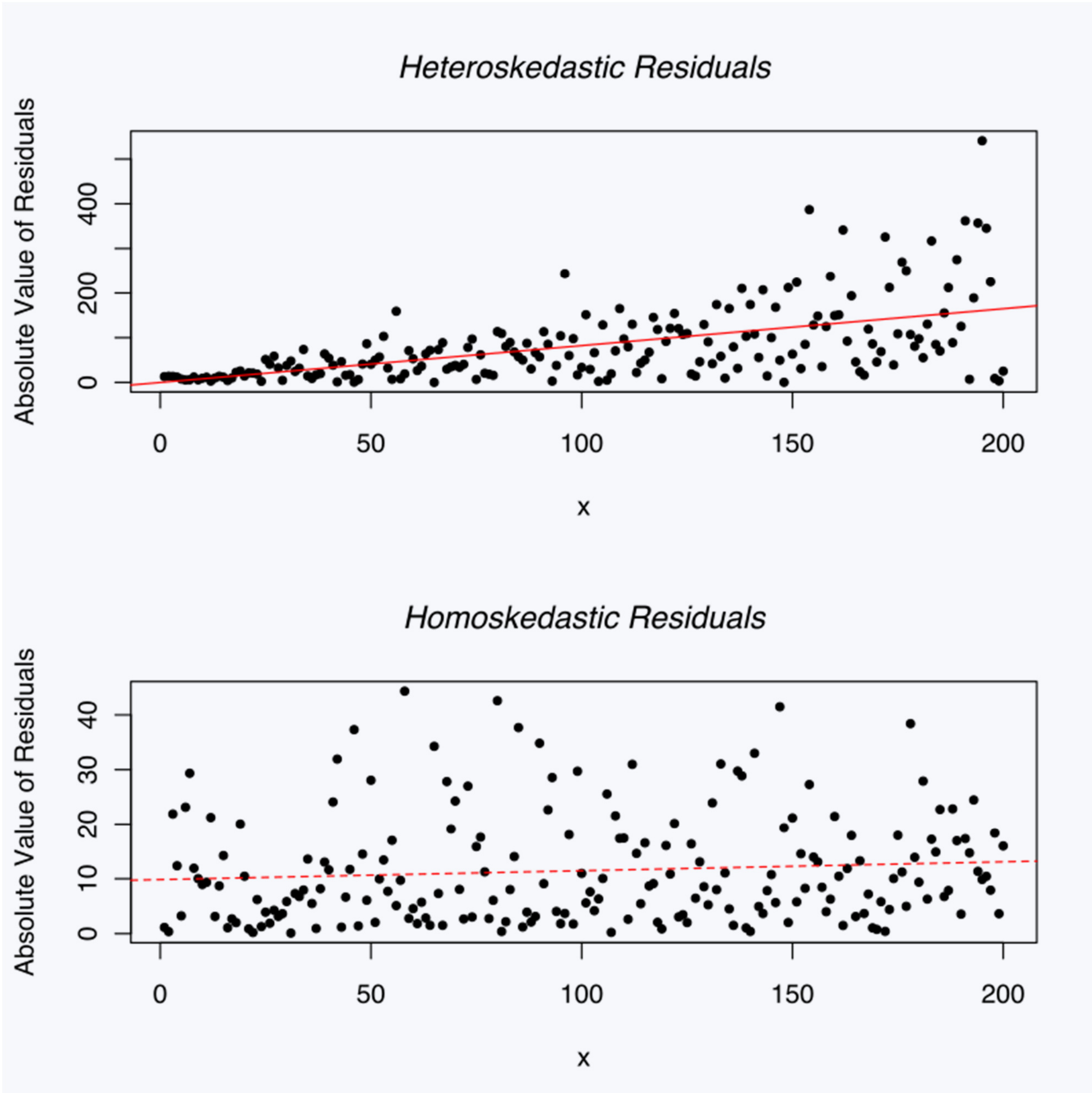
Wild Bootstrapping:
Wild bootstrapping, as a resampling method, respects differences in the conditional variance of the error term. An alternative approach involves resampling observations rather than errors. It's important to note that resampling errors without considering the associated values of the observation enforces homoskedasticity and may lead to incorrect inference.

MINQUE and Customary Estimators:
Using MINQUE or other customary estimators can also be considered. For instance, MINQUE involves estimating variances for independent samples, and its efficiency losses are minimal, particularly when the number of observations per sample is large.

These correction methods provide researchers with a toolkit to address heteroscedasticity in various scenarios, allowing for more robust and accurate statistical analyses.

Testing:

Residuals can undergo a homoscedasticity assessment through the Breusch–Pagan test, which conducts an auxiliary regression by squaring the residuals against independent variables. In this auxiliary regression, the retained explained sum of squares is halved and serves as the test statistic following a chi-squared distribution, with degrees of freedom matching the number of independent variables. The null hypothesis for this chi-squared test posits homoscedasticity, while the alternative hypothesis implies heteroscedasticity. Recognizing the Breusch–Pagan test's sensitivity to deviations from normality or limited sample sizes, the Koenker–Bassett or 'generalized Breusch–Pagan' test is commonly favored. It preserves the R-squared value from the auxiliary regression, multiplying it by the sample size to yield the chi-squared test statistic (with the same degrees of freedom). Although the Koenker–Bassett test doesn't necessitate it, the Breusch–Pagan test demands that squared residuals be divided by the residual sum of squares divided by the sample size. For assessing groupwise heteroscedasticity, the Goldfeld–Quandt test is a viable option.



Bibliography:

Websites:
https://en.wikipedia.org/wiki/Homoscedasticity_and_heteroscedasticity (last seen 27/11/23 at 11:42)
https://uedufy.com/what-is-homoscedasticity-assumption-in-statistics/ (last seen 27/11/23 at 12:45)
https://www.investopedia.com/terms/h/heteroskedasticity.asp#:~:text=Heteroskedastic%20refers%20to%20a%20condition,a%20reg

ression%20model%20varies%20widely.&text=Homoskedastic%20refers%20to%20a%20condition,a%20regression%20model%20is%20constant.

Videos:
https://www.youtube.com/watch?v=zRklTsY9w9c
https://www.youtube.com/watch?v=uJIEpdhB5-Q
https://www.youtube.com/watch?v=VCYTV35QacM
https://www.youtube.com/watch?v=0c8l5pL_WEI

Papers:
Yang. K et al 2019