The assumption of homoscedasticity (literally, same variance) is central to linear regression models. Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables. Heteroscedasticity (the violation of homoscedasticity) is present when the size of the error term differs across values of an independent variable. The impact of violating the assumption of homoscedasticity is a matter of degree, increasing as heteroscedasticity increases.

A simple bivariate example can help to illustrate heteroscedasticity: Imagine we have data on family income and spending on luxury items. Using bivariate regression, we use family income to predict luxury spending (as expected, there is a strong, positive association between income and spending). Upon examining the residuals we detect a problem – the residuals are very small for low values of family income (families with low incomes don’t spend much on luxury items) while there is great variation in the size of the residuals for wealthier families (some families spend a great deal on luxury items while some are more moderate in their luxury spending). This situation represents heteroscedasticity because the size of the error varies across values of the independent variable. Examining the scatterplot of the residuals against the predicted values of the dependent variable would show the classic cone-shaped pattern of heteroscedasticity.

The problem that heteroscedasticity presents for regression models is simple. Recall that ordinary least-squares (OLS) regression seeks to minimize residuals and in turn produce the smallest possible standard errors. By definition OLS regression gives equal weight to all observations, but when heteroscedasticity is present the cases with larger disturbances have more “pull” than other observations. The coefficients from OLS regression where heteroscedasticity is present are therefore inefficient but remain unbiased. In this case, weighted least squares regression would be more appropriate, as it downweights those observations with larger disturbances.

A more serious problem associated with heteroscedasticity is the fact that the standard errors are biased. Because the standard error is central to conducting significance tests and calculating confidence intervals, biased standard errors lead to incorrect conclusions about the significance of the regression coefficients. Many statistical programs provide an option of robust standard error to correct this bias; weighted least squares regression also addresses this concern but requires a number of additional assumptions. Another approach for dealing with heteroscedasticity is to transform the dependent variable using one of the variance stabilizing transformations. A logarithmic transformation can be applied to highly skewed variables, while count variables can be transformed using a square root transformation. Overall, the
violation of the homoscedasticity assumption must be quite severe in order to present a major problem given the robust nature of OLS regression.