Iman Nekooeimehr
Weighted Least Squares in R

Heteroscedasticity is a huge concern in linear regression models which violates the assumption that the model residuals have a constant variance and are uncorrelated. While the Ordinary (Gaussian) Least Squares (OLS) model is an unbiased estimator for this scenario, it is yet insufficient because the correct variance and covariance are not considered.

Heteroscedasticity is defined mathematically as follows: consider a linear model $y_i = a\,.\,x_i + b + e_i$ where e_i are independent noise from a distribution that depends on x_i, while the conditional mean of the e_i given x_i is zero. As an example, it can be seen from figure 1 that the variance of the residuals is not constant and is increasing as x_is are getting further from zero. It actually looks like the residuals are following a polynomial distribution.

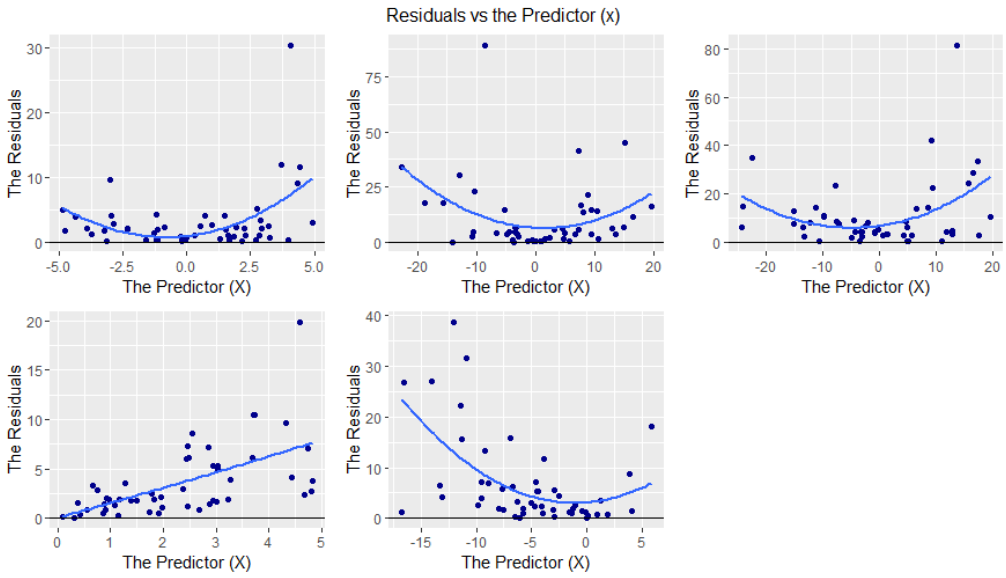

Figure 1. The residuals and their relation with the predictor

An effective way to address the heteroscedasticity is to use the Weighted Least Squares (WLS) method, in which the x_i and y_is are transformed and then OLS is applied on the transformed data. Assuming that for each e_i, $E(e_i) = 0$ and $\sigma^2(e_i) = s_i^2$, if s_i^2 can be estimated, then $y_i = a\,.\,x_i + b + e_i$ can be transformed in a way that $\sigma^2(e_T) = I$, where e_T is the residuals after the transformation. Using this transformation, the OLS can be used without violating any of the assumptions. The loss function in the matrix format would look like below:

$$\min L(B) = (Y - XB)'W(Y - XB) \qquad (1)$$

where

$$W = \begin{bmatrix} 1/s_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/s_n^2 \end{bmatrix} \qquad (2)$$

In order to estimate $L(B)$, its derivative with respect to B is taken and is set equal to zero:

$$(X'WX)B = X'WY \qquad (3)$$

Therefore,

$$B = (X'WX)^{-1}X'WY \qquad (4)$$

The standard deviation of the residuals s_i are estimated using a polynomial distribution of degree 2 as a function of the predictor. The estimated polynomial distribution is used to build the diagonal variance matrix which is later used to build the weight matrix W. The variance matrix is diagonal because it was assumed that the errors are independent from each other and therefore, they do not have any correlation with each other. The process of building a linear model, estimating the weights based on the new residuals and updating the model using WLS is repeated until the difference between the coefficients of two consecutive models are less than a predefined tolerance.

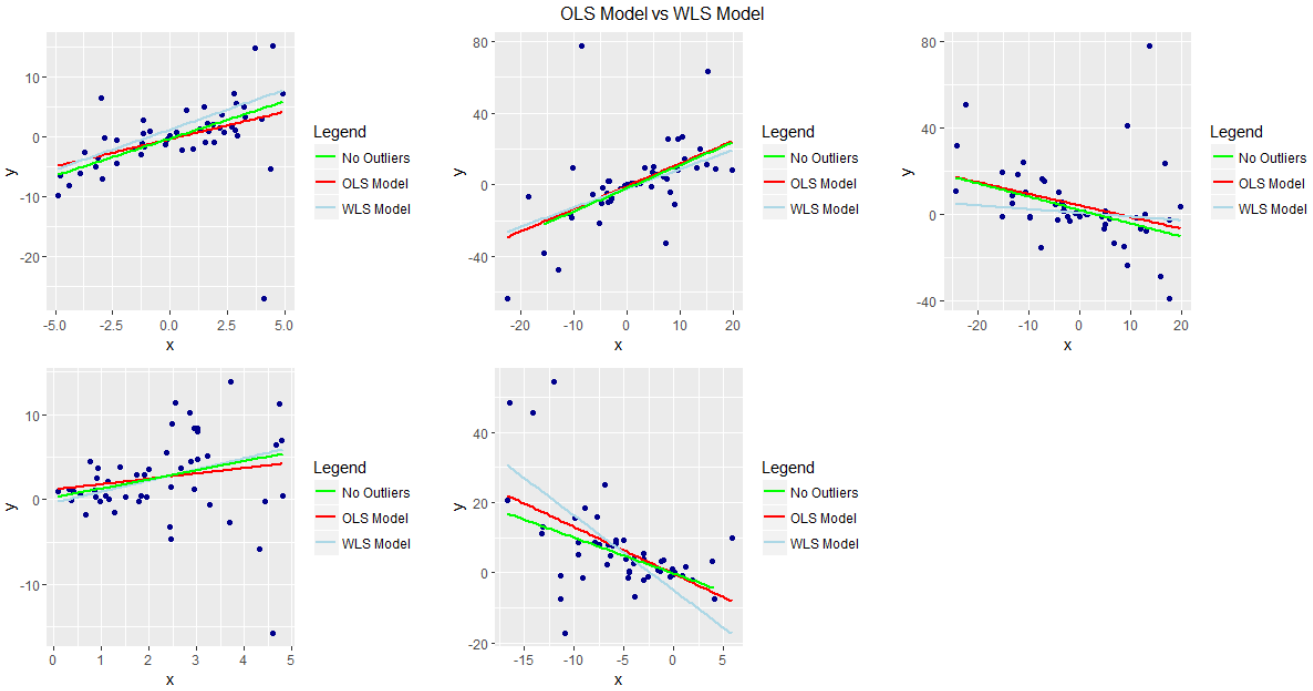

Figure 2. The data points and the three fitted models: 1) OLS Model 2) WLS Model 3) OLS Model after Removing the Outliers

The resulted linear fit for all models are shown in figure 2. The OLS model after removing the outliers is also considered in the results. Therefore, three models are compared: 1) OLS model 2) WLS model 3) OLS model after removing the outliers. The three models are compared based on mean squared error and r-squared on the testing set using 5 fold cross validation. The results for the 5 datasets are shown in table 1 and the best results are bolded. As can be seen from the table 1, it looks like the three models are not much different in terms of both MSE and r-squared.

Table 1. Results of the 3 models on 5 datasets

Measurement	Dataset #	OLS Model	WLS Model	OLS Model after removing outliers
R-Squared	Dataset 1	0.1764	**0.2462**	0.2167
	Dataset 2	**0.5326**	0.4748	0.4522
	Dataset 3	0.1652	0.3493	**0.3970**
	Dataset 4	**0.5882**	0.1129	0.1319
	Dataset 5	**0.3743**	0.1843	0.2271
MSE	Dataset 1	31.0114	29.8794	**29.4715**
	Dataset 2	**245.9360**	246.6825	249.3700
	Dataset 3	356.8396	354.0216	**353.2434**
	Dataset 4	31.7110	30.2135	**29.6320**
	Dataset 5	134.7961	140.0895	**131.7174**

In order to test if there exists any significant difference among the 3 models, the Friedman's test is used which is a non-parametric equivalent to ANOVA. The p-value is 0.5945 for the r-squared and 0.2276 for the MSE which indicates there does not exist enough evidence to reject the null hypothesis. Therefore, we cannot conclude the three methods are significantly different for both measurements.