Assumptions of Linear Regression Model

4. The error terms do not have constant variance.

One of the assumptions necessary for the validity of regression inferences is that the error term ε have constant variance σ^2 for all levels of the independent variable(s). Symbolically,

$$E(\varepsilon_i) = \sigma^2 \qquad i = 1,2,...,n$$

Variances that satisfy this property are called **homoscedastic**. Unequal variances for different settings of the independent variable(s) are said to be **heteroscedastic**. Various statistical tests and graphical method for heteroscedasticity have been developed.

There are several reasons why the variances of ε_i may be variable, some of which are as follows.

1. Heteroscedasticity arise as a result of the presence of **outliers.** The inclusion or exclusion of an observation that is considered outlier, especially if the sample size is small, can substantially alter the results of regression analysis.
2. Heteroscedasticity arises from violating of classical linear regression model, namely, that the regression model is ***correctly specified***.
3. Another source of heteroscedasticity is **skewness** in the distribution of one or more regressors included in the model.
4. Other sources of heteroscedasticity: As David Hendry notes, heteroscedasticity can also arise because of (1) incorrect data transformation (e.g., ratio or first difference transformations) and (2) incorrect functional form (e.g., linear versus log–linear models).

Homoscedasticity is necessary to calculate accurate standard errors for parameter estimates that is why we need to do a remedial measures if this assumption is violated. Some remedial measures to address the issue is to modify the model, fit a generalized linear model, or run a weighted least squares regression.

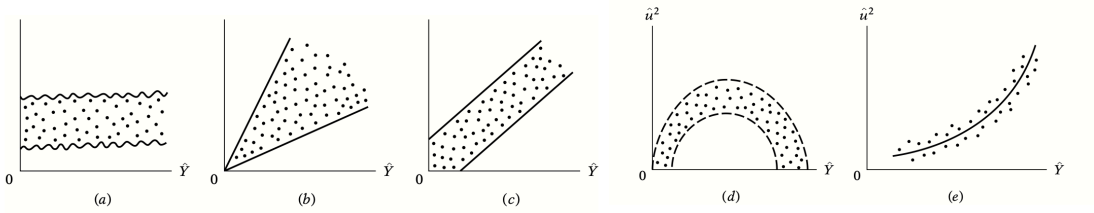

Figure 6.3.1. Hypothetical patterns of estimated squared residuals.

In **Figure 6.3.1**, ε_i^2 are plotted against $\hat{Y}_i$, the estimated Y_i from the regression line, the idea being to find out whether the estimated mean value of Y is systematically related to the squared residual. In **Figure 6.3.1*a*** we see that there is no systematic pattern between the two variables, suggesting that perhaps no heteroscedasticity is present in the data. **Figure 6.3.1*b*** to **Figure 6.3.1*e***, however, exhibits definite patterns. For instance, **Figure 6.3.1*c*** suggests a linear relationship, whereas **Figure 6.3.1*d*** and **Figure 6.3.1*e*** indicates a quadratic relationship between ε_i^2 and $\hat{Y}_i$. Using such knowledge, although informal, one may transform the data in such a manner that the transformed data do not exhibit heteroscedasticity.

Instead of plotting ε_i^2 against $\hat{Y}_i$, one may plot them against one of the explanatory variables, especially if plotting ε_i^2 against $\hat{Y}_i$ results in the pattern shown in **Figure 6.3.1*a***. Such a plot, which is shown in **Figure 6.3.2**, may reveal patterns similar to those given in **Figure 6.3.1**. (In the case of the two-variable model, plotting ε_i^2 against $\hat{Y}_i$ is equivalent to plotting it against X_i, and therefore **Figure 6.3.2** is similar to **Figure 6.3.1**. But this is not the situation when we consider a model involving two or more X variables; in this instance, ε_i^2 may be plotted against any X variable included in the model.)

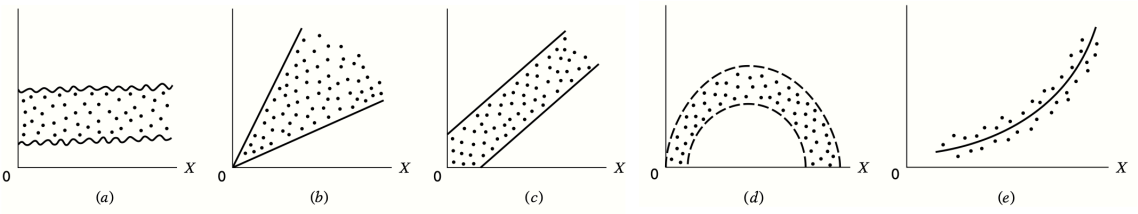

Figure 6.3.2 Scattergram of estimated squared residuals against *X*.

Statistical Test use to Detect Heteroscedasticity

1. Park Test
2. Glejser Test
3. Goldfeld-Quandt Test
4. Breusch–Pagan–Godfrey Test
5. White's General Heteroscedasticity Test

Take Note!!!

- Heteroscedasticity does not destroy the unbiasedness and consistency properties of OLS estimators.
- But these estimators are no longer minimum variance or efficient. That is, they are not BLUE.
- The BLUE estimators are provided by the method of weighted least squares, provided the heteroscedastic error variances, σ_i^2, are known.
- In the presence of heteroscedasticity, the variances of OLS estimators are not provided by the usual OLS formulas. But if we persist in using the usual OLS formulas, the t and F tests based on them can be highly misleading, resulting in erroneous conclusions.
- Even if heteroscedasticity is suspected and detected, it is not easy to correct the problem. If the sample is large, one can obtain White's heteroscedasticity corrected standard errors of OLS estimators and conduct statistical inference based on these standard errors.

5. For given X's, there is autocorrelation, or serial correlation, between the error term.

The term **autocorrelation** may be defined as "correlation between members of series of observations ordered in time (as in time series data) or space (as in cross-sectional data)." In the regression context, the classical linear regression model assumes that such autocorrelation does not exist in the error term ε_i. Symbolically,

$$E(\varepsilon_i \varepsilon_j) = 0 \qquad i \neq j$$

Put simply, the classical model assumes that the disturbance term relating to any observation is not influenced by the disturbance term relating to any other observation.

Durbin - Watson *d* Test

The most celebrated test for detecting serial correlation is that developed by statisticians Durbin and Watson. It is popularly known as the **Durbin– Watson *d* statistic,** which is defined as

$$d = \frac{\sum_{t=2}^{t=n} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{t=n} \hat{\varepsilon}_t^2} \tag{6.1}$$

which is simply the ratio of the sum of squared differences in successive residuals to the sum of squares residual. The actual test procedure can be explained better with the aid of **Figure 6.3.3**, which shows that the limits of d are 0 and 4.

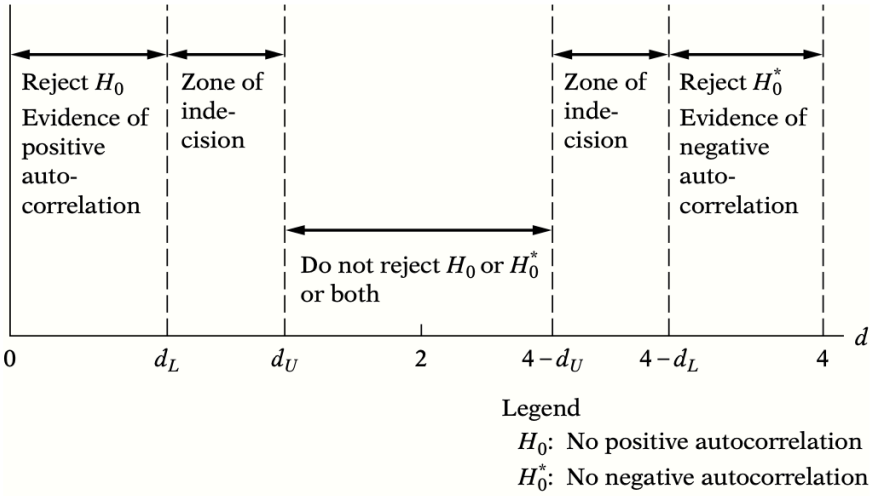

Figure 6.3.3. Durbin-Watson *d* Statistic

(**6.1**) can be written as

$$d \approx 2\left(1 - \frac{\sum_{t=2}^{t=n}(\hat{\varepsilon}_t \hat{\varepsilon}_{t-1})}{\sum_{t=1}^{t=n}\hat{\varepsilon}_t^2}\right) \tag{6.2}$$

Now let us define

$$\hat{\rho} = \frac{\sum_{t=2}^{t=n}(\hat{\varepsilon}_t \hat{\varepsilon}_{t-1})}{\sum_{t=1}^{t=n}\hat{\varepsilon}_t^2} \tag{6.3}$$

Using (**6.3**), we can express (**6.2**) as

$$d \approx 2(1 - \hat{\rho})$$

but since $-1 \leq \rho \leq 1$, implies that $0 \leq d \leq 4$. These are the bounds of d; any estimated d value must lie within these limits.

The Rule of Thumb:

If d is found to be 2 in an application, one may assume that there is no first-order autocorrelation, either positive or negative.

Question?

If $\hat{\rho} = 1$, what will be the value of d, and what will you conclude?

If $\hat{\rho} = -1$, what will be the value of d, and what will you conclude?

Remedial Measures

1. Try to find out if the autocorrelation is **pure autocorrelation** and not the result of misspecification of the model. Sometimes we observe patterns in residuals because the model is misspecified—that is, it has excluded some important variables—or because its functional form is incorrect.
2. If it is pure autocorrelation, one can use appropriate transformation of the original model so that in the transformed model we do not have the problem of (pure) autocorrelation. As in the case of heteroscedasticity, we will have to use some type of **generalized least-square (GLS) method.**
3. In large samples, we can use the **Newey–West** method to obtain standard errors of OLS estimators that are corrected for autocorrelation. This method is actually an extension of White's heteroscedasticity-consistent standard errors method that we discussed in the previous chapter.
4. In some situations we can continue to use the OLS method.

Take Note!!!

In the presence of autocorrelation, the OLS estimators remain unbiased, consistent, and asymptotically normally distributed; however, they are no longer efficient. As a result, the usual t, F, and χ^2 tests cannot be legitimately applied. Therefore, remedial measures may be necessary.

Prepared by:

KATRINA D. ELIZON
Department of Mathematics and Statistics
College of Science