Classical Linear Regression Model: Assumptions and Diagnostic Tests

Yan Zeng

Version 1.1, last updated on 10/05/2016

Abstract

Summary of statistical tests for the Classical Linear Regression Model (CLRM), based on Brooks [1], Greene [5] [6], Pedace [8], and Zeileis [10].

Contents

1 The Classical Linear Regression Model (CLRM)

Let the column vector $\boldsymbol{x}_k$ be the T observations on variable x_k, $k = 1, \cdots, K$, and assemble these data in an $T \times K$ data matrix $\boldsymbol{X}$. In most contexts, the first column of $\boldsymbol{X}$ is assumed to be a column of 1s:

$$\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{T\times 1}$$

so that β_1 is the constant term in the model. Let $\boldsymbol{y}$ be the T observations $y_1, \cdots, y_T$, and let $\boldsymbol{\varepsilon}$ be the column vector containing the T disturbances. The **Classical Linear Regression Model** (CLRM) can be written as

$$\boldsymbol{y} = \boldsymbol{x}_1\beta_1 + \cdots + \boldsymbol{x}_K\beta_K + \boldsymbol{\varepsilon}, \ \boldsymbol{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iT} \end{bmatrix}_{T\times 1}$$

or in matrix form

$$\boldsymbol{y}_{T\times 1} = \boldsymbol{X}_{T\times K}\boldsymbol{\beta}_{K\times 1} + \boldsymbol{\varepsilon}_{T\times 1}.$$

Assumptions of the CLRM (Brooks [1, page 44], Greene [6, page 16-24]):

(1) **Linearity:** The model specifies a linear relationship between y and $x_1, \cdots, x_K$.

$$\boxed{\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$$

(2) **Full rank:** There is no exact linear relationship among any of the ndependent variables in the model. This assumption will be necessary for estimation of the parameters of the model (see formula (1)).

$$\boxed{\boldsymbol{X} \text{ is a } T \times K \text{ matrix with rank } K.}$$

(3) **Exogeneity of the independent variables:** $E[\varepsilon_i | x_{j1}, x_{j2}, \cdots, x_{jK}] = 0$. This states that the expected value of the disturbance at observation i in the sample is not a function of the independent variables observed at any observation, including this one. This means that the independent variables will not carry useful information for prediction of ε_i.

$$\boxed{E[\boldsymbol{\varepsilon}|\boldsymbol{X}] = \boldsymbol{0}.}$$

(4) **Homoscedasticity and nonautocorrelation:** Each disturbance, ε_i has the same finite variance, σ^2, and is uncorrelated with every other disturbance, ε_j.

$$\boxed{E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}] = \sigma^2\boldsymbol{I}.}$$

(5) **Data generation:** The data in $(x_{j1}, x_{j2}, \cdots, x_{jK})$ may be any mixture of constants and random variables.

$$\boxed{\boldsymbol{X} \text{ may be fixed or random.}}$$

(6) **Normal distribution:** The disturbances are normally distributed.

$$\boxed{\boldsymbol{\varepsilon}|\boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}).}$$

In order to obtain estimates of the parameters $\beta_1, \beta_2, \cdots, \beta_K$, the *residual sum of squares* (RSS)

$$RSS = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \sum_{t=1}^{T} \hat{\varepsilon}_t^2 = \sum_{t=1}^{T}\left(y_t - \sum_{i=1}^{K} x_{it}\beta_i\right)^2$$

is minimised so that the coefficient estimates will be given by the *ordinary least squares (OLS) estimator*

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \cdots \\ \hat{\beta}_k \end{bmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{1}$$

In order to calculate the standard errors of the coefficient estimates, the variance of the errors, σ^2, is estimated by the estimator

$$s^2 = \frac{RSS}{T-K} = \frac{\sum_{t=1}^{T} \hat{\varepsilon}_t^2}{T-K} \tag{2}$$

where we recall K is the number of regressors including a constant. In this case, K observations are "lost" as K parameters are estimated, leaving $T-K$ degrees of freedom.

Then the parameter variance-covariance matrix is given by

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{3}$$

And the coefficient standard errors are simply given by taking the square roots of each of the terms on the leading diagonal. In summary, we have (Brooks [1, page 91-92])

$$\begin{cases} \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \\ s^2 = \frac{\sum_{t=1}^{T} \hat{\varepsilon}_t^2}{T-K} \\ \mathrm{Var}(\hat{\boldsymbol{\beta}}) = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \end{cases} \tag{4}$$

The OLS estimator is the best linear unbiased estimator (BLUE), consistent and asymptotically normally distributed (CAN), and if the disturbances are normally distributed, asymptotically efficient among all CAN estimators.

2 Hypothesis Testing: The t-test and The F-test

The t-statistic for hypothesis testing is given by

$$\frac{\hat{\beta}_i - \text{hypothesized value}}{SE(\hat{\beta}_i)} \sim t(T-K)$$

where $SE(\hat{\beta}_i) = \sqrt{\mathrm{Var}(\hat{\boldsymbol{\beta}})_{ii}}$, and is used to test single hypotheses. The F-test is used to test more than one coefficient simultaneously.

Under the F-test framework, two regressions are required. The *unrestricted regression* is the one in which the coefficients are freely determined by the data, and the *restricted regression* is the one in which the coefficients are restricted, i.e. the restrictions are imposed on some βs. Thus the F-test approach to hypothesis testing is also termed *restricted least squares*.

The F-test statistic for testing multiple hypotheses about the coefficient estimate is given by

$$\frac{RRSS - URSS}{URSS} \times \frac{T-K}{m} \sim F(m, T-K) \tag{5}$$

where $URSS$ is the residual sum of squares from unrestricted regression, $RRSS$ is the residual sum of squares from restricted regression, m is the number of restrictions[1], T is the number of observations, and K is the number of regressors in the unrestricted regression.

To see why the test centres around a comparison of the residual sums of squares from the restricted and unrestricted regressions, recall that OLS estimation involved choosing the model that minimised the residual

[1] Informally, the number of restrictions can be seen as "the number of equality signs under the null hypothesis".

sum of squares, with no constraints imposed. Now if, after imposing constraints on the model, a residual sum of squares results that is not much higher than the unconstrained model's residual sum of squares, it would be concluded that the restrictions were supported by the data. On the other hand, if the residual sum of squares increased considerably after the restrictions were imposed, it would be concluded that the restrictions were not supported by the data and therefore that the hypothesis should be rejected.

It can be further stated that $RRSS \geq URSS$.[2] Only under a particular set of very extreme circumstances will the residual sums of squares for the restricted and unrestricted models be exactly equal. This would be the case when the restriction was already present in the data, so that it is not really a restriction at all.

Finally, we note any hypothesis that could be tested with a t-test could also have been tested using an F-test, since

$$t^2(T-K) \sim F(1, T-K).$$

3 Violation of Assumptions: Multicollinearity

If the explanatory variables were orthogonal to one another, adding or removing a variable from a regression equation would not cause the values of the coefficients on the other variables to change. *Perfect multicollinearity* will make it impossible to invert the $(\boldsymbol{X'X})$ matrix since it would not be of full rank. Technically, the presence of high multicollinearity doesn't violate any CLRM assumptions. Consequently, OLS estimates can be obtained and are BLUE with high multicollinearity. The larger variances (and standard errors) of the OLS estimators are the main reason to avoid high multicollinearity.

Causes of multicollinearity include
- You use variables that are lagged values of one another.
- You use variables that share a common time trend component.
- You use variables that capture similar phenomena.

3.1 Detection of multicollinearity

Testing for multicollinearity is surprisingly difficult. *Correlation matrix* is one simple method, but if the relationship involves more variables that are collinear, then multicollinearity would be very difficult to detect.

Rule of thumb for identifying multicollinearity. Because high multicollinearity doesn't violate a CLRM assumption and is a sample-specific issue, researchers typically don't use formal statistical tests to detect multicollinearity. Instead, they use two sample measurements as indicators of a potential multicollinearity problem.

• **Pairwise correlation coefficients**. The sample correlation of two independent variables, x_k and x_j, is calculated as

$$r_{kj} = \frac{s_{kj}}{s_k s_j}.$$

As a rule of thumb, correlation coefficients around 0.8 or above may signal a multicollinearity problem. Other evidence you should also check include insignificant t-statistics, sensitive coefficient estimates, and nonsensical coefficient signs and values.

Note the pairwise correlation coefficients only identify the linear relationship of two variables. It does not check linear relationship among more than two variables.

• **Auxiliary regression and the variance inflation factor (VIF)**. A VIF for any given independent variable is calculated by

$$VIF_k = \frac{1}{1 - R_k^2}$$

where R_k^2 is the R-squared value obtained by regressing independent variable x_k on all the other independent variables in the model.

[2] Recall URSS is the shortest distance from a vector to its projection plane.

As a rule of thumb, VIFs greater than 10 signal a highly likely multicollinearity problem, and VIFs between 5 and 10 signal a somewhat likely multicollinearity issue. Remember to check also other evidence of multicollinearity (insignificant t-statistics, sensitive or nonsensical coefficient estimates, and nonsensical coefficient signs and values). A high VIF is only an indicaotr of potential multicollinearity, but it may not result in a large variance for the estimator if the variance of the independent variable is also large.

3.2 Consequence of ignoring near multicollinearity

First, R^2 will be high but the individual coefficients will have high standard errors, so that the regression "looks good" as a whole, but the individual variables are not significant. This arises in the context of very closely related explanatory variables as a consequence of the difficulty in observing the individual contribution of each variable to the overall fit of the regression.

Second, the regression becomes very sensitive to small changes in the specification, so that adding or removing an explanatory variable leads to large changes in the coefficient values or significance of the other variables. The intuition is that, if the independent variables are highly collinear, the estimates must emphasize small differences in the variables in order to assign an independent effect to each of them.

Third, nonsensical coefficient signs and magnitudes. With higher multicollinearity, the variance of the estimated coefficients increases, which in turn increases the chances of obtaining coefficient estimates with extreme values.

Finally, near multicollinearity will thus make confidence intervals for the parameters very wide, and significance tests might therefore give inappropriate conclusions, and so make it difficult to draw sharp inferences.

3.3 Dealing with multicollinearity

A number of alternative estimation techniques have been proposed that are valid in the presence of multicollinearity – for example, ridge regression, or principal components. Many researchers do not use these techniques, however, as they can be complex, their properties are less well understood than those of the OLS estimator and, above all, many econometricians would argue that multicollinearity is more a problem with the data than with the model or estimation method.

Ad hoc methods include:

- **Ignore it**, if the model is other wise adequate, i.e. statistically and in terms of each coefficient being of a plausible magnitude and having an appropriate sign. The presence of near multicollinearity does not affect the BLUE properties of the OLS estimation. However, in the presence of near multicollinearity, it will be hard to obtain small standard errors.
- **Drop one of the collinear variables**, so that the problem disappears. This may be unacceptable if there were strong *a priori* theoretical reasons for including both variables in the model. Also, if the removed variable was relevant in the data generating process for y, an omitted variable bias would result (see later).
- **Transform the highly correlated variables into a ratio and include only the ratio and not the individual variables in the regression**. This may be unacceptable by financial theory.
- **Increase the sample size**, e.g. by using a pooled sample. This is because near multicollinearity is *more a problem with the data than with the model*.
- **Use a new model**.

First-differencing. Its use is limited to models utilizing time-series or panel data. It also has its cost: 1) losing observations; 2) losing variation in your independent variables (resulting in insignificant coefficients); 3) changing the specification (possibly resulting in misspecification bias).

The composite index variable. But never combine variables into an index that would, individually, be expected to have opposite signs.

4 Violation of Assumptions: Heteroscedasticity

4.1 Detection of heteroscedasticity

This is the situation where $E[\varepsilon_i^2|\boldsymbol{X}]$ is not a finite constant.

4.1.1 The Goldfeld-Quandt test

The **Goldfeld-Quandt test** is based on splitting the total sample of length T into two sub-samples of length T_1 and T_2. The regression model is estimated on each sub-sample and the two residual variances are calculated as

$$s_1^2 = \hat{\boldsymbol{\varepsilon}}_1'\hat{\boldsymbol{\varepsilon}}_1/(T_1 - K),\ s_2^2 = \hat{\boldsymbol{\varepsilon}}_2'\hat{\boldsymbol{\varepsilon}}_2/(T_2 - K)$$

respectively. The null hypothesis is that the variances of the disturbances are equal, against a two-sided alternative. The test statistic, denoted GQ, is simply

$$GQ = \frac{s_1^2}{s_2^2}$$

with $s_1^2 > s_2^2$. The test statistic is distributed as an $F(T_1 - K, T_2 - K)$ under the null hypothesis, and the null of a constant variance is rejected if the test statistic exceeds the critical value.

The GQ test is simple to construct but its conclusions may be contingent upon a particular, and probably arbitrary, choice of where to split the sample. An alternative method that is sometimes used to sharpen the inferences from the test and to increase its power is to omit some of the observations from the centre of the sample so as to introduce a degree of separation between the two sub-samples.

4.1.2 The White's general test

The **White's general test** for heteroscedasticity is carried out as follows.

(1) Assume that the regression model estimated is of the standard linear form, e.g.

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t.$$

To test $\text{Var}(\varepsilon_t) = \sigma^2$, estimate the model above, obtaining the residuals $\hat{\varepsilon}_t$.

(2) Run the auxiliary regression

$$\hat{\varepsilon}_t^2 = \alpha_1 + \alpha_2 x_{2t} + \alpha_3 x_{3t} + \alpha_4 x_{2t}^2 + \alpha_5 x_{3t}^2 + \alpha_6 x_{2t} x_{3t} + \nu_t.$$

The squared residuals are the quantity of interest since $\text{Var}(\varepsilon_t) = E[\varepsilon_t^2]$ under the assumption that $E[\varepsilon_t] = 0$. The reason that the auxiliary regression takes this form is that it is desirable to investigate whether the variance of the residuals varies systematically with any known variables relevant to the model. Note also that this regression should include a constant term, even if the original regression did not. This is as a result of the fact that $\hat{\varepsilon}_t^2$ will always have a non-zero mean.

(3) Given the auxiliary regression, the test can be conducted using two different approaches.

(i) First it is possible to use the F-test framework. This would involve estimating the auxiliary regression as the unrestricted regression and then running a restricted regression of $\hat{\varepsilon}_t^2$ on a constant only. The RSS from each specification would then be used as inputs to the standard F-test formula.

(ii) An alternative approach, called Lagrange Multiplier (LM) test, centres around the value of R^2 for the auxiliary regression and does not require the estimation of a second (restricted) regression. If one or more coefficients in the auxiliary regression is statistically significant, the value of R^2 for that equation will be relatively high, while if none of the variables is significant, R^2 will be relatively low. The LM test would thus operate by obtaining R^2 from the auxiliary regression and multiplying it by the number of observations, T. It can be shown that

$$TR^2 \sim \chi^2(m)$$

where m is the number of regressors in the auxiliary regression (excluding the constant term), equivalent to the number of restrictions that would have to be placed under the F-test approach.

(4) The test is one of the joint null hypothesis that $\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$. For the LM test, if the χ^2-test statistic from step (3) is greater than the corresponding value from the statistical table then reject the null hypothesis that the errors are homoscedastic.

4.1.3 The Breusch-Pagan test

The **Breusch-Pagan test** can be seen as a special case of White's general test. See [11] for a summary.

4.1.4 The Park test

The **Park test** assumes that the heteroscedasticity may be proportional to some power of an independent variable x_k in the model: $\sigma^2_{\varepsilon_t} = \sigma^2_\varepsilon x^\alpha_{kt}$.

(1) Estimate the model $y_t = \beta_1 + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \varepsilon_t$ using OLS.

(2) Obtain the squared residuals, $\hat{\varepsilon}^2_t$, after estimating your model.

(3) Estimate the model $\ln \hat{\varepsilon}^2_t = \gamma + \alpha \ln x_{kt} + \nu_t$ using OLS.

(4) Examine the statistical significance of α using the t-statistic: $t = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}}$. If the estimate of α coefficient is statistically significant, then you have evidence of heteroskedasticity.

4.2 Consequences of using OLS in the presence of heteroscedasticity

When the errors are heteroscedastic, *the OLS estimators will still give unbiased (and also consistent) coefficient estimates, but they are no longer BLUE.* The reason is that the error variance, σ^2, plays no part in the proof that the OLS estimator is consistent and unbiased, but σ^2 does appear in the formulae for the coefficient variances.

If OLS is still used in the presence of heteroscedasticity, the standard errors could be wrong and hence any inferences made could be misleading. In general, the OLS standard errors will be too large for the intercept when the errors are heteroscedastic. The effect of heteroscedasticity on the slope standard errors will depend on its form.

4.3 Dealing with heteroscedasticity

4.3.1 The generalised least squares method

The **generalised least squares** (GLS) method supposes that the error variance was related to z_t by the expression

$$\mathrm{Var}(\varepsilon_t) = \sigma^2 z_t^2.$$

All that would be required to remove the heteroscedasticity would be to divide the regression equation through by z_t

$$\frac{y_t}{z_t} = \beta_1 \frac{1}{z_t} + \beta_2 \frac{x_{2t}}{z_t} + \beta_3 \frac{x_{3t}}{z_t} + \nu_t$$

where $\nu_t = \frac{\varepsilon_t}{z_t}$ is an error term. GLS can be viewed as OLS applied to transformed data that satisfy the OLS assumptions. GLS is also known as *weighted least squares* (WLS) since under GLS a weighted sum of the squared residuals is minimised, whereas under OLS it is an unweighted sum. Researchers are typically unsure of the exact cause of the heteroscedasticity, and hence this technique is usually infeasible in practice.

4.3.2 Transformation

A second "solution" for heteroscedasticity is transforming the variables into logs or reducing by some other measure of "size". This has the effect of re-scaling the data to "pull in" extreme observations.

4.3.3 The White-corrected standard errors

A third “solution” for heteroscedasticity is **robust standard error** (**White-corrected standard errors**, **heteroscedasticity-corrected standard errors**) following White [9]. In a model with one independent variable, the robust standard error is

$$se(\hat{\beta}_1)_{HC} = \sqrt{\frac{\sum_{t=1}^{T}(x_t - \overline{x})^2 \hat{\varepsilon}_t^2}{\left(\sum_{t=1}^{T}(x_t - \overline{x})^2\right)^2}}.$$

Generalizing this result to a multiple regression model, the robust standard error is

$$se(\hat{\beta}_k)_{HC} = \sqrt{\frac{\sum_{t=1}^{T}\hat{\omega}_{tk}^2 \hat{\varepsilon}_t^2}{\left(\sum_{t=1}^{T}\hat{\omega}_{tk}^2\right)^2}}$$

where the $\hat{\omega}_{tk}^2$'s are the squared residuals obtained from the auxiliary regression of x_k on all the other independent variables. Here's how to calculate robust standard errors:

(1) Estimate your original multivariate model, $y_t = \beta_1 + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \varepsilon_t$, and obtain the squared residuals, $\hat{\varepsilon}_t^2$.

(2) Estimate $K - 1$ auxiliary regressions of each independent variable on all the other independent variables and retain all $T \times (K-1)$ squared residuals ($\hat{\omega}_{tpk}^2$).

(3) For any independent variable, calculate the robust standard errors:

$$se(\hat{\beta}_k)_{HC} = \sqrt{\frac{\sum_{t=1}^{T}\hat{\omega}_{tk}^2 \hat{\varepsilon}_t^2}{\left(\sum_{t=1}^{T}\hat{\omega}_{tk}^2\right)^2}}.$$

The effect of using the correction is that, if the variance of the errors is positively related to the square of an explanatory variable, the standard errors for the slope coefficients are increased relative to the usual OLS standard errors, which would make hypothesis testing more “conservative”, so that more evidence would be required against the null hypothesis before it would be rejected.

The results of Fabozzi and Francis [4] strongly suggest the presence of heteroscedasticity in the context of the single index market model. Numerous versions of robust standard errors exist for the purpose of improving the statistical properties of the heteroskedasticity correction; no form of robust standard error is preferred above all others.

5 Violation of Assumptions: Autocorrelation

5.1 Detection of autocorrelation

When you’re drawing conclusions about autocorrelation using the error pattern, all other CLRM assumptions must hold, especially the assumption that the model is correctly specified. If a model isn’ t correctly specified, you may mistakenly identify the model as suffering from autocorrelation. Misspecification is a more serious issue than autocorrelation.

5.1.1 Graphical test

Graphical test. The first step is to consider possible relationships between the current residual and the immediately previous one via a graphical exploration. Thus $\hat{\varepsilon}_t$ is plotted against $\hat{\varepsilon}_{t-1}$, and $\hat{\varepsilon}_t$ is plotted over time. Under *positive autocorrelation*, most of the dots $(\hat{\varepsilon}_{t-1}, \hat{\varepsilon}_t)$ are in the first and third quadrants, while under *negative autocorrelation*, most of the dots $(\hat{\varepsilon}_{t-1}, \hat{\varepsilon}_t)$ are in the second and fourth quadrants. In a graph where $\hat{\varepsilon}_t$ is plotted against time, a positively autocorrelated series of residuals will not cross the time-axis very frequently, while a negatively autocorrelated series of residuals will cross the time-axis more frequently than if they were distributed randomly.

5.1.2 The run test (the Geary test)

The run test (the Geary test). You want to use the run test if you're uncertain about the nature of the autocorrelation.

A *run* is defined as a sequence of positive or negative residuals. The hypothesis of no autocorrelation isn't sustainable if the residuals have too many or too few runs.

The most common version of the test assumes that runs are distributed normally. If the assumption of no autocorrelation is sustainable, with 95% confidence, the number of runs should be between

$$\mu_r \pm 1.96\sigma_r$$

where μ_r is the expected number of runs and σ_r is the standard deviation. These values are calculated by

$$\mu_r = \frac{2T_1T_2}{T_1+T_2} + 1,\ \sigma_r = \sqrt{\frac{2T_1T_2(2T_1T_2 - T_1 - T_2)}{(T_1+T_2)^2(T_1+T_2-1)}}$$

where r is the number of observed runs, T_1 is the number of positive residuals, T_2 is the number of negative residuals, and T is the total number of observations.

If the number of observed runs is below the expected interval, it's evidence of positive autocorrelation; if the number of runs exceeds the upper bound of the expected interval, it provides evidence of negative autocorrelation.

5.1.3 The Durbin-Watson test

The **Durbin-Watson (*DW*) test** is a test for first order autocorrelation. One way to motivate the test and to interpret the test statistic would be in the context of a regression of the time-t error on its previous value

$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t \tag{6}$$

where $\nu_t \sim N(0, \sigma^2_{nu})$.[3] The DW test statistic has as its null and alternative hypotheses

$$H_0 : \rho = 0,\ H_1 : \rho \neq 0.$$

It is not necessary to run the regression given by (6) since the test statistic can be calculated using quantities that are already available after the first regression has been run

$$DW = \frac{\sum_{t=2}^{T}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=2}^{T}\hat{\varepsilon}_t^2} \approx 2(1-\hat{\rho}) \tag{7}$$

where $\hat{\rho}$ is the estimated correlation coefficient that would have been obtained from an estimation of (6). The intuition of the DW statistic is that the numerator "compares" the values of the error at times $t-1$ and t. If there is positive autocorrelation in the errors, this difference in the numerator will be relatively small, while if there is negative autocorrelation, with the sign of the error changing very frequently, the numerator will be relatively large. No autocorrelation would result in a value for the numerator between small and large.

In order for the DW test to be valid for application, three conditions must be fulfilled:

(i) There must be a constant term in the regression.
(ii) The regressors must be non-stochastic.
(iii) There must be no lags on dependent variable in the regresion.[4]

The DW test does not follow a standard statistical distribution. It has two critical values: an upper critical value d_U and a lower critical value d_L. The rejection and non-rejection regions for the DW test are illustrated in Figure 1.

[3] More generally, the $AR(1)$ processes in time series analysis.

[4] If the test were used in the presence of lags of the dependent variable or otherwise stochastic regressors, the test statistic would be biased towards 2, suggesting that in some instances the null hypothesis of no autocorrelation would not be rejected when it should be.

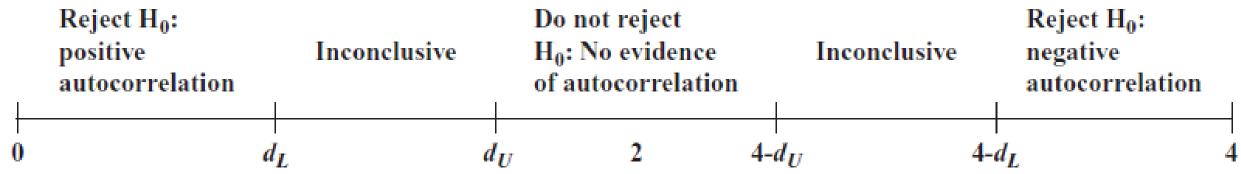
Reject H0: positive autocorrelation
Inconclusive
Do not reject H0: No evidence of autocorrelation
Inconclusive
Reject H0: negative autocorrelation
0
dL
dU
2
4-dU
4-dL
4

Figure 1: Rejection and non-rejection regions for DW test

5.1.4 The Breusch-Godfrey test

The **Breusch-Godfrey test** is a more general test for autocorrelation up to the rth order, whereas DW is a test only of whether consecutive errors are related to one another. The model for the errors under the Breusch-Godfrey test is

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \cdots + \rho_r \varepsilon_{t-r} + \nu_t,\ \nu_t \sim N(0, \sigma_\nu^2). \tag{8}$$

The null and alternative hypotheses are:

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_r = 0,\ H_1 : \rho_1 \neq 0 \text{ or } \rho_2 \neq 0 \text{ or } \cdots \text{ or } \rho_r \neq 0.$$

The Breusch-Godfrey test is carried out as follows:
(1) Estimate the linear regression (8) using OLS and obtain the residuals, $\hat{\varepsilon}_t$.
(2) Obtain R^2 from the auxiliary regression

$$\hat{\varepsilon}_t = \gamma_1 + \sum_{i=2}^{K} \gamma_i x_{it} + \sum_{j=1}^{r} \rho_j \hat{\varepsilon}_{t-j} + \nu_t,\ \nu_t \sim N(0, \sigma_\nu^2).$$

(3) Letting T denote the number of observations, the test statistic is given by

$$(T - r)R^2 \sim \chi_r^2.$$

Note that $(T - r)$ pre-multiplies R^2 in the test for autocorrelation rather than T. This arises because the first r observations will effectively have been lost from the sample in order to obtain the r lags used in the test regression, leaving $(T - r)$ observations from which to estimate the auxiliary regression.

One potential difficulty with Breusch-Godfrey is in determining an appropriate value of r. There is no obvious answer to this, so it is typical to experiment with a range of values, and also to use the frequency of the data to decide. For example, if the data is monthly or quarterly, set r equal to 12 or 4, respectively.

5.2 Consequences of ignoring autocorrelation if it is present

The consequences of ignoring autocorrelation when it is present are similar to those of ignoring heteroscedasticity. The coefficient estimates derived using OLS are still unbiased, but they are inefficient, even at large sample sizes, so that the standard error estimates could be wrong.

In the case of positive serial correlation in the residuals, the OLS standard error estimates will be biased downwards relative to the true standard errors. Furthermore, R^2 is likely to be inflated relative to its "correct" value if autocorrelation is present but ignored, since residual autocorrelation will lead to an underestimate of the true error variance (for positive autocorrelation).

The following example illustrates the statement above. We assume the autocorrelation is represented by a *first-order autocorrelation*:

$$y_t = \beta_1 + \sum_{i=2}^{K} \beta_i x_{it} + \varepsilon_t,\ \varepsilon_t = \rho \varepsilon_{t-1} + \nu_t.$$

where $-1 < \rho < 1$ and ν_t is a random variable that satisfies the CLRM assumptions; namely $E[\nu_t | \varepsilon_{t-1}] = 0$, $\text{Var}(\nu_t | \varepsilon_{t-1}) = \sigma_\nu^2$, and $\text{Cov}(\nu_t, \nu_s) = 0$ for all $t \neq s$. By repeated substitution, we obtain

$$\varepsilon_t = \nu_t + \rho \nu_{t-1} + \rho^2 \nu_{t-2} + \rho^3 \nu_{t-3} + \cdots .$$

Therefore,

$$E[\varepsilon_t] = 0,\ \mathrm{Var}(\varepsilon_t) = \sigma_\nu^2 + \rho^2\sigma_\nu^2 + \cdots = \frac{\sigma_\nu^2}{1-\rho^2}.$$

The stationarity assumption ($|\rho| < 1$) is necessary to constrain the variance from becoming an infinite value. OLS assumes no autocorrelation; that is, $\rho = 0$ in the expression $\sigma_\varepsilon^2 = \frac{\sigma_\nu^2}{1-\rho^2}$. Consequently, in the presence of autocorrelation, the estimated variances and standard errors from OLS are underestimated.

5.3 Dealing with autocorrelation

5.3.1 The Cochrane-Orcutt procedure

The **Cochrane-Orcutt (CO) procedure** works by assuming a particular form for the structure of the autocorrelation, and it is carried out as follows:

(1) Assume that the general model is of the form

$$y_t = \beta_1 + \sum_{i=2}^{K} \beta_i x_{it} + \varepsilon_t$$

$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t. \tag{9}$$

Estimate the equation using OLS, ignoring the residual autocorrelation.

(2) Obtain the residuals, and run the regression

$$\hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + \nu_t.$$

(3) Obtain $\hat{\rho}$ and construct $y_t^* = y_t - \hat{\rho}y_{t-1}$, $\beta_1^* = (1-\hat{\rho})\beta_1$, $x_{2t}^* = (x_{2t} - \hat{\rho}x_{2(t-1)})$, etc., so that the original model can be written as

$$y_t^* = \beta_1^* + \sum_{i=2}^{K} \beta_i x_{it}^* + \nu_t$$

(4) Run the GLS regression

$$y_t^* = \beta_1^* + \sum_{i=2}^{K} \beta_i x_{it}^* + \nu_t$$

Cochrane and Orcutt [2] argue that better estimates can be obtained by repeating steps (2)-(4) until the change in $\hat{\rho}$ between one iteration and the next is less than some fixed amount (e.g. 0.01). In practice, a small number of iterations (no more than 5) will usually suffice. We also note assumptions like (9) should be tested before the Cochrane-Orcutt or similar procedure is implemented.[5]

5.3.2 The Newey-West standard errors

The **Newey-West (NW) standard errors**. Estimating the model using OLS and adjusting the standard errors for autocorrelation has become more popular than other correction methods. There are two reasons for this: (1) The serial correlation robust standard errors can adjust the results in the presence of a basic $AR(1)$ process or a more complex $AR(q)$ process, and (2) only the biased portion of the results (the standard errors) are adjusted, while the unbiased estimates (the coefficients) are untouched, so no model transformation is required.

The White variance-covariance matrix of the coefficients is appropriate when the residuals of the estimated equation are heteroscedastic but serially uncorrelated. Newey and West [7] developed a variance-covariance estimator that is consistent in the presence of both heteroscedasticity and autocorrelation. The corresponding standard error is called the *heteroscedasticity-autocorrelation-corrected (HAC) standard error*, the *serial correlation robust standard error*, or the *Newey-West (NW) standard error*. It can be calculated by applying the following steps:

[5] For a similar procedure called the **Prais-Winsten (PW) transformation**, see [11].

(1) Estimate your original model $y_t = \beta_1 + \sum_{i=2}^{K} \beta_i x_{it} + \varepsilon_t$ and obtain the residuals: $\hat{\varepsilon}_t$.

(2) Estimate the auxiliary regression $x_{2t} = \alpha_1 + \sum_{i=3}^{K} \alpha_i x_{it} + r_t$ and retain the residuals: $\hat{r}_t$.

(3) Find the intermediate adjustment factor, $\hat{\alpha}_t = \hat{r}_t \hat{\varepsilon}_t$, and decide how much serial correlation (the number of lags) you're going to allow. A Breusch-Godfrey test can be useful in making this determination, while EViews uses INTEGER$[4(T/100)^{2/9}]$.

(4) Obtain the error variance adjustment factor, $\hat{v} = \sum_{t=1}^{T} \hat{\alpha}_t^2 + 2\sum_{h=1}^{g} \left[1 - \frac{h}{g+1}\right] \left(\sum_{t=h+1}^{T} \hat{\alpha}_t \hat{\alpha}_{t-h}\right)$, where g represents the number of lags determined in Step 3.

(5) Calculate the serial correlation robust standard error. For variable x_2,

$$se(\hat{\beta}_2)_{HAC} = \left(\frac{se(\hat{\beta}_2)}{\hat{\sigma}_\varepsilon}\right)^2 \sqrt{\hat{v}}.$$

(6) Repeat Steps (2) through (5) for independent variables x_3 through x_K.

5.3.3 Dynamic models

Dynamic models. In practice, assumptions like (9) are likely to be invalid and serial correlation in the errors may arise as a consequence of "misspecified dynamics". Therefore a dynamic model that allows for the structure of y should be used rather than a residual correction on a static model that only allows for a *contemporaneous relationship* between the variables.

5.3.4 First difference

First differences. Another potential "remedy" for autocorrelated residuals would be to switch to a model in first differences rather than in levels.

5.4 Miscellaneous issues

Why might lags be required in a regression? Lagged values of the explanatory variables or of the dependent variable (or both) may capture important dynamic structure in the dependent variable. Two possibilities that are relevant in fiance are as follows.

- *Inertia of the dependent variable.* Often a change in the value of one of the explanatory variables will not affect the dependent variable immediately during one time period, but rather with a lag over several time periods. Many variables in economics and finance will change only slowly as a result of pure psychological factors. Delays in response may also arise as a result of technological or institutional factors.
- *Over-reactions.*

Autocorrelation that would not be remedied by adding lagged variables to the model.

- *Omission of relevant variables, which are themselves autocorrelated*, will induce the residuals from the estimated model to be serially correlated.
- *Autocorrelation owing to unparameterised seasonality.*
- *If "misspecification" error has been committed by using an inappropriate functional form.*

Problems with adding lagged regressors to "cure" autocorrelation.

- *Inclusion of lagged values of the dependent variable violates the assumption that the explanatory variables are non-stochastic.* In small samples, inclusion of lags of the dependent variable can lead to biased coefficient estimates, although they are still consistent.
- *A model with many lags may have solved a statistical problem (autocorrelated residuals) at the expense of creating an interpretational one.*

Note that if there is still autocorrelation in the residuals of a model including lags, then the OLS estimators will not even be consistent.

Autocorrelation in cross-sectional data. Autocorrelation in the context of a time series regression is quite intuitive. However, it is also plausible that autocorrelation could be present in certain types of cross-sectional data.

6 Violation of Assumptions: Non-Stochastic Regressors

The OLS estimator is consistent and unbiased in the presence of stochastic regressors, provided that the regressors are not correlated with the error term of the estimated equation. However, if one or more of the explanatory variables is contemporaneously correlated with the disturbance term, the OLS estimator will not even be consistent. This results from the estimator assigning explanatory power to the variables where in reality it is arising from the correlation between the error term and y_t.

7 Violation of Assumptions: Non-Normality of the Disturbances

The **Bera-Jarque test** statistic is given by

$$W = T\left[\frac{b_1^2}{6} + \frac{(b_2-3)^2}{24}\right]$$

where T is the sample size, b_1 is the coefficient of skewness

$$b_1 = \frac{E[\varepsilon^3]}{(\sigma^2)^{3/2}},$$

and b_2 is the coefficient of kurtosis

$$b_2 = \frac{E[\varepsilon^4]}{(\sigma^2)^2}.$$

The test statistic W asymptotically follows a $\chi^2(2)$ under the null hypothesis that the distribution of the series is symmetric and mesokurtic, properties that a standard normal distribution has.

b_1 and b_2 can be estimated using the residuals from the OLS regression $\hat{\varepsilon}$. The null hypothesis is of normality, and this would be rejected if the residuals from the model were either significantly skewed or leptokurtic/platykurtic (or both).

8 Issues of Model Specification

8.1 Omission of an important variable

The consequence would be that the estimated coefficients on all the other variables will be biased and inconsistent unless the excluded variable is uncorrelated with all the included variables. Even if this condition is satisfied, the estimate of the coefficient on the constant term will be biased, and the standard errors will also be biased (upwards). Further information is offered in Dougherty [3], Chapter 7.

8.2 Inclusion of an irrelevant variable

The consequence of including an irrelevant variable would be that the coefficient estimators would still be consistent and unbiased, but the estimators would be inefficient. This would imply that the standard errors for the coefficients are likely to be inflated. Variables which would otherwise have been marginally significant may no longer be so in the presence of irrelevant variables. In general, it can also be stated that the extent of the loss of efficiency will depend positively on the absolute value of the correlation between the included irrelevant variable and the other explanatory variables.

When trying to determine whether to err on the side of including too many or too few variables in a regression model, there is an implicit trade-off between inconsistency and efficiency; many researchers would argue that while in an ideal world, the model will incorporate precisely the correct variables – no more and no less – the former problem is more serious than the latter and therefore in the real world, one should err on the side of incorporating marginally significant variables.

8.3 Functional form: Ramsey's RESET

Ramsey's **regression specification error test** (RESET) is conducted by adding a quartic function of the fitted values of the dependent variable ($\hat{y}_t^2$, $\hat{y}_t^3$, and $\hat{y}_t^4$) to the original regression and then testing the joint significance of the coefficients for the added variables.

The logic of using a quartic of your fitted values is that they serve as proxies for variables that may have been omitted – higher order powers of the fitted values of y can capture a variety of non-linear relationships, since they embody higher order powers and cross-products of the original explanatory variables.

The test consists of the following steps:

1. Estimate the model you want to test for specification error. E.g. $y_t = \beta_1 + \beta_2 x_{1t} + \cdots + \beta_K x_{Kt} + \varepsilon_t$.

2. Obtain the fitted values after estimating your model and estimate :

$$y_t = \alpha_1 + \alpha_2 \hat{y}_t^2 + \cdots + \alpha_p \hat{y}_t^p + \sum_{i=1}^{K} \beta_i x_{it} + \nu_t. \tag{10}$$

3. Test the joint significance of the coefficients on the fitted values of y_t terms using an F-statistic, or using the test statistic TR^2, which is distributed asymptotically as $\chi^2(p-1)$ (the value of R^2 is obtained from the regression (10)). If the value of the test statistic is greater than the critical value, reject the null hypothesis that the functional form was correct.

A RESET allows you to identify whether misspecification is a serious problem with your model, but it doesn't allow you to determine the source.

8.4 Parameter stability / structural stability tests

The idea is essentially to split the data into sub-periods and then to estimate up to three models, for each of the sub-parts and for all the data and then to "compare" the RSS of each of the models. There are two types of test: the Chow (analysis of variance) test and predictive failure tests.

8.4.1 The Chow test

To apply a **Chow test** for structural stability between any two groups (A and B):

1. Estimate your model combining all data and obtain the residual sum of squares (RSS_r) with degrees of freedom $T - 2K$.

2. Estimate your model separately for each group and obtain the residual sum of squares for group A, $RSS_{ur,A}$, with degrees of freedom $T_A - K$ and the residual sum of squares for group B, $RSS_{ur,B}$, with degrees of freedom $T_B - K$.

3. Compute the F-statistic by using this formula:

$$F = \frac{\frac{RSS_r - (RSS_{ur,A} + RSS_{ur,B})}{K}}{\frac{RSS_{ur,A} + RSS_{ur,B}}{T-2K}}.$$

The null hypothesis for the Chow test is structural stability. The larger the F-statistic, the more evidence you have against structural stability and the more likely the coefficients are to vary from group to group. If the value of the test statistic is greater than the critical value from the F-distribution, which is an $F(K, T-2K)$, then reject the null hypothesis that the parameters are stable over time.

Note the result of the F-statistic for the Chow test assumes homoskedasticity. A large F-statistic only informs you that the parameters vary between the groups, but it doesn't tell you which specific parameter(s) is (are) the source(s) of the structural break.

8.4.2 Predictive failure tests

A problem with the Chow test is that it is necessary to have enough data to do the regression on both sub-samples. An alternative formulation of a test for the stability of the model is the **predictive failure test**, which requires estimation for the full sample and one of the sub-samples only. The predictive failure test works by estimating the regression over a "long" sub-period (i.e. most of the data) and then using those coefficient estimates for predicting values of y for the other period. These predictions for y are then implicitly compared with the actual values. The null hypothesis for this test is that the prediction errors for all of the forecasted observations are zero.

To calculate the test:

1. Run the regression for the whole period (the restricted regression) and obtain the RSS.

2. Run the regression for the "large" sub-period and obtain the RSS (called RSS_1). Note the number of observations for the long estimation sub-period will be denoted by T_1. The test statistic is given by

$$\frac{RSS - RSS_1}{RSS} \times \frac{T_1 - K}{T_2}$$

where T_2 is the number of observations that the model is attempting to "predict". The test statistic will follow an $F(T_2, T_1{-}K)$ distribution.

Forward predictive failure tests are where the last few observations are kept back for forecast testing. *Backward predictive failure tests* attempt to "back-cast" the first few observations. Both types of test offer further evidence on the stability of the regression relationship over the whole sample period

8.4.3 The Quandt likelihood ratio (QLR) test

The Chow and predictive failure tests will work satisfactorily if the date of a structural break in a financial time series can be specified. But more often, a researcher will not know the break date in advance. In such circumstances, a modified version of the Chow test, known as the **Quandt likelihood ratio (QLR) test**, can be used instead. The test works by automatically computing the usual Chow F-test statistic repeatedly with different break dates, then the break date giving the largest F-statistic value is chosen.

While the test statistic is of the F-variety, it will follow a non-standard distribution rather than an F-distribution since we are selecting the largest from a number of F-statistics rather than examining a single one. The test is well behaved only when the range of possible break dates is sufficiently far from the end points of the whole sample, so it is usual to "trim" the sample by (typically) 5% at each end.

8.4.4 Recursive least squares (RLS): CUSUM and CUSUMQ

An alternative to the QLR test for use in the situation where a researcher is unsure of the date is to perform **recursive least squares (RLS)**. The procedure is appropriate only for time-series data or cross-sectional data that have been ordered in some sensible way (e.g., a sample of annual stock returns, ordered by market capitalisation).

Recursive estimation simply involves starting with a sub-sample of the data, estimating the regression, then sequentially adding one observation at a time and re-running the regression until the end of the sample is reached.

It is to be expected that the parameter estimates produced near the start of the recursive procedure will appear rather unstable, but the key question is whether they then gradually settle down or whether the volatility continues through the whole sample. Seeing the latter would be an indication of parameter instability.

It should be evident that RLS in itself is not a statistical test for parameter stability as such, but rather it provides qualitative information which can be plotted and thus gives a very visual impression of how stable the parameters appear to be.

The **CUSUM test** is based on a normalised (i.e. scaled) version of the cumulative sums of the residuals. Under the null hypothesis of perfect parameter stability, the CUSUM statistic is zero. A set of ±2 standard error bands is usually plotted around zero and any statistic lying outside the bands is taken as evidence of parameter instability.

The **CUSUMSQ test** is based on a normalised version of the cumulative sums of squared residuals. The scaling is such that under the null hypothesis of parameter stability, the CUSUMSQ statistic will start at zero and end the sample with a value of 1. Again, a set of ±2 standard error bands is usually plotted around zero and any statistic lying outside these is taken as evidence of parameter instability.

For full technical details of CUSUM and CUSUMQ, see Greene [5], Chapter 7.

9 The Generalized Linear Regression Model (GLRM)

The **generalized linear regression model** is

$$\begin{cases} \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ E[\boldsymbol{\varepsilon}|\boldsymbol{X}] = \boldsymbol{0} \\ E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}] = \sigma^2\boldsymbol{\Omega} = \boldsymbol{\Sigma}, \end{cases} \tag{11}$$

where $\boldsymbol{\Omega}$ is a positive definite matrix.

In this model, heteroscedasticity usually arises in volatile high frequency time-series data and in cross-section data where the scale of the dependent variable and the explanatory power of the model tend to vary across observations. Autocorrelation is usually found in time-series data. Panel data sets, consisting of cross sections observed at several points in time, may exhibit both characteristics.

Convention on notation: Throughout this section, we shall use n in place of T to stand for the number of observations. This is to be consistent with popular textbooks like Greene [5].

9.1 Properties of OLS in the GLRM

Recall under the assumptions of CLRM, the OLS estimator

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

is the best linear unbiased estimator (BLUE), consistent and asymptotically normally distributed (CAN), and if the disturbances are normally distributed, asymptotically efficient among all CAN estimators.

In the GLRM, the OLS estimator remains unbiased, consistent, and asymptotically normally distributed. It will, however, no longer be efficient and the usual inference procedures based on the F and t distributions are no longer appropriate.

Theorem 1 (**Finite Sample Properties of $\hat{\boldsymbol{\beta}}$ in the GLRM).** *If the regressors and disturbances are uncorrelated, then the least squares estimator is unbiased in the generalized linear regression model. With non-stochastic regressors, or conditional on $\boldsymbol{X}$, the sampling variance of the least squares estimator is*

$$\begin{aligned} Var[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] &= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'|\boldsymbol{X}] \\ &= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{\Omega})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \frac{\sigma^2}{n}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{\Phi}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \end{aligned} \tag{12}$$

where $\Phi = \frac{\sigma^2}{n}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X} = Cov\left(\frac{1}{\sqrt{n}}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\beta)\right)$ *is essentially the covariance matrix of the scores or estimating functions. If the regressors are stochastic, then the unconditional variance is* $E_{\boldsymbol{X}}[Var[\boldsymbol{b}|\boldsymbol{X}]]$. $\hat{\boldsymbol{\beta}}$ *is a linear function of* $\boldsymbol{\varepsilon}$. *Therefore, if* $\boldsymbol{\varepsilon}$ *is normally distributed, then*

$$\hat{\boldsymbol{\beta}}|\boldsymbol{X} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

If $\mathrm{Var}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$ converges to zero, then $\hat{\boldsymbol{\beta}}$ is mean square consistent. With well-behaved regressors, $(\boldsymbol{X}'\boldsymbol{X}/n)^{-1}$ will converge to a constant matrix. But $(\sigma^2/n)(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}/n)$ need not converge at all.

Theorem 2 (**Consistency of OLS in the GLRM).** *If* $\boldsymbol{Q} = p\lim(\boldsymbol{X}'\boldsymbol{X}/n)$ *and* $p\lim(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}/n)$ *are both finite positive definite matrices, then* $\hat{\boldsymbol{\beta}}$ *is consistent for* $\boldsymbol{\beta}$. *Under the assumed conditions,*

$$p\lim\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}.$$

The conditions in the above theorem depend on both $\boldsymbol{X}$ and $\boldsymbol{\Omega}$. An alternative formula that separates the two components can be found in Greene [5, page 194-195].

Theorem 3 (**Asymptotic Distribution of $\hat{\boldsymbol{\beta}}$ in the GLRM).** *If the regressors are sufficiently well behaved and the off-diagonal terms in $\boldsymbol{\Omega}$ diminish sufficiently rapidly, then the least squares estimator is asymptotically normally distributed with covariance matrix*

$$Asy.\,Var[\hat{\boldsymbol{\beta}}] = \frac{\sigma^2}{n}\boldsymbol{Q}^{-1} p\lim\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}\right)\boldsymbol{Q}^{-1}.$$

9.2 Robust estimation of asymptotic covariance matrices for OLS

In view of formula (12), is it necessary to discard OLS as an estimator? If $\boldsymbol{\Omega}$ is known, then as will be shown later, there is a simple and efficient estimator based on $\boldsymbol{\Omega}$, and the answer is yes. If $\boldsymbol{\Omega}$ is unknown but its structure is known and we can estimate $\boldsymbol{\Omega}$ using sample information, then the answer is less clear-cut. The third possibility is that $\boldsymbol{\Omega}$ is completely unknown, both as to its structure and the specific values of its elements. In this situation, least squares or instrumental variables may be the only estimator available, and as such, the only available strategy is to try to devise an estimator for the appropriate asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$.

If $\sigma^2\boldsymbol{\Omega}$ were known, then the *estimator* of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ would be

$$\boldsymbol{V}_{OLS} = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'[\sigma^2\boldsymbol{\Omega}]\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

The matrices of sums of squares and cross products in the left and right matrices are sample data that are readily estimable, and the problem is the center matrix that involves the unknown $\sigma^2\boldsymbol{\Omega} = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}]$. For estimation purposes, we will assume that $\mathrm{tr}(\boldsymbol{\Omega}) = n$, as it is when $\sigma^2\boldsymbol{\Omega} = \sigma^2\boldsymbol{I}$ in the CLRM.

Let $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j} = \sigma^2\boldsymbol{\Omega} = \sigma^2(\omega_{ij})_{i,j}$. What is required is an estimator of the $K(K+1)/2$ unknown elements in the matrix

$$\boldsymbol{Q}_* = \frac{1}{n}\boldsymbol{X}'\boldsymbol{\Sigma}\boldsymbol{X} = \frac{1}{n}\sum_{i,j=1}^{n}\sigma_{ij}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_j'.$$

where $\tilde{\boldsymbol{x}}_i$ is the column vector formed by the transpose of row i of $\boldsymbol{X}$ (see Greene [5, page 805]). To verify this formula of $\boldsymbol{Q}_*$, recall we have the convention

$$\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K],\ \boldsymbol{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix}$$

with $i = 1, \cdots, K$. So

$$\tilde{\boldsymbol{x}}_i = [x_{1i}, x_{2i}, \cdots, x_{Ki}]' \text{ and } \tilde{\boldsymbol{x}}_j' = [x_{1j}, x_{2j}, \cdots, x_{Kj}],\ 1 \le i, j \le n.$$

Consequently,

$$\boldsymbol{X} = \begin{bmatrix} \tilde{\boldsymbol{x}}_1' \\ \tilde{\boldsymbol{x}}_2' \\ \vdots \\ \tilde{\boldsymbol{x}}_n' \end{bmatrix},\ \boldsymbol{X}' = [\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_n], \text{ and } \boldsymbol{X}'\boldsymbol{\Sigma}\boldsymbol{X} = \sum_{i,j=1}^{n}\sigma_{ij}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_j'$$

The least squares estimator $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, which implies that the least squares residuals $\hat{\varepsilon}_i$ are "pointwise" consistent estimators of their population counterparts ε_i. The general approach, then, will be to use $\boldsymbol{X}$ and $\hat{\boldsymbol{\varepsilon}}$ to devise an estimator of $\boldsymbol{Q}_*$.

9.2.1 HC estimator

Consider the heteroscedasticity case first. White [9] has shown that under very general conditions, the estimator

$$\boldsymbol{S}_0 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i'$$

has

$$p\lim \boldsymbol{S}_0 = p\lim \boldsymbol{Q}_*.$$

Therefore, the **White heteroscedasticity consistent (HC) estimator**

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X'X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i'\right)\left(\frac{1}{n}\boldsymbol{X'X}\right)^{-1} = n\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{S_0}\left(\boldsymbol{X'X}\right)^{-1}.$$

can be used to estimate the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$.

This result is extremely important and useful. It implies that without actually specifying the type of heteroscedasticity, we can still make appropriate inferences based on the results of least squares. This implication is especially useful if we are unsure of the precise nature of the heteroscedasticity.

9.2.2 HAC estimator

In the presence of both heteroscedasticity and autocorrelation, as a natural extension of White's result, the natural counterpart for estimating

$$\boldsymbol{Q}_* = \frac{1}{n}\sum_{i,j=1}^{n} \sigma_{ij} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_j'$$

would be

$$\hat{\boldsymbol{Q}}_* = \frac{1}{n}\sum_{i,j=1}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_j \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_j'$$

But there are two problems with this estimator. The first one is that it is difficult to conclude yet that $\hat{\boldsymbol{Q}}_*$ will converge to anything at all, since the matrix is $1/n$ times a sum of n^2 terms. We can achieve the convergence of $\hat{\boldsymbol{Q}}_*$ by assuming that the rows of $\boldsymbol{X}$ are well behaved and that the correlations diminish with increasing separation in time.

The second problem is a practical one, that $\hat{\boldsymbol{Q}}_*$ needs not be positive definite. Newey and West [7] have devised an estimator, the **Newey–West autocorrelation consistent (AC) covariance estimator**, that overcomes this difficulty:

$$\hat{\boldsymbol{Q}}_* = \boldsymbol{S}_0 + \frac{1}{n}\sum_{l=1}^{L}\sum_{t=l+1}^{n} w_l \hat{\varepsilon}_t \hat{\varepsilon}_{t-l}(\tilde{\boldsymbol{x}}_t \tilde{\boldsymbol{x}}_{t-l}' + \tilde{\boldsymbol{x}}_{t-l}\tilde{\boldsymbol{x}}_t'),\ w_l = 1 - \frac{l}{L+1}.$$

It must be determined in advance how large L is to be. In general, there is little theoretical guidance. Current practice specifies $L \approx T^{1/4}$. Unfortunately, the result is not quite as crisp as that for the heteroscedasticity consistent estimator.

9.2.3 R package for robust covariance estimation of OLS

See Zeileis [10] for a survey.

References

[1] Chris Brooks. *Introductory econometrics for finance*, 2ed.. New York, Cambridge University Press, 2008. 1, 3, 4

[2] Cochrane, D. and Orcutt, G. H. (1949). "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms", *Journal of the American Statistical Association* 44, 32–61. 12

[3] Christopher Dougherty. *Introduction to econometrics*, 3ed.. Oxford University Press, 2007. 14

[4] Fabozzi, F. J. and Francis, J. C. (1980). "Heteroscedasticity in the Single Index Model", *Journal of Economics and Business* 32, 243–8. 9

[5] William H. Greene. *Econometric analysis*, 5ed.. Prentice Hall, 2002. 1, 16, 17, 18

[6] William H. Greene. *Econometric analysis*, 7ed.. Prentice Hall, 2012. 1, 3

[7] Newey, W. K. and West, K. D. (1987). "A Simple Positive-Definite Heteroskedasticity and Autocorrelation-Consistent Covariance Matrix", *Econometrica* 55, 703–8. 12, 19

[8] Roberto Pedace. *Econometrics for dummies.* Hoboken, John Wiley & Sons Inc., 2013. 1

[9] White, H. (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity", *Econometrica* 48, 817–38. 9, 19

[10] Zeileis, A. (2004). "Econometric Computing with HC and HAC Covariance Matrix Estimators", *Journal of Statistical Software*, **11**:10. 1, 19

[11] Zeng, Y. (2016). "Book Summary: *Econometrics for Dummies*", version 1.0.5. Unpublished manuscript. 8, 12