# PE I: Multivariable Regression

## Generalized Least Squares, Heteroskedastic and Autocorrelated Errors (Chapters 4.6, 4.7 & 4.8)

Andrius Buteikis, andrius.buteikis@mif.vu.lt
http://web.vu.lt/mif/a.buteikis/

# Multiple Regression: Model Assumptions

Much like in the case of the univariate regression with one independent variable, the multiple regression model has a number of required assumptions:

**(MR.1): Linear Model** The Data Generating Process (**DGP**), or in other words, the population, is described by a linear (*in terms of the coefficients*) model:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon \tag{MR.1}$$

**(MR.2): Strict Exogeneity** Conditional expectation of $\varepsilon$, given all observations of the explanatory variable matrix $\boldsymbol{X}$, is zero:

$$\mathbb{E}\left(\varepsilon|\boldsymbol{X}\right) = \mathbf{0} \tag{MR.2}$$

This assumption also implies that $\mathbb{E}(\varepsilon) = \mathbb{E}\left(\mathbb{E}(\varepsilon|\boldsymbol{X})\right) = \mathbf{0}$, $\mathbb{E}(\varepsilon\boldsymbol{X}) = \mathbf{0}$ and $\mathbb{C}\mathrm{ov}(\varepsilon, \boldsymbol{X}) = \mathbf{0}$. Furthermore, this property implies that: $\mathbb{E}\left(\boldsymbol{Y}|\boldsymbol{X}\right) = \boldsymbol{X}\beta$

**(MR.3): Conditional Homoskedasticity** The variance-covariance matrix of the error term, conditional on $\boldsymbol{X}$ is constant:

$$\mathbb{V}\text{ar}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right) = \begin{bmatrix} \mathbb{V}\text{ar}(\epsilon_1) & \mathbb{C}\text{ov}(\epsilon_1, \epsilon_2) & ... & \mathbb{C}\text{ov}(\epsilon_1, \epsilon_N) \\ \mathbb{C}\text{ov}(\epsilon_2, \epsilon_1) & \mathbb{V}\text{ar}(\epsilon_2) & ... & \mathbb{C}\text{ov}(\epsilon_2, \epsilon_N) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\text{ov}(\epsilon_N, \epsilon_1) & \mathbb{C}\text{ov}(\epsilon_N, \epsilon_2) & ... & \mathbb{V}\text{ar}(\epsilon_N) \end{bmatrix} = \sigma_\epsilon^2 \boldsymbol{I} \qquad \text{(MR.3)}$$

**(MR.4): Conditionally Uncorrelated Errors** The covariance between different error term pairs, conditional on $\boldsymbol{X}$, is zero:

$$\mathbb{C}\text{ov}\left(\epsilon_i, \epsilon_j|\boldsymbol{X}\right) = 0, \quad i \neq j \qquad \text{(MR.4)}$$

This assumption implies that all error pairs are uncorrelated. For cross-sectional data, this assumption implies that there is no spatial correlation between errors.

**(MR.5)** There exists no exact linear relationship between the explanatory variables. This means that:

$$c_1 X_{i1} + c_2 X_{i2} + ... + c_k X_{ik} = 0, \ \forall i = 1, ..., N \iff c_1 = c_2 = ... = c_k = 0 \quad \text{(MR.5)}$$

This assumption is violated if there exists some $c_j \neq 0$.
Alternatively, this requirement means that:

$$\text{rank}\,(\boldsymbol{X}) = k + 1$$

or, alternatively, that:

$$\det\left(\boldsymbol{X}^\top \boldsymbol{X}\right) \neq 0$$

This assumption is important, because a linear relationship between independent variables means that we cannot separately estimate the effects of changes in each variable separately.

**(MR.6) (optional)** The residuals are normally distributed:

$$\varepsilon | \boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}\right) \quad \text{(MR.6)}$$

# Generalized Least Squares

# Generalized Least Squares

Let our multiple regression be defined as:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

So far, one of our assumptions about the error term was defined by (MR.3) - (MR.4), namely that:

$$\mathbb{V}\mathrm{ar}\left(\varepsilon|\boldsymbol{X}\right) = \mathbb{E}\left(\varepsilon\varepsilon^{\top}\right) = \sigma_{\epsilon}^{2}\boldsymbol{I}$$

However, sometimes the variance-covariance matrix of the residuals $\mathbb{V}\mathrm{ar}\left(\varepsilon|\boldsymbol{X}\right) \neq \sigma_{\epsilon}^{2}\boldsymbol{I}$.

In this lecture we will examine how this change effects parameter estimation, as well as the possible solutions for such cases.

## A General Multiple Linear Regression

Consider the following model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right) = \mathbf{0}, \quad \mathbb{V}\mathrm{ar}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right) = \mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\top}\right) = \boldsymbol{\Sigma} = \sigma_{\epsilon}^{2}\boldsymbol{\Omega}$$

where $\boldsymbol{\Omega}$ is symmetric and positive definite $N \times N$ matrix. This model allows for the errors to be **heteroskedastic** or **autocorrelated** (or both) and is often referred to as **special case of** the **generalized linear (regression) model (GLM)**.
Consequently, we may refer to this type of a models as a **general multiple linear regression** (GMLR), to distinguish it as only a special case of a GLM. We will examine GLM's in more detail in a later lecture.

- If $\boldsymbol{\Omega} = \boldsymbol{I}$, then the GMLR is just the simple multiple linear regression model that we are already familiar with;
- If $\boldsymbol{\Omega}$ is diagonal with non-constant diagonal elements, then the error terms are **uncorrelated** but they are **heteroskedastic**;
- If $\boldsymbol{\Omega}$ is **not diagonal** then $\mathbb{C}\text{ov}(\epsilon_i, \epsilon_j) = \Omega_{i,j} \neq 0$ for some $i \neq j$. In econometrics, non-diagonal covariance matrices are most commonly encountered in **time-series** data, with higher correlations for observations closer in time (usually when $i$ and $j$ are differ by 1 or 2).
  - If $\boldsymbol{\Omega}$ is **not diagonal** and the diagonal elements are **constant**, then the errors are autocorrelated and homoskedastic;
  - If $\boldsymbol{\Omega}$ is **not diagonal** and the diagonal elements are **not constant**, then the errors are autocorrelated and heteroskedastic;

As such, we will now assume that $\boldsymbol{\Omega} \neq \boldsymbol{I}$, since we have already covered this case in previous chapters.

## Consequences when OLS is used on GMLR

Since the OLS estimator can be expressed as:

$$\widehat{\boldsymbol{\beta}}_{OLS} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

then:

- The expected value of the OLS estimator:

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = \boldsymbol{\beta} + \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \mathbb{E}\left(\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta}$$

  i.e. the OLS estimators of $\boldsymbol{\beta}$ are **unbiased**;
- The variance-covariance matrix of the OLS estimator:

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) &= \mathbb{E}\left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right] \\ &= \sigma_\epsilon^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \neq \sigma_\epsilon^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\end{aligned}$$

- Since $\mathbb{V}\mathrm{ar}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) \neq \sigma_\epsilon^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$, statistical inference based on the following assumptions:

$$\begin{cases} \widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = \widehat{\sigma}^2_{OLS} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \\ \widehat{\sigma}^2_{OLS} = \dfrac{\widehat{\boldsymbol{\varepsilon}}^\top_{OLS} \widehat{\boldsymbol{\varepsilon}}_{OLS}}{N-(k+1)} = \dfrac{1}{N-(k+1)} \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{OLS}\right)^\top \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{OLS}\right) \end{cases}$$

  are **invalid** since, in general, $\widehat{\sigma}^2_{OLS}$ is a **biased** and **inconsistent** estimator of $\sigma^2$ for GMLR. Consequently $\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right)$ are also biased and inconsistent.
- The usual OLS $t$-statistics **do not** have Students $t$ distribution, even in large samples;
- The $F$-statistic no longer has Fisher distribution; the $LM$ statistic (see the previously described heteroskedasticity tests) no longer has an asymptotic $\chi^2$ distribution.
- $\widehat{\boldsymbol{\beta}}_{OLS} \to \boldsymbol{\beta}$, if the largest eigenvalue of $\boldsymbol{\Omega}$ is bounded for all $N$, and the largest eigenvalue of $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$ tends to zero as $N \to \infty$.

In other words, the OLS estimates are **unbiased** and **consistent**, but **their variance estimators are biased and inconsistent**, which leads to incorrect results for statistical tests.

## Generalized Least Squares (GLS)

The general idea behind GLS is that in order to obtain an efficient estimator of $\widehat{\beta}$, we need to transform the model, so that the transformed model satisfies the Gauss-Markov theorem (which is defined by our (MR.1) - (MR.5) assumptions).
Then, estimating the transformed model by OLS yields efficient estimates. The transformation is expressed in terms of a (usually) *triangular* matrix $\boldsymbol{\Psi}$, such that:

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}$$

Then, premultiplying our initial model via this matrix yields:

$$\boldsymbol{\Psi}^{\top}\boldsymbol{Y} = \boldsymbol{\Psi}^{\top}\boldsymbol{X}\beta + \boldsymbol{\Psi}^{\top}\varepsilon \tag{1}$$

Because $\boldsymbol{\Omega}$ is non-singular - so is $\boldsymbol{\Psi}$, therefore we can always multiply the previous expression by $\left(\boldsymbol{\Psi}^{\top}\right)^{-1}$ to arrive back at the initial model.
In other words - we have simply scaled both sides of the initial equation.

Then, the following properties hold for the residuals:

$$\mathbb{E}\left(\boldsymbol{\Psi}^\top \boldsymbol{\varepsilon}\right) = \boldsymbol{\Psi}^\top \mathbb{E}\left(\boldsymbol{\varepsilon}\right) = \mathbf{0}$$

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left(\boldsymbol{\Psi}^\top \boldsymbol{\varepsilon}\right) &= \mathbb{V}\mathrm{ar}\left(\boldsymbol{\Psi}^\top \boldsymbol{\varepsilon}\right) = \boldsymbol{\Psi}^\top \mathbb{V}\mathrm{ar}\left(\boldsymbol{\varepsilon}\right) \boldsymbol{\Psi} \\
&= \sigma^2 \boldsymbol{\Psi}^\top \boldsymbol{\Omega} \boldsymbol{\Psi} = \sigma^2 \boldsymbol{\Psi}^\top \left(\boldsymbol{\Omega}^{-1}\right)^{-1} \boldsymbol{\Psi} \\
&= \sigma^2 \boldsymbol{\Psi}^\top \left(\boldsymbol{\Psi} \boldsymbol{\Psi}^\top\right)^{-1} \boldsymbol{\Psi} = \sigma^2 \boldsymbol{\Psi}^\top \left(\boldsymbol{\Psi}^\top\right)^{-1} \boldsymbol{\Psi}^{-1} \boldsymbol{\Psi} \\
&= \sigma^2 \boldsymbol{I}
\end{aligned}$$

which means that now the assumptions (MR.3) - (MR.4) hold true for the model, defined in (1).

## The Parameter GLS Estimator

Consequently, the OLS estimator of $\beta$ from regression in (1) is:

$$\widehat{\beta} = \left(\boldsymbol{X}^\top \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \boldsymbol{Y} = \left(\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{Y}$$

This estimator is called the **generalized least squares (GLS)** estimator of $\beta$.
So, the GLS estimator can be formulated as follows:

Let $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$ and let $\boldsymbol{Y}^* = \boldsymbol{X}^*\beta + \varepsilon^*$, where $\boldsymbol{Y}^* = \boldsymbol{\Psi}^\top \boldsymbol{Y}$, $\boldsymbol{X}^* = \boldsymbol{\Psi}^\top \boldsymbol{X}$ and $\varepsilon^* = \boldsymbol{\Psi}^\top \varepsilon$. then the **GLS** estimator of $\beta$ is:

$$\widehat{\beta}_{GLS} = \left(\boldsymbol{X}^{*\top} \boldsymbol{X}^*\right)^{-1} \boldsymbol{X}^{*\top} \boldsymbol{Y}^* = \left(\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{Y}$$

Alternatively, it can also be obtained as a solution to the minimization of the **GLS criterion function**:

$$(\boldsymbol{Y} - \boldsymbol{X}\beta)^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\beta) \to \min_{\beta}$$

This criterion function can be thought of as a generalization of the *RSS* function, which is minimized in the OLS case. The effect of such **weighting** is clear when $\boldsymbol{\Omega}$ is diagonal - each observation is simply given a weight proportional to the inverse of the variance of its error term.

## The Error Variance Estimator

Next, we need to estimate the error variance:

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{N-(k+1)}\left(\boldsymbol{Y}^* - \boldsymbol{X}^*\widehat{\beta}_{GLS}\right)^\top\left(\boldsymbol{Y}^* - \boldsymbol{X}^*\widehat{\beta}_{GLS}\right)\\ &= \frac{1}{N-(k+1)}\left(\boldsymbol{\Psi}^\top\boldsymbol{Y} - \boldsymbol{\Psi}^\top\boldsymbol{X}\widehat{\beta}_{GLS}\right)^\top\left(\boldsymbol{\Psi}^\top\boldsymbol{Y} - \boldsymbol{\Psi}^\top\boldsymbol{X}\widehat{\beta}_{GLS}\right)\\ &= \frac{1}{N-(k+1)}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}_{GLS}\right)^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}_{GLS}\right)\end{aligned}$$

which leads to the following **unbiased** and **consistent** estimator of $\sigma^2$:

$$\widehat{\sigma}^2 = \frac{1}{N-(k+1)}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}_{GLS}\right)^\top\boldsymbol{\Omega}^{-1}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}_{GLS}\right)$$

## Properties of the GLS Estimator

Because of our conveniently written GMLR model form $\boldsymbol{Y}^* = \boldsymbol{X}^*\beta + \varepsilon^*$, the following GLS properties can be derived following the same principles as in the OLS case:

- The GLS is an **unbiased** estimator:

$$\mathbb{E}\left(\widehat{\beta}_{GLS}\right) = \beta + \left(\boldsymbol{X}^{*\top}\boldsymbol{X}^*\right)^{-1}\boldsymbol{X}^{*\top}\mathbb{E}\left(\varepsilon^*\right) = \beta$$

- The GLS variance-covariance matrix is:

$$\mathbb{V}\mathrm{ar}\left(\widehat{\beta}_{GLS}\right) = \sigma^2\left(\boldsymbol{X}^{*\top}\boldsymbol{X}^*\right)^{-1} = \sigma^2\left(\boldsymbol{X}^{\top}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\boldsymbol{X}\right)^{-1} = \sigma^2\left(\boldsymbol{X}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}$$

- If the errors are normally distributed, then:

$$\widehat{\beta}_{GLS}|\boldsymbol{X} \sim \mathcal{N}\left(\beta,\ \sigma^2\left(\boldsymbol{X}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\right)$$

- $\widehat{\sigma}^2$ is unbiased and consistent.

The GLS estimator is BLUE for the GMLR.

Consequently, for the GMLR, the OLS estimator is also inefficient.

Sometimes, we may decompose the whole variance-covariance matrix:

$$\mathbb{V}\mathrm{ar}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right) = \sigma_{\epsilon}^{2}\boldsymbol{\Omega} = \boldsymbol{\Sigma}, \quad \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{-1}$$

Then the variance-covariance matrix is: $\mathbb{V}\mathrm{ar}\left(\boldsymbol{\Phi}^{\top}\boldsymbol{\varepsilon}\right) = \boldsymbol{\Phi}^{\top}\mathbb{V}\mathrm{ar}\left(\boldsymbol{\varepsilon}\right)\boldsymbol{\Phi} = \boldsymbol{I}$

In some econometric software, the variance-covariance matrix may be decomposed in this way. The difference is that in this case the GLS error variance is **known** to be 1, while before, we needed to estimate $\sigma^2$.

Either way, **the GLS estimates will be the same, regardless of the error variance-covariance specification used**.

## Multiple Linear Restriction Test

$M$ multiple Linear Restrictions can be tested via the following hypothesis:

$$\begin{cases} H_0 & : \boldsymbol{L}\beta = \boldsymbol{r} \\ H_1 & : \boldsymbol{L}\beta \neq \boldsymbol{r} \end{cases}$$

Then, the $F$-statistic is:

$$F = \frac{\left(\boldsymbol{L}\widehat{\beta}_{GLS} - \boldsymbol{r}\right)^{\top} \left[\boldsymbol{L}\left(\boldsymbol{X}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{L}^{\top}\right]^{-1} \left(\boldsymbol{L}\widehat{\beta}_{GLS} - \boldsymbol{r}\right)}{M\widehat{\sigma}^2} \sim F_{(M,N-(k+1))}$$

## Weighted Least Squares

It is particularly easy to obtain GLS estimates, when the error terms are **heteroskedastic** but **uncorrelated** - this implies that the matrix is $\mathbf{\Omega}$ diagonal, with non-constant diagonal elements:

$$\mathbf{\Omega} = \begin{bmatrix} \omega_1^2 & 0 & 0 & \dots & 0 \\ 0 & \omega_2^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \omega_N^2 \end{bmatrix} \iff \mathbf{\Omega}^{-1} = \begin{bmatrix} \omega_1^{-2} & 0 & 0 & \dots & 0 \\ 0 & \omega_2^{-2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \omega_N^{-2} \end{bmatrix}$$

So, if we select matrix $\mathbf{\Psi}$ as:

$$\mathbf{\Psi} = \begin{bmatrix} \omega_1^{-1} & 0 & 0 & \dots & 0 \\ 0 & \omega_2^{-1} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \omega_N^{-1} \end{bmatrix}$$

Then, our typical regression expression $\boldsymbol{\Psi}^\top \boldsymbol{Y} = \boldsymbol{\Psi}^\top \boldsymbol{X}\beta + \boldsymbol{\Psi}^\top \varepsilon$ can be written as:

$$Y_i/\omega_i = \beta_0 \cdot (1/\omega_i) + \beta_1 \cdot (X_{1,i}/\omega_i) + ... + \beta_k \cdot (X_{k,i}/\omega_i) + \epsilon_i/\omega_i$$

In other words, **all** of the variables, **including the constant term**, are multiplied by the same **weights** $\omega_i^{-1}$.

Consequently, $\beta_0$ is no longer multiplied by a constant, so **when estimating the model with these transformed variables via OLS, create a new variable for $1/\omega_i$, and exclude a constant from your model**.

There are various ways of determining the weights used in weighted least squares estimation. In the simplest case, either from economic theory, or from some preliminary testing, we may assume that $\mathbb{E}(\epsilon_i^2)$ is proportional to $Z_i^2$, where $Z_i$ is some **observed variable**.

For example $Z_i$ may be the population, or income. Alternatively, sometimes $\mathbb{E}(\epsilon_i^2)$ may be proportional the sample size used to obtain an average (or total) value of observation $i$, which is saved in the dataset.

- If observation $Y_i$ is an **average** of $N_i$ equally variable (uncorrelated) observations, then $\mathbb{V}\mathrm{ar}\left(Y_i\right) = \sigma^2/N_i$ and $\omega_i = 1/\sqrt{N_i}$;
- If observation $Y_i$ is an aggregated **total** of $N_i$ (uncorrelated) observations, then $\mathbb{V}\mathrm{ar}\left(Y_i\right) = N_i\sigma^2$ and $\omega_i = \sqrt{N_i}$;
- If the variance of $Y_i$ is proportional to some predictor $Z_i$, then $\mathbb{V}\mathrm{ar}\left(Y_i\right) = Z_i^2\sigma^2$ and $\omega_i = Z_i$;

It is possible to report various summary statistics, like $R^2$, *ESS* and *RSS* in terms of $Y_i$, or $Y_i/\omega_i$, however, $R^2$ is only valid for the transformed variables $Y_i/\omega_i$ (since the coefficients are estimated on the transformed dependent variables).

## Feasible Generalized Least Squares

In practice the true form of $\mathbf{\Omega}$ is **unknown**.

Even in the simplest case of $\mathbf{\Omega} = \text{diag}\left(\omega_1^2, ..., \omega_N^2\right)$, we would need to estimate a total of $k + 1 + N$ parameters ($\beta_0, ..., \beta_k$ and $\omega_1^2, ..., \omega_N^2$), which is **impossible** since we only have $N$ observations (and a rule of thumb says that we would need a minimum of 5 - 20 observations per parameters in order to get satisfactory estimates of $\beta$).

Therefore, we need to assume some simplifying conditions for $\mathbf{\Omega}$. For example, if $\mathbf{\Omega}$ is diagonal, suppose that $\omega_i^2 = \alpha_0 + \alpha_1 X_{m,i}$ for some $m = 1, ..., k$ - now we only need to estimate two additional parameters (as opposed to $N$).

The case, where we use an estimated matrix $\widehat{\mathbf{\Omega}}$, is known as the **feasible (or, estimable) generalized least squares (FGLS)**. The estimators have good properties in large samples.

Consequently, using the estimated matrix of $\boldsymbol{\Omega}$ results in the following FGLS estimator:

**The FGLS estimator is defined as:**

$$\widehat{\beta}_{FGLS} = \left(\boldsymbol{X}^\top \widehat{\boldsymbol{\Omega}}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{\Omega}}^{-1} \boldsymbol{Y}$$

where $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}}(\boldsymbol{\theta})$ is a parametric estimation (with parameter vector $\boldsymbol{\theta}$) of the true unknown matrix $\boldsymbol{\Omega}$.

For the **weighted least squares** case we would need to:

1. Estimate a model via OLS and calculate the residuals $\widehat{\epsilon}_{i,(OLS)}$. If the regression is correctly specified, an estimate of $\omega_i^2$ is the OLS squared residual $\widehat{\epsilon}^2_{i,(OLS)}$.
2. Generally, the residuals are much too variable to be used directly in estimating the weights, so instead we could use:
   - some other kind of transformation of the residuals, or their squares. For example, $\log(\widehat{\epsilon}_i^2)$. Then regressing these variables on some selected predictors and using the fitted values as weights.
   - regressing the squared residuals against the fitted values. Then the square root of the fitted values could be used as $\omega_i$;
   - regressing the absolute residuals against the fitted values. Then the fitted values could be used as $\omega_i$;

Plotting the residuals (or their squared values, or their absolute values, or the root square of their absolute value) against the independent variable $X$ would help in determining the regression form.

The resulting **fitted values** from this regression are estimates of $\omega_i$.

Generally, the structure can be imposed in two most popular ways: by assuming error heteroskedasticity, or by assuming error serial correlation.

If the assumption about the error covariance structure is incorrect - heteroskedasticity still remains and FGLS is is no longer BLUE. In this case:

- FGLS it is still unbiased, just like OLS;
- the FGLS estimator of the error variance is biased, just like OLS (but the magnitude of the bias is not the same);

Furthermore, regarding the use of FGLS instead of GLS, it can be stated that:

Since we do not know the true $\mathbf{\Omega}$, but instead estimate $\widehat{\mathbf{\Omega}}$ from the sample, then it becomes a random variable (just like when we estimate other parameters, like $\widehat{\beta}$). This fact affects the distribution of the GLS estimator. **Very little is known about the finite-sample properties of the FGLS estimator**.

## A Note on Coefficient Interpretation

Since we are using the following model:

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\beta + \varepsilon^*, \text{ where } \boldsymbol{Y}^* = \boldsymbol{\Psi}^\top \boldsymbol{Y}, \quad \boldsymbol{X}^* = \boldsymbol{\Psi}^\top \boldsymbol{X}, \quad \varepsilon^* = \boldsymbol{\Psi}^\top \varepsilon$$

to estimate $\widehat{\beta}$ with GLS (or FGLS), then the predicted values are:

$$\widehat{\boldsymbol{Y}^*} = \boldsymbol{X}^* \widehat{\beta}_{GLS} = \boldsymbol{\Psi}^\top \boldsymbol{X} \widehat{\beta}_{GLS} \quad \left( i.e. \ = \boldsymbol{\Psi}^\top \widehat{\boldsymbol{Y}} \right)$$

In other words multiplying both sides by $\left(\boldsymbol{\Psi}^\top\right)^{-1}$ yields:

$$\left(\boldsymbol{\Psi}^\top\right)^{-1} \widehat{\boldsymbol{Y}^*} = \boldsymbol{X} \widehat{\beta}_{GLS} = \widehat{\boldsymbol{Y}}$$

If we were to use $\boldsymbol{X}^*$ to calculate the fitted/predicted values, then we would need to multiply the fitted values $\widehat{\boldsymbol{Y}^*}$ by $\left(\boldsymbol{\Psi}^\top\right)^{-1}$ to get the fitted values for the original data $\widehat{\boldsymbol{Y}}$. Alternatively, this expression also means that we can use the the **original design matrix** $\boldsymbol{X}$ with the GLS estimates of $\beta$ to get the fitted/predicted values of $\boldsymbol{Y}$. In other words:

> The GLS (as well as WLS and FGLS) estimates of $\beta$ retain their coefficient interpretation of how a unit increase in one explanatory variable $X_j$ affects the dependent variable $Y$, ceteris paribus.

# Heteroskedastic Errors

# Heteroskedastic Errors

Consider the case where assumption (MR.3) **does not hold**, but assumption (MR.4) (and the other remaining assumptions (MR.1), (MR.2), (MR.5) and, optionally, (MR.6)) **are still valid**. Then we can write the following model as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right) = \boldsymbol{0}, \quad \mathbb{V}\mathrm{ar}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right) = \mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\top}\right) = \boldsymbol{\Sigma} \neq \sigma_{\epsilon}^{2}\boldsymbol{I}$$

The case when the error variance-covariance matrix is diagonal, but not equal to $\sigma_{\epsilon}^{2}\boldsymbol{I}$, expressed as:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_N^2 \end{bmatrix}$$

is referred to as **heteroskedasticity**. As mentioned before:

- the OLS estimators will remain unbiased;
- the OLS variance estimator is biased and inconsistent;
- the usual $t$-statistics of the OLS estimates do not have Student's $t$ distribution, even in large samples.

1. Assume that $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ is the true model.
2. Test the null hypothesis that **the residuals are homoskedastic**:

$$H_0 : \mathbb{V}\mathrm{ar}\left(\varepsilon | \boldsymbol{X}\right) = \sigma_\epsilon^2 \boldsymbol{I}$$

3. If we fail to reject the null hypothesis - we can use the usual OLS estimators.
4. If we reject the null hypothesis, there are two ways we can go:
   - Use the OLS estimators, but correct their variance estimators (i.e. make them consistent);
   - Instead of OLS, use the weighted least squares (WLS) to estimate the parameters;
   - Attempt to specify a different model, which would hopefully, be able to account for heteroskedasticity (this is the least preferred method - our initial model should already be the one we want in terms of variables, signs, interpretation, etc.).

## Example

We will simulate the following model:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$$
$$u_i = i \cdot \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \left[ \text{ or } u_i \sim \mathcal{N}(0, i^2 \cdot \sigma^2) \right]$$

```
set.seed(123)
#
N <- 200
beta_vec <- c(10, 5, -3)
#
x1 <- seq(from = 0, to = 5, length.out = N)
x2 <- sample(seq(from = 3, to = 17, length.out = 80), size = N, replace = TRUE)
e  <- rnorm(mean = 0, sd = 1:N, n = N)
x_mat <- cbind(1, x1, x2)
y <- x_mat %*% beta_vec + e
#
data_mat <- data.frame(y, x1, x2)
```

## Testing For Heteroskedasticity

We can examine the presence of heteroskedasticity from the residuals plots, as well as conducting a number of formal tests.

We will begin by estimating our model via OLS, as we usually would.

```
mdl_1 <- lm(y ~ x1 + x2, data = data_mat)
print(round(coef(summary(mdl_1)), 5))
```

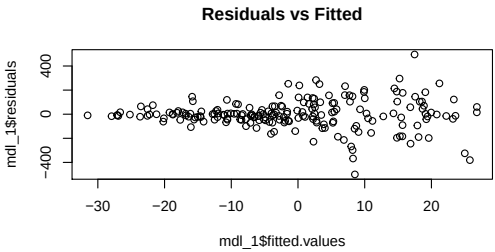
```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -0.89066   28.42373 -0.03134  0.97503
## x1           7.21929    5.93429  1.21654  0.22524
## x2          -1.83213    2.18885 -0.83703  0.40359
```

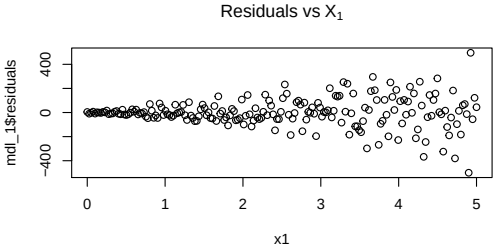
## Residual Plot Diagnostic

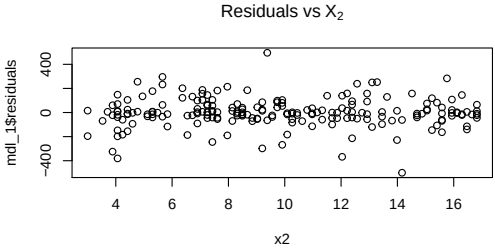
One way of investigating the existence of heteroskedasticity is to visually examine the OLS model residuals. If they are homoskedastic, there should be no pattern in the residuals. If the errors are heteroskedastic, they would exhibit increasing or decreasing variation in some systematic way.

For example, variation may increase with larger values of $\widehat{Y}$, or with larger values of $X_j$.

```
par(mfrow = c(2, 2))
#
plot(mdl_1$residuals, main = "Run-Sequence Plot")
plot(mdl_1$fitted.values, mdl_1$residuals, main = "Residuals vs Fitted")
plot(x1, mdl_1$residuals, main = bquote("Residuals vs"~X[1]))
plot(x2, mdl_1$residuals, main = bquote("Residuals vs"~X[2]))
```


The residuals appear to have different variance, depending on the value of $\widehat{Y}$ and $X_1$.

## Heteroskedasticity Tests

Most of the tests are identical to the ones described in @ref(OLS-Test-Heteroskedastic), but with a bit more mathematical detail.

## The Goldfeld–Quandt Test

This test is designed for the case where we have two sub-samples with possibly different variances. The sub-samples can be created in a number of ways:

- create data sub-samples based on an indicator variable;
- sort the data along a known explanatory variable, from lowest to highest;

The test outline is as follows . . .

Let the the sub-samples $j = 1, 2$ contain $N_j$ observations and let the regression on sub-sample $j$ have $k_j + 1$ parameters (including the intercept). Let $\widehat{\varepsilon}_j$ be the residuals of the regression on the $j$-th sub-sample and let the true variance of the sub-sample errors be $\sigma_j^2$. Then the sub-sample variance can be estimated by:

$$\widehat{\sigma}_j^2 = \frac{\widehat{\varepsilon}_j^\top \widehat{\varepsilon}_j}{N_j - (k_j + 1)}$$

We want to test the null hypothesis:

$$\begin{cases} H_0 : \widehat{\sigma}_1^2 / \widehat{\sigma}_2^2 = 1 \\ H_1 : \widehat{\sigma}_1^2 / \widehat{\sigma}_2^2 \neq 1 \end{cases}$$

Then, the **GQ** statistic can be defined as:

$$GQ = \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} \sim F_{(N_1 - (k_1+1), N_2 - (k_2+1))}$$

If we reject the null hypothesis, for a chosen significance level $\alpha$, then the variances in the sub-samples are different, hence we cannot reject the null hypothesis of *heteroskedasticity*.

In our example model, we see that the residuals can be ordered by either the fitted values, or by $X_1$. If we specify to order by $X_1$:

```
GQ_t <- lmtest::gqtest(mdl_1, alternative = "two.sided", order.by = ~ x1)
print(GQ_t)
```

```
##
##  Goldfeld-Quandt test
##
## data:  mdl_1
## GQ = 10.217, df1 = 97, df2 = 97, p-value < 2.2e-16
## alternative hypothesis: variance changes from segment 1 to 2
```

since the $p$-value is less than 0.05, we **reject the null hypothesis** and conclude that the residuals are heteroskedastic.

- Furthermore, we can order by the fitted values:

```
GQ_t <- lmtest::gqtest(mdl_1, alternative = "two.sided", order.by = order(mdl_1$fitted.values))
print(GQ_t)
```

```
##
##  Goldfeld-Quandt test
##
## data:  mdl_1
## GQ = 5.7826, df1 = 97, df2 = 97, p-value = 3.736e-16
## alternative hypothesis: variance changes from segment 1 to 2
```

and we would arrive at the same conclusions.

On the other hand, if we would have specified to order by $X_2$:

```
GQ_t <- lmtest::gqtest(mdl_1, alternative = "two.sided", order.by = ~ x2)
print(GQ_t)
```

```
##
##  Goldfeld-Quandt test
##
## data:  mdl_1
## GQ = 1.1747, df1 = 97, df2 = 97, p-value = 0.4292
## alternative hypothesis: variance changes from segment 1 to 2
```

We would have no grounds to reject the null hypothesis, and would have concluded that the errors are homoskedastic.

**This is why the residual plots are important.**

Without them, we may have blindly selected one explanatory variable at random to order by, and would have arrived at a completely different conclusion!

By default, if `order.by` is not specified, then the data is taken as **ordered** - i.e. it would be the equivalent of the residual run-sequence plot.

## A General Heteroskedasticity Test

The next test is used for conditional heteroskedasticity, which is related to explanatory variables. This test is a generalization of the **Breusch–Pagan Test**.

Consider the following regression:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

We assume that the general expression for the conditional variance can be expressed as:

$$\mathbb{V}\mathrm{ar}\left(\varepsilon|\boldsymbol{Z}\right) = \mathbb{E}\left(\varepsilon\varepsilon^{\top}|\boldsymbol{Z}\right) = \mathrm{diag}\left(h\left(\boldsymbol{Z}\boldsymbol{\alpha}\right)\right)$$

where $\boldsymbol{Z}$ is some explanatory variable matrix, which may include some (or all) of the explanatory variables from $\boldsymbol{X}$. $h(\cdot)$ is some smooth function, and $\boldsymbol{\alpha} = [\alpha_0, ..., \alpha_m]^{\top}$.
We want to test the null hypothesis:

$$\begin{cases} H_0 : \alpha_1 = ... = \alpha_m = 0 \\ H_1 : \text{ at least one } \alpha_j \neq 0,\ j \in \{1, ..., m\} \end{cases}$$

Then the associated **BP** test specifies the following **linear** model for the squared residuals:

$$\widehat{\varepsilon}\widehat{\varepsilon}^{\top} = \mathrm{diag}\left(\boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{v}\right)$$

Estimating the squared residuals model via OLS yields the parameter estimates. Then, using the $F$-test for the joint significance of $\alpha_1, ..., \alpha_m$ would be equivalent to testing for homoskedasticity. Alternatively, and more conveniently, there is a test based on the $R^2_{\widehat{\epsilon}}$:

$$LM = N \cdot R^2_{\widehat{\epsilon}} \sim \chi^2_m$$

If we reject the null hypothesis, for a chosen significance level $\alpha$, then the error variance is heteroskedastic.

```r
BP_t <- lmtest::bptest(mdl_1)
print(BP_t)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mdl_1
## BP = 35.896, df = 2, p-value = 1.604e-08
```

The $p$-value is less than the chosen 0.05 significance level, so the residuals are heteroskedastic.

## The White Test

An even more generalized test, which proposes to include not only the exogenous variables, but also their polynomial and interaction terms.

The associated OLS squared residual regression is similar to the general case:

$$\widehat{\varepsilon}\widehat{\varepsilon}^{\top} = \text{diag}\,(\boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{v})$$

except now, the matrix $\boldsymbol{Z}$ also includes $s$ additional **polynomial** and **interaction** terms of the explanatory variables. The parameter vector is $\alpha_1, ..., \alpha_{m+s}$. We want to test the null hypothesis:

$$\begin{cases} H_0 : \alpha_1 = ... = \alpha_{m+s} = 0 \\ H_1 : \text{ at least one } \alpha_j \neq 0,\ j \in \{1, ..., m+s\} \end{cases}$$

Then the test statistic is calculated in the same way as before:

$$LM = N \cdot R^2_{\widehat{\epsilon}} \sim \chi^2_{m+s}$$

One difficulty with the White test is that it can detect problems other than heteroskedasticity.
Thus, while it is a useful diagnostic, be careful about interpreting the test result - instead of heteroskedasticity, it may be that you have an incorrect functional form, or an omitted variable.

The test is also performed similar to how it was for the univariate regression with one explanatory variable:

```
W_t <- lmtest::bptest(mdl_1, ~ x1 + I(x1^2) + x2 + I(x2^2) + x1:x2)
print(W_t)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mdl_1
## BP = 43.195, df = 5, p-value = 3.373e-08
```

We again reject the null hypothesis of homoskedasticity and conclude that the residuals are heteroskedastic.

## Heteroskedasticity-Consistent Standard Errors (HCE)

Once, we have determined that the errors are heteroskedastic, we want to have a way to account for that.
One alternative is to stick with OLS estimates, but correct their variance. This is known as the **White correction** - it *will not* change the $\widehat{\beta}_{OLS}$, which are unbiased and ineffective, but it *will* correct $\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\beta}_{OLS}\right)$.

If the errors from the error vector $\varepsilon$ are independent, but have distinct variances, so that $\mathbb{V}\mathrm{ar}(\varepsilon|\boldsymbol{X}) = \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, ..., \sigma_N^2)$.
Then the true underlying variance-covariance matrix of the OLS estimates would be:

$$\mathbb{V}\left(\widehat{\beta}_{OLS}\right) = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\Sigma} \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

Since $\mathbb{E}(\epsilon_i) = 0$, then $\mathbb{V}\mathrm{ar}(\epsilon_i) = \mathbb{E}(\epsilon_i^2) = \sigma_i^2$. Which means that we can estimate the variance diagonal elements as:

$$\widehat{\sigma_i^2} = \widehat{\epsilon_i^2}$$

Let $\widehat{\boldsymbol{\Sigma}} = \mathrm{diag}(\widehat{\epsilon_1^2}, ..., \widehat{\epsilon_N^2})$. Then the **Whites Estimators**, also known as the **H**eteroscedasticity-**C**onsistent **E**stimator (**HCE**), can be specified as:

$$\mathbb{V}_{HCE}\left(\widehat{\beta}_{OLS}\right) = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

Furthermore, alternative specifications of $\widehat{\boldsymbol{\Sigma}}$ are also possible.

In our simulated case, the white correction can be estimated either **manually**:

```
xtx_inv <- solve(t(x_mat) %*% x_mat)
V_HC <- xtx_inv %*% t(x_mat) %*% diag(mdl_1$residuals^2) %*% x_mat %*% xtx_inv
#
print(V_HC)
```

```
##                       x1         x2
##    693.25117 -91.740409 -52.244062
## x1 -91.74041  46.570134   2.341761
## x2 -52.24406   2.341761   4.749053
```

or, via the built-in functions:

```
V_HC_1 <- sandwich::vcovHC(mdl_1, type = "HC0")
print(V_HC_1)
```

```
##             (Intercept)         x1         x2
## (Intercept)   693.25117 -91.740409 -52.244062
## x1            -91.74041  46.570134   2.341761
## x2            -52.24406   2.341761   4.749053
```

Having estimated the standard errors is nice, but we would also like to get the associated $p$-values.

```
print(lmtest::coeftest(mdl_1, vcov. = V_HC_1))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.89066   26.32966 -0.0338   0.9730
## x1           7.21929    6.82423  1.0579   0.2914
## x2          -1.83213    2.17923 -0.8407   0.4015
```

compared to the biased standard errors:

```
print(coef(summary(mdl_1)))
```

```
##               Estimate Std. Error     t value  Pr(>|t|)
## (Intercept) -0.8906595  28.423734 -0.03133506 0.9750341
## x1           7.2192910   5.934290  1.21653830 0.2252355
## x2          -1.8321252   2.188846 -0.83702804 0.4035910
```

- We see that after correcting the standard errors, the $p$-values are also slightly different. The extent of the difference depends of the severity of heteroskedasticity and the method used to correct for it.
- While the corrected variance estimated helps us avoid incorrect $t$-statistics (as well as avoid incorrect confidence intervals) in case of heteroskedasticity, **it does not address the problem that OLS estimates are no longer the best** (in terms of variance).
- If we have a large number of observations (i.e. thousands upon thousands), then the **robust** corrected standard errors are enough. On the other hand, if we have s smaller sample size, then we would like to have a more efficient estimator - this is where our previously presented GLS and FGLS come in handy.

## HCE Alternative Specifications

A nice summary on alternative HCE specifications are presented in this paper by Hayes and Cai 2007. Several alternative $\widehat{\boldsymbol{\Sigma}}$ specifications have been proposed in literature. Here we will briefly summarize them.

Taking $\widehat{\boldsymbol{\Sigma}} = \text{diag}(\hat{\epsilon}_1^2, ..., \hat{\epsilon}_N^2)$, where $\epsilon_i$ are the OLS residuals, leads to the following **White's** HCE:

$$\text{HC0}_{\widehat{\boldsymbol{\beta}}_{OLS}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \text{diag}(\hat{\epsilon}_1^2, ..., \hat{\epsilon}_N^2) \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

- for small sample sizes, the standard errors from $\text{HC0}_{\widehat{\boldsymbol{\beta}}_{OLS}}$ are quite biased. This results in overly liberal inferences in regression models.
- $\text{HC0}_{\widehat{\boldsymbol{\beta}}_{OLS}}$ is consistent when the errors are heteroskedastic - the bias shrinks as the sample size increases.

Estimators in this family have come to be known as **sandwich estimators** - $\boldsymbol{X}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{X}$ is the filling between two matrices $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$.

Consequently, alternative estimators to $\text{HC0}_{\widehat{\boldsymbol{\beta}}_{OLS}}$ were proposed, which are asymptotically equivalent to $\text{HC0}_{\widehat{\boldsymbol{\beta}}_{OLS}}$, but have superior small sample properties, when compared to $\text{HC0}_{\widehat{\boldsymbol{\beta}}_{OLS}}$.

- An alternative HC estimator adjusts the degrees of freedom of $\text{HC0}_{\widehat{\beta}_{OLS}}$:

$$\text{HC1}_{\widehat{\beta}_{OLS}} = \frac{N}{N-(k+1)} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \text{diag}(\widehat{\epsilon}_1^2, ..., \widehat{\epsilon}_N^2) \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

Another HC estimator is defined based on extensive research, that finite-sample bias is a result of the existence of points of **high leverage**. It uses the weight matrix, which is defined as:

$$\boldsymbol{H} = \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$$

- Then the HC estimator is constructed by by weighting the $i$-th squared OLS residual by using the $i$-th diagonal elements of the weight matrix $\boldsymbol{H}$:

$$\text{HC2}_{\widehat{\beta}_{OLS}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \text{diag}\left(\frac{\widehat{\epsilon}_1^2}{1-h_{11}}, ..., \frac{\widehat{\epsilon}_N^2}{1-h_{NN}}\right) \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

- Similarly, using $1/(1-h_{11})^2$, instead of $1/(1-h_{11})$ for the weights leads to yet another HCE specification:

$$\text{HC3}_{\widehat{\beta}_{OLS}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \text{diag}\left(\frac{\widehat{\epsilon}_1^2}{(1-h_{11})^2}, ..., \frac{\widehat{\epsilon}_N^2}{(1-h_{NN})^2}\right) \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

**Evaluating the empirical power of the four methods: HC0, HC1, HC2 and HC3 suggested that HC3 is a superior estimate, regardless of the presence, or absence, of heteroskedasticity.**

Nevertheless, the performance of HC3 depends on the presence, or absence, of points of high leverage in $\boldsymbol{X}$ and it may fail for certain forms of heteroskedasticity (for example, when the predictors are from heavy-tailed distributions, and the errors are from light-tailed distributions).

- To account for large leverage values, another HC estimator was proposed:

$$\text{HC4}_{\widehat{\beta}_{OLS}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \text{diag}\left(\frac{\widehat{\epsilon}_1^2}{(1-h_{11})^{\delta_1}}, ..., \frac{\widehat{\epsilon}_N^2}{(1-h_{NN})^{\delta_N}}\right) \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

where:

$$\delta_i = \min\left\{4, \frac{N \cdot h_{ii}}{k+1}\right\}$$

Simulation tests indicated that HC4 can outperform HC3 when there are high leverage points and non-normal errors.

Nevertheless, even with these alternative specifications, the variability of HCE are often larger than model-based estimators, like WLS, GLS of FGLS, if the residual covariance-matrix is correctly specified. On the other hand, HCE are derived under a minimal set of assumptions about the errors. As such, it is useful when heteroskedasticity if of an unknown form and cannot be adequately evaluated from the data.

**Therefore, when using HCE (instead of some other model-based estimation methods), we are trading efficiency for consistency.**

Furthermore, there is also another specification, HC5 (Source).

In our example dataset different HCE can be easily specified with the built-in functions:

```
print(lmtest::coeftest(mdl_1, vcov. = sandwich::vcovHC(mdl_1, type = "HC0")))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.89066   26.32966 -0.0338   0.9730
## x1           7.21929    6.82423  1.0579   0.2914
## x2          -1.83213    2.17923 -0.8407   0.4015
```

```
print(lmtest::coeftest(mdl_1, vcov. = sandwich::vcovHC(mdl_1, type = "HC1")))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.89066   26.52939 -0.0336   0.9733
## x1           7.21929    6.87600  1.0499   0.2950
## x2          -1.83213    2.19576 -0.8344   0.4051
```

```r
print(lmtest::coeftest(mdl_1, vcov. = sandwich::vcovHC(mdl_1, type = "HC2")))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.89066   26.59922 -0.0335   0.9733
## x1           7.21929    6.89407  1.0472   0.2963
## x2          -1.83213    2.20143 -0.8322   0.4063
```

```r
print(lmtest::coeftest(mdl_1, vcov. = sandwich::vcovHC(mdl_1, type = "HC3")))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.89066   26.87208 -0.0331   0.9736
## x1           7.21929    6.96475  1.0365   0.3012
## x2          -1.83213    2.22390 -0.8238   0.4110
```

```r
print(lmtest::coeftest(mdl_1, vcov. = sandwich::vcovHC(mdl_1, type = "HC4")))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.89066   26.72908 -0.0333   0.9735
## x1           7.21929    6.92585  1.0424   0.2985
## x2          -1.83213    2.21168 -0.8284   0.4085
```

Note: `Python` currently does not have `HC4` specification. Available specifications can be found [here](https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.RegressionResults.html

As per the sandwich::vcovHC documentation, we can even calculate the HC5 specification:

```
print(lmtest::coeftest(mdl_1, vcov. = sandwich::vcovHC(mdl_1, type = "HC5")))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.89066   26.52816 -0.0336   0.9733
## x1           7.21929    6.87476  1.0501   0.2950
## x2          -1.83213    2.19536 -0.8345   0.4050
```

## Weighted Least Squares (WLS)

After detecting heteroskedasticity in the errors, we may want to impose a structure on the residual covariance matrix and estimate the coefficients via GLS. If we know that there is no serial correlation in the errors, then the covariance is diagonal. This leads to a specific case of GLS - weighted least squares (WLS).

In reality, we do not know the true form of heteroskedasticity. So, even knowing that the covariance matrix is diagonal still does not say anything about the diagonal elements, they could be either of the following:

$$\mathbf{\Sigma} = \sigma^2 \cdot \text{diag}\left(X_{j,1}, X_{j,2}, ..., X_{j,N}\right) \quad \text{i.e., variance is proportional to } X_j$$

$$\mathbf{\Sigma} = \sigma^2 \cdot \text{diag}\left(X_{j,1}^2, X_{j,2}^2, ..., X_{j,N}^2\right)$$

$$\mathbf{\Sigma} = \sigma^2 \cdot \text{diag}\left(\sqrt{X_{j,1}}, \sqrt{X_{j,2}}, ..., \sqrt{X_{j,N}}\right), \text{ if } X_{j,i} \geq 0, \ \forall i = 1, ..., N$$

Furthermore, in a multiple regression setting, heteroskedasticity pattern may depend on more than one explanatory variable - it could even be related to variables, not included in the model.

**So, how do we select the most likely form for heteroskedasticity?**

In general, one specification of heteroskedasticity, which works quite well is:

$$\begin{aligned}\sigma_i^2 &= \exp\left(\alpha_0 + \alpha_1 Z_{1,i} + ... + \alpha_m Z_{m,i}\right) \\ &= \sigma^2 \exp\left(\alpha_1 Z_{1,i} + ... + \alpha_m Z_{m,i}\right), \quad \text{where } \exp\left(\alpha_0\right) = \sigma^2\end{aligned}$$

taking logarithms of both sides and adding/removing the OLS residual yields:

$$\log(\sigma_i^2) = \alpha_0 + \alpha_1 Z_{1,i} + ... + \alpha_m Z_{m,i} \pm \log(\widehat{\epsilon}_i^2)$$

which simplifies to:

$$\begin{aligned}\log(\widehat{\epsilon}_i^2) &= \alpha_0 + \alpha_1 Z_{1,i} + ... + \alpha_m Z_{m,i} + \log\left(\frac{\widehat{\epsilon}_i^2}{\sigma_i^2}\right) \\ &= \alpha_0 + \alpha_1 Z_{1,i} + ... + \alpha_m Z_{m,i} + v_i\end{aligned}$$

using this model we can estimate $\alpha_0, ..., \alpha_m$ via OLS. the properties of this model depend on the introduced error term $v_i$ - whether it is homoskedastic, with zero-mean. In smaller samples it is not, but in larger samples it is closer to what we expect from the error term.

**Feasible GLS Procedure**:

1. Estimate the regression $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ via OLS.
2. Use the residuals $\widehat{\varepsilon}_{OLS}$ and create $\log(\epsilon_i^2)$.
3. Estimate the regression $\log(\epsilon_i^2) = \alpha_0 + \alpha_1 Z_{1,i} + ... + \alpha_m Z_{m,i} + v_i$ and calculate the fitted values $\widehat{\log(\epsilon_i^2)}$. In practice we use the same variables from $\boldsymbol{X}$, unless we know for sure that there are additional explanatory variables, which may determine heteroskedasticity.
4. Take the exponent of the fitted values: $\widehat{h}_i = \exp\left(\widehat{\log(\epsilon_i^2)}\right)$.
5. Estimate the regression $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ via **WLS** using weights $\omega_i^{-1} = 1/\sqrt{\widehat{h}_i}$, i.e. use FGLS with $\boldsymbol{\Psi}^\top = \boldsymbol{\Psi} = \text{diag}\left(1/\sqrt{\widehat{h}_1}, ..., 1/\sqrt{\widehat{h}_N}\right)$.

As noted before, applying WLS is equivalent to dividing each observation by $\sqrt{\widehat{h}_i}$ and estimating OLS on the transformed data:

$$Y_i/\sqrt{\widehat{h}_i} = \beta_0 \cdot \left(1/\sqrt{\widehat{h}_i}\right) + \beta_1 \cdot \left(X_{1,i}/\sqrt{\widehat{h}_i}\right) + ... + \beta_k \cdot \left(X_{k,i}/\sqrt{\widehat{h}_i}\right) + \epsilon_i/\sqrt{\widehat{h}_i}$$

The **downside** is that our model no longer contains a constant - $\beta_0$ is now used for the new (**non-constant**) variable $1/\sqrt{\widehat{h}_i} \neq 1$.

We begin by manually transforming the data and estimating OLS on the transformed variables:

```
resid_data <- data.frame(log_e2 = log(mdl_1$residuals^2), x1, x2)
#
resid_mdl  <- lm(log_e2 ~ x1 + x2, data = resid_data)
h_est <- exp(resid_mdl$fitted.values)
#
data_weighted = data.frame(data_mat / sqrt(h_est),
                           weighted_intercept = 1 / sqrt(h_est))
mdl_w_ols <- lm(y ~ -1 + weighted_intercept + x1 + x2, data = data_weighted)
print(round(coef(summary(mdl_w_ols)), 5))
```

```
##                    Estimate Std. Error  t value Pr(>|t|)
## weighted_intercept -1.39237   10.19482 -0.13658  0.89151
## x1                  7.97645    4.00325  1.99249  0.04770
## x2                 -2.02459    0.88157 -2.29656  0.02270
```

Next, we carry out WLS (where $\mathbf{\Omega}^{-1} = \text{diag}\left(\widehat{h_1^{-1}}, ..., \widehat{h_N^{-1}}\right)$). Manually:

```
beta_wls <- solve(t(x_mat) %*% diag(1 / h_est) %*% x_mat) %*% t(x_mat) %*% diag(1 / h_est) %*% y
#
resid_wls   <- y - x_mat %*% beta_wls
sigma2_wls  <- (t(resid_wls) %*% diag(1 / h_est) %*% resid_wls) / (N - length(beta_wls))
beta_wls_se <- c(sigma2_wls) * solve(t(x_mat) %*% diag(1 / h_est) %*% x_mat)
#
print(data.frame(est = beta_wls, se = sqrt(diag(beta_wls_se))))
```

```
##          est         se
##    -1.392372 10.1948202
## x1  7.976445  4.0032528
## x2 -2.024586  0.8815747
```

Note: **in most econometric software, you need to pass weights, which are inversely proportional to the variances**. In other words you need to supply weights $= 1/\sigma_i^2 = 1/\omega_i^2$ which are the diagonal elements of $\mathbf{\Omega}^{-1}$, **not** $\mathbf{\Psi}$ (regardless of your specification of $\mathbf{\Sigma}$). Then, the software automatically takes the square root of the specified values when calculating.

Using built-in functions:

```
mdl_wls <- lm(y ~ x1 + x2, data = data_mat, weights =  1 / h_est)
print(round(coef(summary(mdl_wls)), 5))
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -1.39237   10.19482 -0.13658  0.89151
## x1           7.97645    4.00325  1.99249  0.04770
## x2          -2.02459    0.88157 -2.29656  0.02270
```

We see that we get identical results in all cases.

Finally, we would like to compare the **residuals from the OLS and GLS procedures**. If we have indeed accounted for heteroskedasticity, then the residual plots should indicate so as well. We note that the GLS estimates are for the model:

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\beta + \varepsilon^*, \quad \text{where } \boldsymbol{Y}^* = \boldsymbol{\Psi}^\top \boldsymbol{Y}, \quad \boldsymbol{X}^* = \boldsymbol{\Psi}^\top \boldsymbol{X}, \quad \varepsilon^* = \boldsymbol{\Psi}^\top \varepsilon$$

In other words we need to calculate: $\widehat{\varepsilon}^*_{WLS} = \boldsymbol{Y}^* - \boldsymbol{X}^*\widehat{\beta}_{WLS}$.

In most software, using built-in WLS estimation, the residuals are calculated as: $\widehat{\varepsilon}_{WLS} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}_{WLS}$.

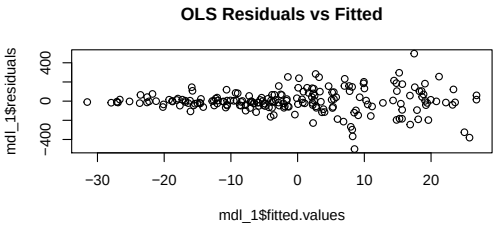
Consequently, since we have specified $\boldsymbol{\Psi}^\top = \boldsymbol{\Psi} = \text{diag}\left(1/\sqrt{\widehat{h}_1}, ..., 1/\sqrt{\widehat{h}_N}\right)$, we can calculate the residuals for the transformed model as:

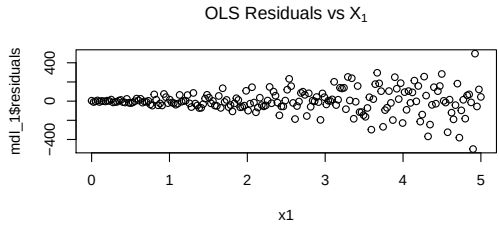
$$\widehat{\varepsilon^*}_{WLS} = \boldsymbol{\Psi}^\top \widehat{\varepsilon}_{WLS} = \text{diag}\left(1/\sqrt{\widehat{h}_1}, ..., 1/\sqrt{\widehat{h}_N}\right) \widehat{\varepsilon}_{WLS}$$

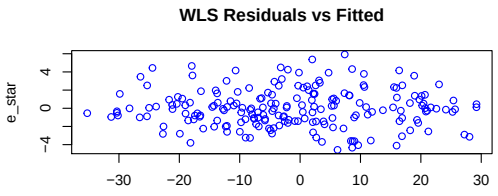
```
e_star <- 1 / sqrt(h_est) * mdl_wls$residuals
```

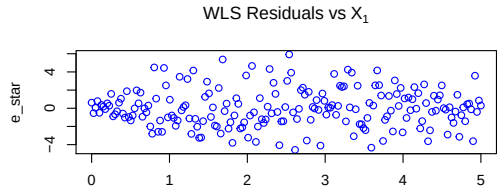
Now we can compare the plots:

```
par(mfrow = c(2, 2))
#
plot(mdl_1$fitted.values, mdl_1$residuals, main = "OLS Residuals vs Fitted")
plot(x1, mdl_1$residuals, main = bquote("OLS Residuals vs"~X[1]))
plot(mdl_wls$fitted.values, e_star, col = "blue", main = "WLS Residuals vs Fitted")
plot(x1, e_star, col = "blue", main = bquote("WLS Residuals vs"~X[1]))
```



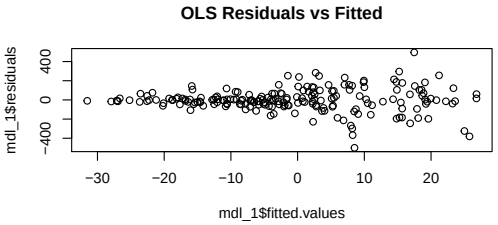



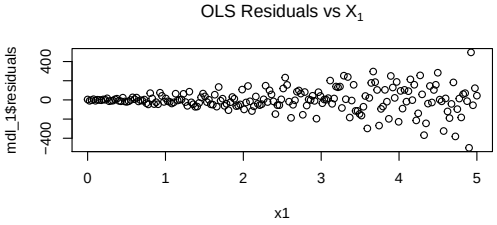
We see that:

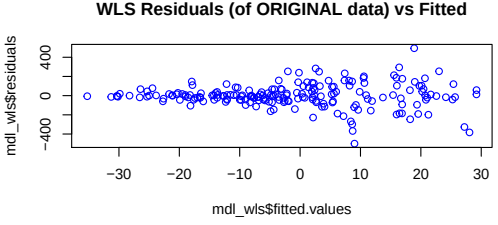
- the magnitude of the residuals is smaller;
- there still appears to be some (albeit possibly insignificant) heteroskedasticity.

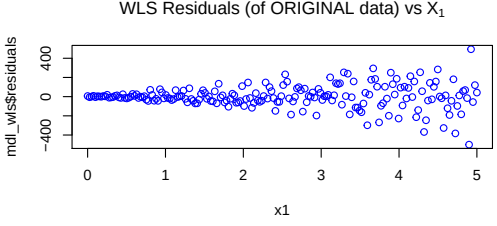
What would have happened, if we were to plot the WLS residuals **of the non-transformed data**?

```r
par(mfrow = c(2, 2))
#
plot(mdl_1$fitted.values, mdl_1$residuals, main = "OLS Residuals vs Fitted")
plot(x1, mdl_1$residuals, main = bquote("OLS Residuals vs"~X[1]))
plot(mdl_wls$fitted.values, mdl_wls$residuals, col = "blue", main = "WLS Residuals (of ORIGINAL data)
plot(x1, mdl_wls$residuals, col = "blue", main = bquote("WLS Residuals (of ORIGINAL data) vs"~X[1]))
```



Looking at the original, untransformed, data it would **appear** that we did not account for heteroskedasticity but this is not the case. As mentioned, **we have created a model on the transformed data, hence we should analyse the residuals of the transformed data**.

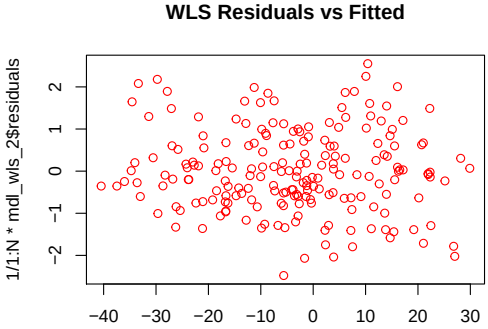
As mentioned before - we have attempted to approximately calculate the weights, using the logarithms of the squared residuals. Therefore, we may not always be able to capture all of the heteroskedasticity. Nevertheless, it is much better than it was initially.

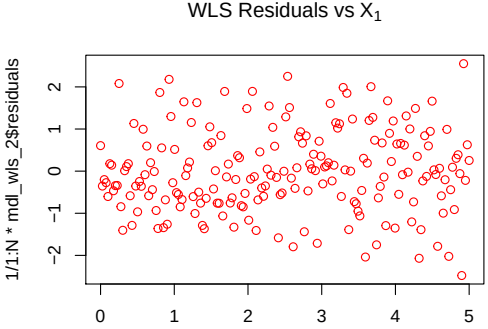
On the other hand, if we were to use $1/i$ as the weights, instead of $1/\sqrt{\widehat{h}_i}$:

```
mdl_wls_2 <- lm(y ~ x1 + x2, data = data_mat, weights =  1 / (1:N)^2)
print(round(coef(summary(mdl_wls_2)), 5))
```

```
##             Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 15.34254    2.37257   6.46663  0.00000
## x1           5.33401    2.90696   1.83491  0.06803
## x2          -3.32553    0.22286 -14.92183  0.00000
```

```
par(mfrow = c(1, 2))
plot(mdl_wls_2$fitted.values, 1 / 1:N * mdl_wls_2$residuals, col = "red", main = "WLS Residuals vs Fi
plot(x1, 1 / 1:N * mdl_wls_2$residuals, col = "red", main = bquote("WLS Residuals vs"~X[1]))
```



The resulting errors of the transformed model are no longer heteroskedastic.

Unfortunately, if we would have used $1/\sqrt{X_{1,i}}$, or $1/X_{1,i}$, then the residuals would have likely remained heteroskedastic.

This is an important conclusion regarding WLS (and FGLS in general): if we attempt to approximate the results, then, no matter the approximation, there is still a chance that the WLS residuals would still contain some kind of (weak) heteroskedasticity. As such, we should examine different residual plots to determine what kind of weights are appropriate.

*In this example* the true (unknown) variance of the error term is $\mathbb{V}\mathrm{ar}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right) = \boldsymbol{\Sigma} = \sigma_{\epsilon}^2 \cdot \mathrm{diag}(1^2, 2^2, ..., N^2)$. This would be equivalent to the case where the $i$-th response is an aggregated total of $N_i = i^2$ observations.

Regarding $R^2$:
In most econometric software, the coefficient of determination for WLS uses a **weighted mean** when calculating **TSS**. As a result, the reported $R^2$ now measures the proportion of total variation in **weighted** $Y$, $Y^*$ explained by the **weighted** X, $X^*$.
To get a more conventional expression of $R^2$ - use the general (or pseudo-) expression:

$$R_g^2 = \mathbb{C}\mathrm{orr}(Y, \widehat{Y})^2$$

where $\widehat{Y}$ are the fitted values of the **original** (i.e. non-weighted) dependent variable).

Note that the weighted mean is calculated as:

$$\overline{Y}^{(w)} = \frac{\sum_{i=1}^{N} Y_i/h_i}{\sum_{i=1}^{N} 1/h_i}$$

We can verify this by examining the manual results with the output:

```
w <- 1 / h_est
RSS_W <- sum(w * mdl_wls$residuals^2)
TSS_W <- sum(w * (y - weighted.mean(y, w))^2)
print(1 - RSS_W / TSS_W)
```

```
## [1] 0.04576885
```

```
print(summary(mdl_wls)$r.squared)
```

```
## [1] 0.04576885
```

```
r2g_ols <- cor(y, mdl_1$fitted)^2
r2g_wls <- cor(y, mdl_wls$fitted)^2
print(paste0("OLS pseudo-R^2 = ", r2g_ols))
```

```
## [1] "OLS pseudo-R^2 = 0.011880153644293"
```

```
print(paste0("WLS pseudo-R^2 = ", r2g_wls))
```

```
## [1] "WLS pseudo-R^2 = 0.0118801535929629"
```

As was mentioned before regarding $R^2$ - it is not always a good measure of the **overall model adequacy**. Even if the OLS $R^2$ is higher, if the residuals do not conform to (MR.2) - (MR.6) assumptions, then the model and its hypothesis test results are not valid.

## Choosing between HCE and WLS

The generalized least squares estimator require that we know the underlying form of the variance-covariance matrix.

Regarding **HCE**:

- The variance estimator is quite **robust** because **it is valid whether heteroskedasticity is present or not**, but only in a matter that is appropriate asymptotically.
- In other words, if we are not sure whether the random errors are heteroskedastic or homoskedastic, then we can use a robust variance estimator and be confident that our standard errors, $t$-tests, and interval estimates are valid **in large samples.**
- In **small samples**, whether we modify the covariance estimator or not, the usual statistics will still be **unreliable**.
- This estimator needs to be modified, if we suspect that the errors may exhibit **autocorrelation (of some unknown form)**.

Regarding **WLS**:

- If we know the underlying form of the residual variance-covariance matrix, then FGLS would result in more efficient estimators;
- If we specify the covariance structure incorrectly, i.e. we do not completely remove heteroskedasticity, then the FGLS estimator will be unbiased, but not the best and the standard errors will still be biased (as in the OLS case).

So, both methods have their benefits and their drawbacks. However, we can **combine them both**:

- Attempt to correct for heteroskedasticity via the WLS;
- Test the residuals from WLS for heteroskedasticity;
- If the WLS residuals are homoskedastic - the WLS has improved the precision of our model;
- If the WLS residuals are heteroskedastic - we may use HCE on the WLS residuals.
- This way we protect ourselves from a possible misspecification of the unknown variance-covariance structure.

## Monte Carlo Simulation: OLS vs FGLS

To illustrate the effects of heteroskedasticity on the standard errors of the estimates, and efficiency between OLS, GLS and FGLS, we will carry out a Monte Carlo simulation. We will simulate the following model:

$$Y_i^{(m)} = \beta_0 + \beta_1 X_{1,i}^{(m)} + \beta_2 X_{2,i}^{(m)} + \epsilon_i^{(m)}, \quad \epsilon_i^{(m)} \sim \mathcal{N}(0, i^2 \cdot \sigma^2), \quad i = 1, ..., N, \quad m = 1, ..., MC$$

We will simulate a total of $MC = 1000$ samples from this model with specific coefficients and estimate the parameters via OLS, WLS, as well as correct the standard errors of OLS via HCE. We will do so with the following code:

```
set.seed(123)
# Number of simulations:
MC <- 1000
# Fixed parameters
N <- 100
beta_vec <- c(10, 5, -3)
# matrix of parameter estimates for each sample:
beta_est_ols  <- NULL
beta_pval_ols <- NULL
beta_pval_hce <- NULL
#
beta_est_wls  <- NULL
beta_pval_wls <- NULL
#
beta_est_gls  <- NULL
beta_pval_gls <- NULL
```

```r
for(i in 1:MC){
  # simulate the data:
  x1 <- seq(from = 0, to = 5, length.out = N)
  x2 <- sample(seq(from = 3, to = 17, length.out = 80), size = N, replace = TRUE)
  e  <- rnorm(mean = 0, sd = 1:N, n = N)
  y  <- cbind(1, x1, x2) %*% beta_vec + e
  data_mat <- data.frame(y, x1, x2)
  # estimate via OLS
  mdl_0 <- lm(y ~ x1 + x2, data = data_mat)
  # correct OLS se's:
  mdl_hce <- lmtest::coeftest(mdl_0, vcov. = sandwich::vcovHC(mdl_0, type = "HC0"))
  # estimate via WLS
  resid_data <- data.frame(log_e2 = log(mdl_0$residuals^2), x1, x2)
  h_est   <- exp(lm(log_e2 ~ x1 + x2, data = resid_data)$fitted.values)
  mdl_wls <- lm(y ~ x1 + x2, data = data_mat, weights =  1 / h_est)
  # estimate via GLS (by knowing the true covariance matrix)
  mdl_gls <- lm(y ~ x1 + x2, data = data_mat, weights =  1 / (1:N)^2)
  # Save the estimates from each sample:
  beta_est_ols  <- rbind(beta_est_ols, coef(summary(mdl_0))[, 1])
  beta_est_wls  <- rbind(beta_est_wls, coef(summary(mdl_wls))[, 1])
  beta_est_gls  <- rbind(beta_est_gls, coef(summary(mdl_gls))[, 1])
  # Save the coefficient p-values from each sample:
  beta_pval_ols <- rbind(beta_pval_ols, coef(summary(mdl_0))[, 4])
  beta_pval_hce <- rbind(beta_pval_hce, mdl_hce[, 4])
  beta_pval_wls <- rbind(beta_pval_wls, coef(summary(mdl_wls))[, 4])
  beta_pval_gls <- rbind(beta_pval_gls, coef(summary(mdl_gls))[, 4])
}
```

Firstly, it is interesting to see how many times we would have **rejected the null hypothesis that a parameter is insignificant** with significance level $\alpha = 0.05$. We will divide the number of times that we would have rejected the null hypothesis by the total number of samples $MC$ to get the rejection rate:

```
alpha = 0.05
a1 <- colSums(beta_pval_ols < alpha) / MC
a2 <- colSums(beta_pval_hce < alpha) / MC
a3 <- colSums(beta_pval_wls < alpha) / MC
a4 <- colSums(beta_pval_gls < alpha) / MC
#
a <- t(data.frame(a1, a2, a3, a4))
colnames(a) <- names(a1)
rownames(a) <- c("OLS: H0 rejection rate", "HCE: H0 rejection rate", "WLS: H0 rejection rate", "GLS:
print(a)
```

```
##                          (Intercept)    x1    x2
## OLS: H0 rejection rate         0.062 0.254 0.569
## HCE: H0 rejection rate         0.109 0.216 0.584
## WLS: H0 rejection rate         0.195 0.369 0.982
## GLS: H0 rejection rate         0.856 0.605 1.000
```

- We see that we would have rejected the null hypothesis that $X_2$ is insignificant in around 55.5% of the simulated samples with OLS.
- If we were to correct for heteroskedasticity, this would have increased to 57.4%.
- On the other hand, if we were to re-estimate the model with WLS, then we would have rejected the null hypothesis that $X_2$ is insignificant around in around 97.6% of the simulated samples.

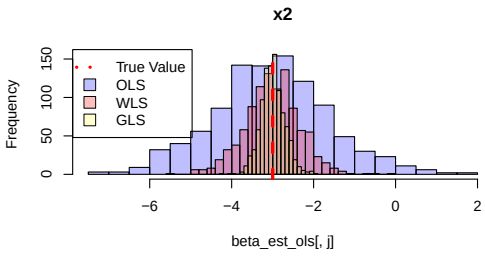
One possible explanation for the relatively poor performance of HCE is the fact that it uses $\widehat{\mathbf{\Sigma}} = \text{diag}(\widehat{\epsilon}_1^2, ..., \widehat{\epsilon}_N^2)$ for the covariance matrix. In this case it is clearly inferior to the covariance matrix specification used in WLS. On the other hand, as we have already mentioned - HCE are only useful in large samples.

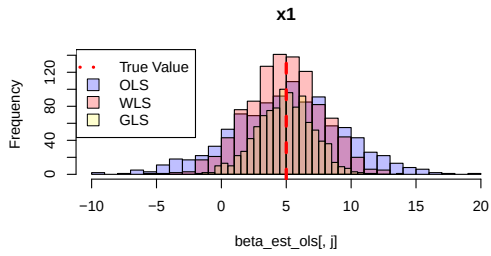
Because of the large variance of $\epsilon_i$, we see that we would often not reject the null hypothesis that $X_1$ is insignificant. On the other hand, as we can see, using WLS reduces this risk significantly.

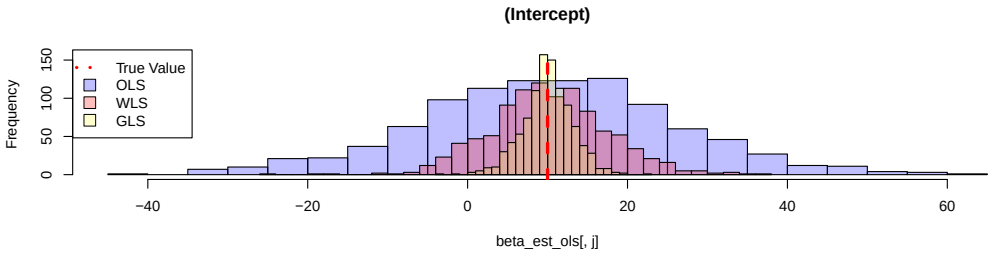
Finally, if we were to know the **true covariance structure** - we could incorporate GLS - this would mean that we would almost never reject the null hypothesis that the coefficient of $X_2$ is insignificant! Furthermore, the rejection rate of the null hypothesis for $\beta_0$ and $\beta_1$ are also significantly larger.

Unfortunately, in empirical applications we will never know the true covariance structure, so the GLS results are only presented as the *best* case scenario, which we would hope to achieve with FGLS.

We can look at the coefficient estimate histograms:



- OLS estimates are less efficient than WLS in terms of the estimate variance.
- On the other hand, both estimates are unbiased - their average is equal to the true parameter value. - Consequently, because OLS estimators are less efficient, we would need a larger sample to achieve a similar level of precision as the WLS.

As mentioned before, GLS is the best case scenario, when we *exactly* know the true covariance structure and **do not need to estimate it**.

# Autocorrelated (Serially Correlated) Errors

# Autocorrelated (Serially Correlated) Errors

Consider the case where assumption (MR.4) **does not hold**, but assumption (MR.3) (and the other remaining assumptions (MR.1), (MR.2), (MR.5) and, optionally, (MR.6)) **are still valid**. Then we can write the following model as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{0}, \quad \mathbb{V}\mathrm{ar}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\top}\right) = \boldsymbol{\Sigma} \neq \sigma_{\epsilon}^{2}\boldsymbol{I}$$

The case when the error variance-covariance matrix is no longer diagonal, but with equal diagonal elements, expressed as:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma_{1,2} & \sigma_{1,3} & \dots & \sigma_{1,N} \\ \sigma_{2,1} & \sigma^2 & \sigma_{2,3} & \dots & \sigma_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{N,1} & \sigma_{N,2} & \sigma_{N3,} & \dots & \sigma^2 \end{bmatrix}, \quad \sigma_{i,j} = \sigma_{j,i} = \mathbb{C}\mathrm{ov}(\epsilon_i, \epsilon_j) \neq 0, \quad i, j = 1, ..., N$$

is referred to as **autocorrelation** (or **serial correlation**). Just like before with heteroskedasticity:

- the OLS estimators remain unbiased and consistent;
- OLS estimators are no longer efficient;
- the variance estimator of the OLS estimates is biased and inconsistent;
- $t$-statistics of the OLS estimates are invalid.

It should be stressed that serial correlation is usually present in **time-series** data. For **cross-sectional** data, the errors may be correlated in terms of social, or geographical distance. For example, the distance between cities, towns, neighborhoods, etc.

1. Assume that $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ is the true model.
2. Create a model for the OLS residuals:

$$\widehat{\epsilon_i} = \alpha_0 + \alpha_1 X_{1,i} + ... + \alpha_k X_{k,i} + \rho_1 \widehat{\epsilon}_{i-1} + \rho_2 \widehat{\epsilon}_{i-2} + ... + \rho_p \widehat{\epsilon}_{i-p} + u_t$$

3. Test the null hypothesis that **the residuals are serially correlated**: create a model on the OLS residuals

$$H_0 : \rho_1 = \rho_2 = ... = \rho_p = 0$$

4. If we fail to reject the null hypothesis - we can use the usual OLS estimators.
5. If we reject the null hypothesis, there are two ways we can go:
   - Use the OLS estimators, but correct their variance estimators (i.e. make them consistent);
   - Instead of OLS, use FGLS (and its variations).
   - Attempt to specify a different model, which would hopefully, be able to account for serial correlation (autocorrelation may be the cause of a misspecified model).

## Example

We will simulate the following model:

$$\begin{cases} Y_i &= \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i \\ \epsilon_i &= \rho \epsilon_{i-1} + u_i, \ |\rho| < 1, \ u_i \sim \mathcal{N}(0, \sigma^2), \quad \epsilon_0 = 0 \end{cases}$$

```
set.seed(123)
#
N <- 200
beta_vec <- c(10, 5, -3)
rho <- 0.8
#
x1 <- seq(from = 0, to = 5, length.out = N)
x2 <- sample(seq(from = 3, to = 17, length.out = 80), size = N, replace = TRUE)
# serially correlated residuals:
ee <- rnorm(mean = 0, sd = 3, n = N)
for(i in 2:N){
  ee[i] <- rho * ee[i-1] + ee[i]
}
#
x_mat <- cbind(1, x1, x2)
y <- x_mat %*% beta_vec + ee
data_mat <- data.frame(y, x1, x2)
```

## Testing For Serial Correlation

We can examine the presence of autocorrelation from the residuals plots, as well as conducting a number of formal tests.
We will begin by estimating our model via OLS, as we usually would.

```
mdl_1 <- lm(y ~ x1 + x2, data = data_mat)
print(round(coef(summary(mdl_1)), 5))
```

```
##             Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 10.61601    0.97983  10.83457        0
## x1           4.89018    0.20457  23.90497        0
## x2          -2.93686    0.07545 -38.92241        0
```

## Residual Correlogram

We begin by calculating the residuals from the OLS model:

$$\widehat{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}_{\text{OLS}}$$

we will use these residuals as estimates for the true (unobserved) error term $\varepsilon$ and examine their autocorrelations. Note that from the OLS structure it holds true that $\mathbb{E}(\widehat{\varepsilon}|\boldsymbol{X}) = \boldsymbol{0}$. So, assume that we want to calculate the sample correlation between $\widehat{\epsilon}_i$ and $\widehat{\epsilon}_{i-k}$ - we want to calculate the autocorrelation at **lag k**:

$$\widehat{\rho}(k) = \widehat{\mathbb{C}\text{orr}}(\widehat{\epsilon}_i, \widehat{\epsilon}_{i-k}) = \frac{\widehat{\mathbb{C}\text{ov}}(\widehat{\epsilon}_i, \widehat{\epsilon}_{i-k})}{\sqrt{\widehat{\mathbb{V}\text{ar}}(\widehat{\epsilon}_i)}\sqrt{\widehat{\mathbb{V}\text{ar}}(\widehat{\epsilon}_{i-k})}} = \frac{\sum_{i=k+1}^{N} \widehat{\epsilon}_i \widehat{\epsilon}_{i-k}}{\sum_{i=1}^{N} \widehat{\epsilon}_i^2}$$

Furthermore, assume that we are interested in testing whether the sample (auto)correlation $\widehat{\rho}(k)$ is significantly different from zero:

$$H_0 : \widehat{\rho}(k) = 0$$
$$H_1 : \widehat{\rho}(k) \neq 0$$

Under the null hypothesis it holds true that $\widehat{\rho}(k) \overset{a}{\sim} \mathcal{N}(0, 1/N)$, where $\overset{a}{\sim}$ indicates asymptotic distribution - the distribution as the sample size $N \to \infty$. In this case, it means that for large samples the distribution is approximately normal. Consequently, a suitable statistic can be constructed as:

$$Z = \frac{\widehat{\rho}(k) - 0}{\sqrt{1/N}} = \widehat{\rho}(k) \cdot \sqrt{N} \overset{a}{\sim} \mathcal{N}(0, 1)$$

So, at a 5% significance level, the critical value is $Z_c \approx 1.96$. In this case, w would reject the null hypothesis when:

$$\widehat{\rho}(k) \cdot \sqrt{N} \geq 1.96, \text{ or } \widehat{\rho}(k) \cdot \sqrt{N} \leq 1.96$$

or alternatively, if we want to have the notation similar to "estimate $\pm$ critical value $\cdot$ se(estimate)", we can write it as:

$$\widehat{\rho}(k) \geq 1.96/\sqrt{N}, \text{ or } \widehat{\rho}(k) \leq 1.96/\sqrt{N}$$

As a rule of thumb, we can sometimes take 2 instead of 1.96. We can also calculate this via software:

```
z_c <- qnorm(p = 1 - 0.05 / 2, mean = 0, sd = 1)
print(z_c)
```

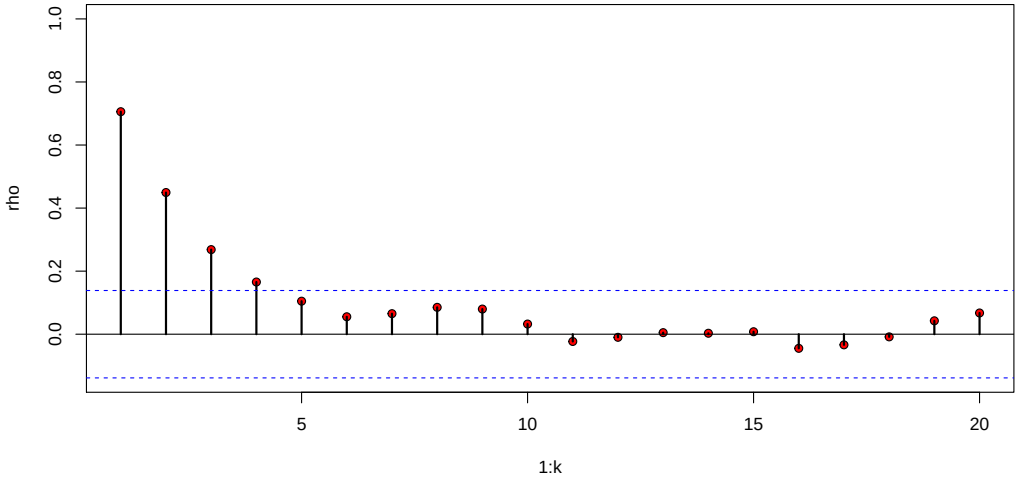
```
## [1] 1.959964
```

We will begin by manually calculating the autocorrelations and their confidence bounds for $k = 1, ..., 20$:

```
rho <- NULL
e <- mdl_1$residuals
for(k in 1:20){
  r <- sum(e[-c(1:k)] * e[1:(N-k)]) / sum(e^2)
  rho <- c(rho, r)
}
rho_lower <- - z_c / sqrt(N)
rho_upper <- + z_c / sqrt(N)
```

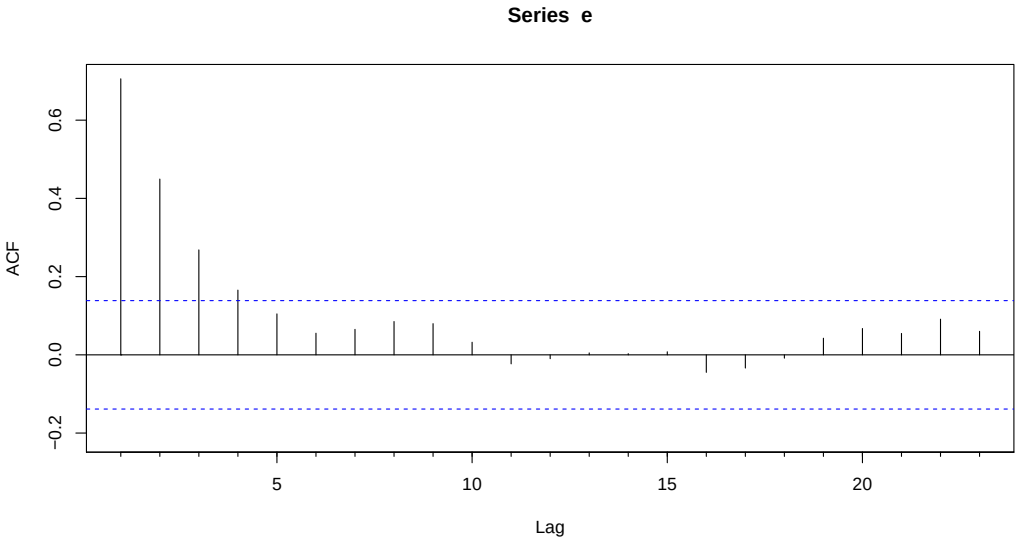
and we can plot them:

```
plot(1:k, rho, ylim = c(min(rho_lower), 1), pch = 21, bg = "red")
segments(x0 = 1:k, y0 = rho, x1 = 1:k, y1 = 0, lwd = 2)
abline(h = 0)
# Draw the confidence bounds:
abline(h = rho_upper, lty = 2, col = "blue")
abline(h = rho_lower, lty = 2, col = "blue")
```


rho
1.0
0.8
0.6
0.4
0.2
0.0
5
10
15
20
1:k

Alternatively, we can use the built-in functions:

```
forecast::Acf(e)
```



We can see from the plots that there are some sample autocorrelations $\widehat{\rho}(k)$, which are statistically significantly different from zero (i.e. their values are *above* the blue horizontal line).

As with most visual diagnostics tools - it may not always be a clear-cut answer from the plots alone. So, we can apply a number of different statistical tests to check whether there are statistically significant sample correlations.

## Autocorrelation Tests

The tests are identical to the ones described in the univariate regression, but re-visited for the multiple regression model case. These tests will be re-examined when dealing with time-series data models.

## Durbin-Watson Test

The **DW** test is used to test the hypothesis that the residuals are serially correlated at lag 1, i.e. that in the following model:

$$\epsilon_i = \rho\epsilon_{i-1} + v_i$$

the hypothesis being tested is:

$$\begin{cases} H_0 : \rho = 0 & \text{(no serial correlation)} \\ H_1 : \rho \neq 0 & \text{(serial correlation at lag 1)} \end{cases}$$

Since we do not observe the true error terms, we use the OLS residuals $\widehat{\epsilon}_i$ and calculate the **Durbin-Watson** statistic as:

$$DW = \frac{\sum_{i=2}^{N}(\widehat{\epsilon}_i - \widehat{\epsilon}_{i-1})^2}{\sum_{i=1}^{N}\widehat{\epsilon}_i^2}$$

The DW test statistics critical values may not be available in some econometric software. Furthermore, its distribution no longer holds, when the equation of $Y_i$ contains a lagged **dependent variable**, $Y_{i-1}$.
As a quick rule of thumb, if the DW statistic is near 2, then we do not reject the null hypothesis of no serial correlation.

If we assume that $\sum_{i=2}^{N} \widehat{\epsilon}_i^2 \approx \sum_{i=2}^{N} \widehat{\epsilon}_{i-1}^2$, then we can re-write the DW statistic as:

$$DW = \frac{\sum_{i=2}^{N} \widehat{\epsilon}_i^2 - 2\sum_{i=2}^{N} \widehat{\epsilon_i}\widehat{\epsilon}_{i-1} + \sum_{i=2}^{N} \widehat{\epsilon}_{i-1}^2}{\sum_{i=1}^{N} \widehat{\epsilon}_i^2} \approx \frac{2\left[\sum_{i=2}^{N} \widehat{\epsilon}_i^2 - \sum_{i=2}^{N} \widehat{\epsilon_i}\widehat{\epsilon}_{i-1}\right]}{\sum_{i=1}^{N} \widehat{\epsilon}_i^2} = 2\left[1 - \widehat{\rho}(1)\right] = \begin{cases} 4, & \text{if } \widehat{\rho}(1) = -1 \\ 2, & \text{if } \widehat{\rho}(1) = 0 \\ 0, & \text{if } \widehat{\rho}(1) = 1 \end{cases}$$

which helps in understanding why we expect the DW statistic to be *close* to 2 under the null hypothesis.

```
print(lmtest::dwtest(mdl_1, alternative = "two.sided"))
```

```
##
##  Durbin-Watson test
##
## data:  mdl_1
## DW = 0.56637, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is not 0
```

because $p - value < 0.05$, we reject the null hypothesis at the 5% significance level and conclude that the residuals are serially correlated at lag order 1.

# Breusch-Godfrey (BG or LM) Test

The **BG** test can be applied to a model, with a lagged response variable on the right-hand side, for example: $Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,i} + \beta_{k+1} Y_{i-1} + \epsilon_i$.
In general, we estimate the model parameters via OLS and calculate the residuals:

$$\widehat{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}$$

where $\boldsymbol{X}$ can contain $X_{1,i}, ..., X_{k,i}, Y_{i-1}$ for $i = 1, ..., N$.
Then, we estimate an auxiliary regression on $\widehat{\varepsilon}$ as:

$$\widehat{\epsilon_i} = \alpha_0 + \alpha_1 X_{1,i} + ... + \alpha_k X_{k,i} + \rho_i \widehat{\epsilon_{i-1}} + ... + \rho_p \widehat{\epsilon_{i-p}} + u_i$$

(Note: if we included $Y_{i-1}$ in the initial regression, then we need to also include them in the auxiliary regression).
We want to test the null hypothesis that the lagged residual coefficients are insignificant:

$$\begin{cases} H_0 : \rho_1 = ... = \rho_p = 0 \\ H_1 : \rho_j \neq 0, \text{ for some } j \end{cases}$$

We can either carry out an $F$-test, or a Chi-squared test:

$$LM = (N - p) \underset{\epsilon}{R^2} \sim \chi^2_p$$

where $\underset{\epsilon}{R^2}$ is the $R$-squared from the auxiliary regression on $\widehat{\varepsilon}$.
The **BG** test is a more general test compared to **DW**

Looking at the correlogram plots Let's test the null hypothesis that at least on of the first three lags of the residuals is not zero, i.e. $p = 3$:

```
bg_t <- lmtest::bgtest(mdl_1, order = 3)
print(bg_t)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 3
##
## data:  mdl_1
## LM test = 100.72, df = 3, p-value < 2.2e-16
```

Since the $p - value < 0.05$, we reject the null hypothesis and conclude that there is serial correlation in the residuals.

## Heteroskedasticity-and-Autocorrelation-Consistent Standard Errors (HAC)

So far, we have assumed that the diagonal elements of $\mathbf{\Sigma}$ are constant - i.e. that the residuals are serially correlated but homoskedastic. We can further generalize this for the case of **heteroskedastic serially correlated standard errors**:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \dots & \sigma_{1,N} \\ \sigma_{2,1} & \sigma_2^2 & \sigma_{2,3} & \dots & \sigma_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{N,1} & \sigma_{N,2} & \sigma_{N3,} & \dots & \sigma_N^2 \end{bmatrix}, \quad \sigma_{i,j} = \sigma_{j,i} = \mathbb{C}\text{ov}(\epsilon_i, \epsilon_j) \neq 0, \quad i, j = 1, ..., N$$

Similarly to dealing with heteroskedasticity, we can use OLS estimator and correct the standard errors. In this case the corrected standard errors are known as **HAC** (**H**eteroskedasticity-and-**A**utocorrelation-**C**onsistent) **standard errors**, or **Newey–West standard errors**.

The White covariance matrix assumes serially uncorrelated residuals. On the other hand, the Newey-West proposed a more general covariance estimator, which is robust to both heteroskedasticity and autocorrelation, **of the residuals of unknown covariance form**. The HAC coefficient covariance estimator handles autocorrelation with lags up to $p$ - it is assumed that lags larger than $p$ are insignificant and thus can be ignored. It is defined as:

$$HAC_{\widehat{\beta}_{OLS}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} N\widehat{\boldsymbol{\Sigma}} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

where:

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Gamma}}(0) + \sum_{j=1}^{p} \left[1 - \frac{j}{p+1}\right] \left[\widehat{\boldsymbol{\Gamma}}(j) + \widehat{\boldsymbol{\Gamma}}(-j)\right]$$

$$\widehat{\boldsymbol{\Gamma}}(j) = \frac{1}{N} \left( \sum_{i=1}^{N} \widehat{\epsilon_i}\widehat{\epsilon_{i-j}} \boldsymbol{X}_i \boldsymbol{X}_{i-j}^\top \right), \quad \boldsymbol{X}_i = \left[1,\ X_{1,i},\ X_{2,i},\ ...,\ X_{k,i}\right]^\top$$

In the absence of serial correlation we would have that:

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Gamma}}(0)$$

which is equivalent to the White Estimators (HCE).
Note: HAC not only corrects for autocorrelation, but also for heteroskedasticity.
Do not be alarmed if you see slightly different HAC standard errors in different statistical programs - there are a number of different variations of $\widehat{\boldsymbol{\Sigma}}$.

Using the built-in functions we have that:

```
#V_HAC <- sandwich::vcovHAC(mdl_1)
V_HAC <- sandwich::NeweyWest(mdl_1, lag = 1)
print(V_HAC)
```

```
##             (Intercept)           x1           x2
## (Intercept)  1.43560709 -0.303808295 -0.043572462
## x1          -0.30380830  0.183608167 -0.002018108
## x2          -0.04357246 -0.002018108  0.004421496
```

Following the documentation, `NeweyWest()` is a convenience interface to `vcovHAC()` using Bartlett kernel weights. In comparison `vcovHAC()` allows choosing weights as either `weightsAndrews`, or `weightsLumley`, or a custom function to calculate the weights.

Then the coefficients, their standard errors and $p$-values can be summarized as:

```
# print(lmtest::coeftest(mdl_1, sandwich::vcovHAC(mdl_1)))
print(lmtest::coeftest(mdl_1, sandwich::NeweyWest(mdl_1, lag = 1))[, ])
```

```
##              Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) 10.616005 1.19816822   8.860196  4.719858e-16
## x1           4.890185 0.42849524  11.412460  1.741811e-23
## x2          -2.936860 0.06649433 -44.167068 3.799730e-104
```

Compared with the biased residuals in the OLS output - HAC standard errors are somewhat larger:

```
print(coef(summary(mdl_1)))
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 10.616005 0.97982714  10.83457 9.311699e-22
## x1           4.890185 0.20456771  23.90497 3.887306e-60
## x2          -2.936860 0.07545421 -38.92241 1.933624e-94
```

## Feasible GLS - Cochrane-Orcutt Procedure (CORC)

Alternatively to HAC, we can make some assumptions regarding the nature of autocorrelation and employ a more efficient GLS estimator.
For the case, when the residuals are serially correlated at lag 1, but not heteroskedastic:

$$\epsilon_i = \rho\epsilon_{i-1} + u_i, \;\; |\rho| < 1, \;\; u_i \sim \mathcal{N}(0, \sigma^2), \quad i \in \{0, \pm 1, \pm 2, ...\}$$

we note that:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\epsilon_i) &= \mathbb{V}\mathrm{ar}(\rho\epsilon_{i-1} + u_i) = \mathbb{V}\mathrm{ar}(\rho(\rho\epsilon_{i-2} + u_{i-1}) + u_i) \\
&= \mathbb{V}\mathrm{ar}(\rho^2(\rho\epsilon_{i-3} + u_{i-2}) + u_i + \rho u_{i-1}) \\
&= ... \\
&= \mathbb{V}\mathrm{ar}\left(\sum_{j=0}^{\infty} \rho^j u_{i-j}\right) = \sum_{j=0}^{\infty} \rho^{2j} \cdot \mathbb{V}\mathrm{ar}\left(u_{i-j}\right) \\
&= \frac{\sigma^2}{1-\rho^2}
\end{aligned}$$

and:

$$\begin{aligned}
\mathbb{C}\text{ov}(\epsilon_i,\ \epsilon_{i-k}) &= \mathbb{C}\text{ov}(\rho(\rho\epsilon_{i-2} + u_{i-1}) + u_i,\ \epsilon_{i-k}) \\
&= \mathbb{C}\text{ov}(\rho^2(\rho\epsilon_{i-3} + u_{i-2}) + \rho u_{i-1},\ \epsilon_{i-k}) \\
&= ... \\
&= \mathbb{C}\text{ov}(\rho^k\epsilon_{i-k},\ \epsilon_{i-k}) \\
&= \rho^k\mathbb{C}\text{ov}(\epsilon_{i-k},\ \epsilon_{i-k}) \\
&= \rho^k\sigma^2, \qquad \text{since } \mathbb{C}\text{ov}(u_i, \epsilon_j) = 0, i \neq j
\end{aligned}$$

Consequently, we can re-write the covariance matrix as:

$$\mathbf{\Sigma} = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & ... & \rho^{N-1} \\ \rho & 1 & \rho & ... & \rho^{N-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & ... & 1 \end{bmatrix}$$

In this case, the knowledge of parameter $\rho$ allows us to empirically apply the GLS - as FGLS.

**Cochrane–Orcutt (CORC) estimator:**

1. Estimate $\beta$ via OLS from:
$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$
and calculate the residuals $\widehat{\varepsilon}^{(1)} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}$, where (1) denotes the first iteration.
2. Estimate the following residual regression via OLS:
$$\widehat{\epsilon}_i^{(1)} = \rho\widehat{\epsilon}_{i-1}^{(1)} + \widehat{u}_i$$
and obtain $\widehat{\rho}^{(1)}$.
3. Calculate the following transformed variables:
$$Y_i^* = Y_i - \widehat{\rho}^{(1)} Y_{i-1}^*, \qquad \boldsymbol{X}_i^* = \boldsymbol{X}_i - \widehat{\rho}^{(1)}\boldsymbol{X}_{i-1}^*, \text{ where } \boldsymbol{X}_i = \left[1,\ X_{1,i},\ X_{2,i},\ ...,\ X_{k,i}\right]^\top$$
4. Apply OLS to the following model:
$$(Y_i - \widehat{\rho}^{(1)} Y_{i-1}) = \beta_0(1 - \widehat{\rho}^{(1)}) + \beta_1(X_{1,i} - \widehat{\rho}^{(1)}) + ... + \beta_k(X_{k,i} - \widehat{\rho}^{(1)}) + u_i$$
or, more compactly:
$$\boldsymbol{Y}^* = \boldsymbol{X}^*\beta + \boldsymbol{u}$$
to get the OLS estimates $\widetilde{\beta}$.
5. Having estimated the model, calculate the residuals on the non-transformed data: $\widehat{\varepsilon}^{(2)} = \boldsymbol{Y} - \boldsymbol{X}\widetilde{\beta}$ and go to step (2).

Repeat this procedure until $\widehat{\rho}$ converges: if the change in $\widehat{\rho}^{(K)}$, compared to the previous iteration $\widehat{\rho}^{(K-1)}$ is no more than $\Delta$ (for example $\Delta = 0.001$) - stop the procedure.

Note on step (4): if we decide to use a column of ones for the constant $\beta_0$, then we will actually have estimated $\widetilde{\beta}_0^* = \widetilde{\beta}_0(1-\widehat{\rho}^{(1)})$ and will need to transform the intercept term as $\widetilde{\beta}_0 = \widehat{\beta}_0^*/(1-\widehat{\rho}^{(1)})$. On the other hand, if we transform the intercept column to have $1-\widehat{\rho}^{(1)}$ in the design matrix instead, then we will not need to transform the intercept coefficient (this is similar to how we had to carry out WLS for the heteroskedastic error case in the previous section).

The final value $\widehat{\rho}^{(K)}$ is then used to get the **FGLS** (**CORC**) estimates of $\widehat{\boldsymbol{\beta}}$. Again, depending on your specification, you may need to transform the intercept coefficient: $\widehat{\beta}_0 = \widehat{\beta}_0^*/(1-\widehat{\rho}^{(K)})$.
**These FGLS estimators are not unbiased but they are consistent and asymptotically efficient.**

The procedure can be carried out with the built-in functions as follows:

```
mdl_1_CORC <- orcutt::cochrane.orcutt(mdl_1)
#
print(coef(summary(mdl_1_CORC)))
```

```
##               Estimate Std. Error t value      Pr(>|t|)
## (Intercept) 10.402800 1.50824063   6.897  7.124332e-11
## x1           5.082365 0.48877934  10.398  1.899433e-20
## x2          -2.978490 0.04202203 -70.879 1.185668e-141
```

the estimated $\rho$ is:

```
print(mdl_1_CORC$rho)
```

```
## [1] 0.7088817
```

which is close to the true value of $\rho$ used in the data generation.

Alternative procedures to CORC are **Hildreth-Lu Procedure** and **Prais–Winsten Procedure**.

## Choosing Between HAC and FGLS

It has become more popular to estimate models by OLS and correct the standard errors for fairly arbitrary forms of serial correlation and heteroskedasticity.

Regarding **HAC**:

- Computation of robust standard errors works generally well in large datasets;
- With increase in computational power, not only has it become possible to (quickly) estimate models on large datasets, but also to calculate their robust covariance (HAC) estimators;
- While FGLS offers a theoretical efficiency, it involves making additional assumptions on the error covariance matrix, which may not be easy to test/verify, which may threaten the consistency of the estimator.

Regarding **FGLS**:

- if the explanatory variables are not strictly exogenous (i.e. if we include $Y_{i-1}$ on the right-hand-side of the equation) - the FGLS is not only inefficient, but it is also inconsistent;
- in most applications of FGLS, the errors are assumed to follow a first order autoregressive process. It may be better to evaluate OLS estimates and use a robust correction on their standard errors for more general forms of serial correlation;
- in addition to imposing an assumption of the residual covariance structure in regard to autocorrelation, GLS also requires an exogeneity assumption **(MR.3)** to hold, unlike HAC.

Generally, serial correlation is usually encountered in time-series data, which has its own set of models that specifically deal address serial correlation of either the residuals $\epsilon$, the endogenous variable $Y$, or the exogeneous variables $X_j$, or even all at once.
It is worth noting that autocorrelated residuals are more frequently the result of a misspecified regression equation, rather than a genuine autocorrelation.

A final thing to note:

In many cases, the presence of autocorrelation, especially in cross-sectional data, is not an indication that the model has autocorrelated errors, but rather that it:

- is misspecified;
- is suffering from omitted variable bias;
- has an incorrect functional form for either $Y$, or $X$.

## Monte Carlo Simulation: OLS vs FGLS

To illustrate the effects of heteroskedasticity on the standard errors of the estimates, and efficiency between OLS and FGLS, we will carry out a Monte Carlo simulation. We will simulate the following model:

$$\begin{cases} Y_i^{(m)} & = \beta_0 + \beta_1 X_{1,i}^{(m)} + \beta_2 X_{2,i}^{(m)} + \epsilon_i^{(m)} \\ \epsilon_i^{(m)} & = \rho \epsilon_{i-1}^{(m)} + u_i^{(m)}, \ |\rho| < 1, \ u_i^{(m)} \sim \mathcal{N}(0, \sigma^2), \quad \epsilon_0^{(m)} = 0, \quad i = 1, ..., N, \quad m = 1, ..., MC \end{cases}$$

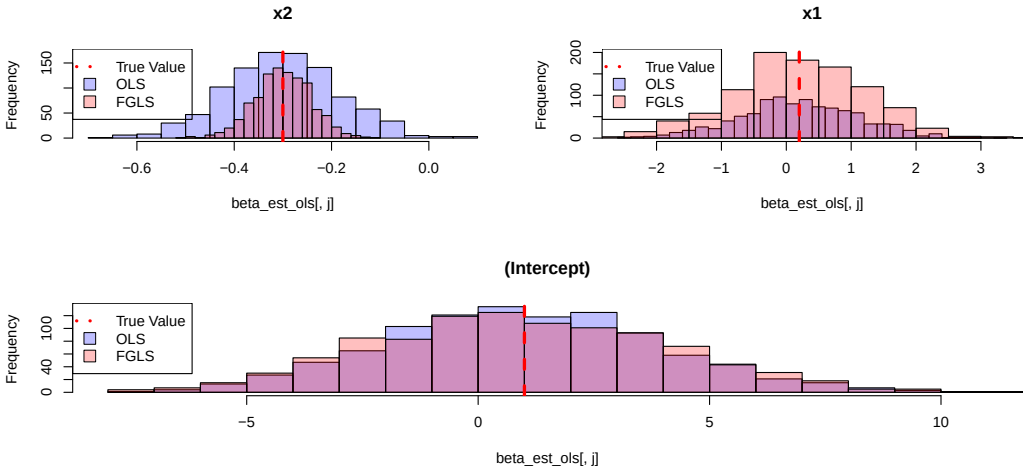
We will simulate a total of $MC = 1000$ samples from this model with specific coefficients and estimate the parameters via OLS, WLS, as well as correct the standard errors of OLS via HCE.
See the lecture notes for the code sample

Regarding the rejection rate of the null hypothesis that a parameter is insignificant, we have that:

```
##                          (Intercept)    x1    x2
## OLS: H0 rejection rate         0.373 0.505 0.757
## HAC: H0 rejection rate         0.457 0.365 0.919
## FGLS: H0 rejection rate        0.135 0.122 0.999
```

So, the FGLS seems to be better in some parameter cases, while HAC may be similar to OLS. All in all, if we only have an autocorrelation of order 1 problem - the corrections may not have a huge impact on our conclusions in **some** cases.

We can look at the coefficient estimate histograms:




For higher order and more complex correlation structure of the residuals, this may not always be the case, hence, if we detect serial correlation, we should account for it in some way.