Advanced Econometrics

Professor: Julie Shi

Timekeeper: Rui Zhou

Fall 2023

Contents

5 Generalized Regression Model and Heteroskedasticity

Generalized linear regression model applies when homoskedasticity is violated, i.e.,

$$E[\varepsilon\varepsilon'|X] = \sigma^2\Omega = \Sigma$$

Remarks: For mathematical convenience, we normalize the variance matrix so that $\operatorname{tr}(\Omega) = n$.

The generalized linear regression model is

$$y = X\beta + \varepsilon$$

with assumptions of

$$\begin{cases} E[\varepsilon|X] = 0 \\ E[\varepsilon\varepsilon'|X] = \sigma^2\Omega = \Sigma \end{cases}$$

The two leading cases we will consider in detail are heteroskedasticity and autocorrelation. Disturbances are heteroskedastic when they have different variances. Heteroskedasticity arises in volatile high-frequency time-series data and in cross-section data. Microeconomic data usually has heteroskedasticity. In heteroskedasticity, the disturbances are still assumed to be uncorrelated across observations, so $\sigma^2\Omega$ would be

$$\sigma^2\Omega = \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Autocorrelation is usually found in time-series data. Economic time series often display a "memory" in that variation around the regression function is not independent from one period to the next. Time-series data are usually homoskedastic, so $\sigma^2\Omega$ might be:

$$\sigma^2\Omega = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}$$

Panel data may exhibit both heteroskedasticity and autocorrelation.

$$\Gamma = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix}$$

$$\text{and } \sigma^2\Omega = \begin{bmatrix} \Gamma_1 & 0 & \cdots & 0 \\ 0 & \Gamma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Gamma_n \end{bmatrix}$$

In comparison with classical linear model, the quick conclusion is that, classical model with spherical disturbances, i.e., $\mathrm{E}\left[\varepsilon|X\right] = \mathbf{0}$ and $\mathrm{E}\left[\varepsilon\varepsilon'|X\right] = \sigma^2\mathbf{I}$, then the OLS estimator is best linear unbiased, consistent, and asymptotically normally distributed; **when homoskedasticity is violated, OLS estimator is still unbiased, consistent and asymptotically normally distributed, but no longer efficient, so the usual inference procedures are no longer appropriated.**

5.1 Properties of OLS Estimator

5.1.1 Unbiasedness

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

If $E[\varepsilon|X] = \mathbf{0}$, then

$$\mathrm{E}\left[b\right] = \mathrm{E}_X\left[\mathrm{E}\left[b|X\right]\right] = \beta$$

The covariance matrix of the disturbance vector has played no role here; unbiasedness is a property of the means.

5.1.2 (No-Longer) Efficiency

$$\begin{aligned} b &= \beta + (X'X)^{-1}X'\varepsilon \\ \Longrightarrow \mathrm{Var}\left[b|X\right] &= \mathrm{E}\left[(b - E[b])(b - E[b])'|X\right] \\ &= \mathrm{E}\left[(b-\beta)(b-\beta)'|X\right] \\ &= \mathrm{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X\right] \\ &= (X'X)^{-1}X'(\sigma^2\Omega)X(X'X)^{-1} \\ &= \frac{\sigma^2}{n}\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\Omega X\right)\left(\frac{1}{n}X'X\right)^{-1} \end{aligned}$$

As we know, under classical linear model's assumptions, especially the homoskedasticity assumption, the OLS estimator has variance matrix $\mathrm{Var}\left[b|X\right] = \sigma^2(X'X)^{-1}$ and is efficient. In the case of heteroskedasticity, the variance matrix is

$\sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}$, which is different from $\sigma^2(X'X)^{-1}(X'\mathbf{I}X)(X'X)^{-1}$, therefore, the estimator is no longer efficient.

5.1.3 Consistency

5.1.3.1 Consistency of b Since b is still unbiased, one natural idea is that, if $Var[b|X]$ converges to zero, then b is mean square consistent. With well-behaved regressors, $\left(\frac{X'X}{n}\right)^{-1}$ will converge to a constant matrix, so we consider

$$\frac{\sigma^2}{n}\left(\frac{X'\Omega X}{n}\right) = \left(\frac{\sigma^2}{n}\right)\left(\frac{\sum_{i=1}^n \sum_{j=1}^n \omega_{ij} x_i x_j'}{n}\right)$$

We see that though the leading constant will by itself converge to 0; the following matrix is a sum of n^2 terms, divided by n. Thus, the product is a scalar that if $O(\frac{1}{n})$ times a matrix that is at least at this juncture, $O(n)$, which results in $O(1)$. So, it does appear at first blush that if the product of $Var[b|X]$ does converge, it might converge to a matrix of nonzero constants. In this case, the covariance matrix of the least squares estimator would not converge to zero, and consistent would be hard to establish through mean square convergence. However, if only $\left(\frac{X'\Omega X}{n}\right)$ converges to a constant matrix, then $Var[b|X]$ would then converge to zero.

Consistency of OLS in the Generalized Regression Model If $Q = \text{plim}\left(\frac{X'X}{n}\right)$ and $\text{plim}\left(\frac{X'\Omega X}{n}\right)$ are both finite positive definite matrices, then b is consistent for β. Under the assumed conditions,

$$\text{plim } b = \beta$$

5.1.3.2 Consistency of s^2 Under assumption of plim $\frac{X'\Omega X}{n}$ being a finite positive matrix, we inspect the consistency of s^2:

$$s^2 = \frac{e'e}{n-K} = \frac{\varepsilon' M \varepsilon}{n-K}$$

As $M = \mathbf{I} - X(X'X)^{-1}X'$, we have

$$s^2 = \frac{\varepsilon'\varepsilon}{n-K} - \frac{\varepsilon X(X'X)^{-1}X'\varepsilon}{n-K}$$

For the first part, as $\text{tr}(\Omega) = n$ (which is a man-made normalization), we have

$$\text{E}\left[\frac{\varepsilon'\varepsilon}{n-K}\middle| X\right] = \frac{\text{tr}\left(\text{E}\left[\varepsilon\varepsilon'|X\right]\right)}{n-K} = \frac{n\sigma^2}{n-K}$$

For the second part, we have

$$\begin{aligned}
\mathrm{E}\left[\frac{\varepsilon' X(X'X)^{-1}X'\varepsilon}{n-K}\middle| X\right] &= \frac{\mathrm{tr}\left\{\mathrm{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon' X|X\right]\right\}}{n-K} \\
&= \frac{\mathrm{tr}\left\{\sigma^2\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\Omega X}{n}\right)\right\}}{n-K} \\
&= \frac{\sigma^2}{n-K}\mathrm{tr}\left\{\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\Omega X}{n}\right)\right\}
\end{aligned}$$

As $n \to \infty$, the first part will converge to σ^2; the second part will converge to zero because both matrices in the product are finite. Therefore, if b is consistent, then $\lim_{n\to\infty} s^2 = \sigma^2$.

In addition, if the fourth moment of every disturbance is finite and all other assumptions are met, then

$$\lim_{n\to\infty} \mathrm{Var}\left[\frac{e'e}{n-K}\right] = \lim_{n\to\infty} \mathrm{Var}\left[\frac{\varepsilon'\varepsilon}{n-K}\right] = 0$$

The result implies that

$$\mathrm{plim}\ b = \beta \Longrightarrow \mathrm{plim}\ s^2 = \sigma^2$$

5.1.3.3 Heteroskedasticity and Var $[b|X]$

$$\begin{aligned}
\mathrm{Var}\left[b|X\right] &= (X'X)^{-1}X'(\sigma^2\Omega)X(X'X)^{-1} \\
&= (X'X)^{-1}\left(\sigma^2\sum_{i=1}^{n}\omega_i x_i x_i'\right)(X'X)^{-1}
\end{aligned}$$

The difference between the conventional estimator and the appropriate (true) covariance matrix for b is

$$\mathrm{Est.Var}\left[b|X\right] - \mathrm{Var}\left[b|X\right] = s^2(X'X)^{-1} - \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}$$

In large sample cases, $s^2 \approx \sigma^2$, so the difference is approximately equal to

$$\begin{aligned}
\sigma^2\frac{X'X}{n} - \sigma^2\frac{X'X}{n}(X'\Omega X)\frac{X'X}{n} &= \sigma^2\frac{X'X}{n}(X'X)\frac{X'X}{n} - \sigma^2\frac{X'X}{n}(X'\Omega X)\frac{X'X}{n} \\
&= \sigma^2\frac{X'X}{n}\left\{(X'X) - (X'\Omega X)\right\}\frac{X'X}{n} \\
&= \frac{\sigma^2}{n}\left(\frac{X'X}{n}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(1-\omega_i)x_i x_i'\right)\left(\frac{X'X}{n}\right)^{-1}
\end{aligned}$$

If the heteroskedasticity is not correlated with the **variables** (instead of observations) in the model, then at least in large samples, the ordinary least squares computations, although not the optimal way to use the data, will not be misleading.

The preceding is a useful result, but one should not be overly optimistic. First, it remains true that ordinary least squares is demonstrably inefficient. Second, if the primary assumption of the analysis—that the heteroskedasticity is unrelated to the variables in the model—is incorrect, then the conventional standard errors may be quite far from the appropriate values.

5.1.4 Asymptotic Normal Distribution

Asymptotic Distribution of b in the Generalized Regression Model If the regressors are sufficiently well behaved and the off-diagonal terms in Ω diminish sufficiently, then the OLS estimator is asymptotically normally distributed with mean β and covariance matrix

$$\text{Asy.Var}\,[b] = \frac{\sigma^2}{n} Q^{-1} \text{plim}\, \left(\frac{X'\Omega X}{n} \right) Q^{-1}$$

Remarks: The condition "off-diagonal tears in ω diminish sufficiently" is required for applying the CLT, which requires a sequence of independent random variables.

5.1.5 IV Estimator

$$\begin{aligned} b_{IV} &= (\hat{X}'X)^{-1}\hat{X}'y \\ &= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'y \\ &= \beta + \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'\varepsilon \end{aligned}$$

Define $Q_{XX,Z}$ for convenience of matrix expression:

$$Q_{XX,Z} = \text{plim}\, \left\{ \left(\frac{X'Z}{n}\right) \left(\frac{Z'Z}{n}\right)^{-1} \left(\frac{Z'X}{n}\right) \right\}^{-1} \left(\frac{X'Z}{n}\right) \left(\frac{Z'Z}{n}\right)^{-1}$$

If the random term in b_{IV}, $\left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'\varepsilon$, vanishes asymptotically, then

$$\text{plim}\; b_{IV} = \beta + Q_{XX,Z}\text{plim}\, \left(\frac{Z'\varepsilon}{n}\right) = \beta$$

And

$$b_{IV} \overset{a}{\sim} N\left[\beta, \frac{\sigma^2}{n}(Q_{XX,Z})\text{plim}\, \left(\frac{Z'\Omega Z}{n}\right) (Q'_{XX,Z})\right]$$

5.2 Efficient Estimation by GLS

We start by considering the case in which Ω is a **known**, symmetric and positive definite matrix.

We have $\Omega = C\Lambda C'$. Let $T = C\Lambda^{1/2}$, and thus $\Omega = TT'$. Also, let $P = \Lambda^{-1/2}C'$, $P' = C\Lambda^{-1/2}$, we have

$$\begin{cases} \Omega^{-1} & = \left(C\Lambda C'\right)^{-1} = \left(C'\right)^{-1}\Lambda^{-1}C^{-1} = C\Lambda^{-1}C' \\ P'P & = C\Lambda^{-1}C' \end{cases} \Longrightarrow \Omega^{-1} = P'P$$

Consider the linear model of $y = X\beta + \varepsilon$, with $\mathrm{E}\left[\varepsilon\varepsilon'|X\right] = \sigma^2\Omega$. We do the following transformation:

$$\begin{aligned} y &= X\beta + \varepsilon \\ \Longrightarrow Py &= PX\beta + P\varepsilon \\ (y_*) &= (X_*\beta) + (\varepsilon_*) \end{aligned}$$

Luckily, we will show that **the transformed linear model is homoskedastic**:

$$\begin{aligned} \mathrm{E}\left[\varepsilon_*\varepsilon_*'|X\right] &= \mathrm{E}\left[P\varepsilon\varepsilon'P'|X_*\right] \\ &= \mathrm{E}\left[P\varepsilon\varepsilon'P'|X\right] \\ &= P\left(\sigma^2\Omega\right)P' \\ &= \sigma^2\left(\Lambda^{-1/2}C'\right)\left(C\Lambda C'\right)\left(C\Lambda^{-1/2}\right) \\ &= \sigma^2\mathbf{I} \end{aligned}$$

Remarks: From here we can understand the construction of P more intuitively. Our goal is to make the transformed model stick with the homoskedasticity assumption. To realize this, we shall "scale" the disturbances in a way that make the scaled version has an identity variance matrix. That is, we scale ε by P, and hope $\mathrm{Var}\left[P\varepsilon|X\right] = P\mathrm{Var}\left[\varepsilon|X\right]P' = \sigma^2 P\Omega P' = \sigma^2\mathbf{I}$. Hence, P works like the inverse of the square root of Ω, so $P = \Lambda^{-1/2}C'$.

The OLS estimator b is then

$$\begin{aligned} b &= (X_*'X_*)^{-1}X_*'y_* \\ &= \left[(PX)'(PX)\right]^{-1}(PX)'(Py) \\ &= \left(X'P'PX\right)^{-1}\left(X'P'Py\right) \\ &= \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}y \end{aligned}$$

The variance of this estimator is then

$$
\begin{aligned}
b &= \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}y \\
&= \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}\left(X\beta+\varepsilon\right) \\
&= \beta + \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}\varepsilon \\
\text{Or equivalently, } b &= (X_*'X_*)^{-1}X_*'y_* \\
&= (X_*'X_*)^{-1}X_*'\left(X\beta+\varepsilon\right) \\
&= \beta + (X_*'X_*)^{-1}X_*'\varepsilon \\
\Longrightarrow \operatorname{Var}\left[b|X\right] &= \sigma^2(X_*'X_*)^{-1} \\
&= \sigma^2\left(X'P'PX\right)^{-1} \\
&= \sigma^2\left(X'\Omega X\right)^{-1}
\end{aligned}
$$

Since $\mathrm{E}\left[\varepsilon_*\varepsilon_*'|X\right] = \sigma^2\mathbf{I}$, the classical regression model applies to this transformed model. Hence the $b = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega y$ is the efficient estimator of β. This estimator is the **generalized least squares (GLS)** or Aitken estimator of β. This estimator is in contrast to the ordinary least squares estimator, which uses a "weighting matrix", $\mathbf{I}$ instead of Ω^{-1}.

Properties of the GLS Estimator

1. If $\mathrm{E}\left[\varepsilon|X\right] = \mathbf{0}$, then the GLS estimator is unbiased.
2. The GLS estimator is consistent if $\operatorname{plim} \dfrac{X_*'X_*}{n} = Q_*$, where Q_* is a finite positive definite matrix. (That is, we require the transformed data $X_* = PX$, not the original data, to be well-behaved.)
3. The GLS estimator is asymptotically normally distributed, with mean β and sampling variance $\operatorname{Var}\left[b|X\right] = \sigma^2(X_*'X_*)^{-1} = \sigma^2\left(X'\Omega^{-1}X\right)^{-1}$.
4. The GLS estimator is the minimum variance linear unbiased estimator in the generalized regression model. (This statement follows by applying the Gauss-Markov theorem to the transformed model. In fact, the broad result includes the Gauss-Markov theorem as a special case when $\Omega = \mathbf{I}$.)

Moreover, all problems regarding this way of heteroskedasticity with known Ω can be addressed by the transformed linear model, because the transformed linear regression model again satisfies the classical linear model assumptions.

There is no precise counterpart to R^2 in the generalized regression model. Alternatives have been proposed, but care must be taken when using them. For example, one choice is the R^2 in the transformed regression. But this regression need not have a constant term, so the R^2 is not bounded by zero and one. Even if there is a constant term, the transformed regression is a computational device, not the model of interest. That a good (or bad) fit is obtained in the transformed model may be of no interest; the dependent variable in that model, y_*, is different from the one y in the model as originally specified. The usual R^2 often suggests that the fit of the model is improved by a correction for heteroskedasticity and degraded by a correction for autocorrelation, but both changes can often be attributed to the computation of y_*. A more appealing fit measure might be based on the residuals from the original model once the GLS estimator is in hand, such as

$$R_G^2 = 1 - \frac{(y - Xb_{GLS})'(y - Xb_{GLS})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Like the earlier contender, however, this measure is not bounded in the unit interval. In addition, this measure cannot be reliably used to compare models. The generalized least squares estimator minimizes the generalized sum of squares

$$\varepsilon_*'\varepsilon_* = (y - X\beta)\,\Omega^{-1}\,(y - X\beta)$$

not $\varepsilon'\varepsilon$. As such, there is no assurance, for example, that dropping a variable from the model will result in a decrease in R_G^2 , as it will in R^2. The R^2-like measures in this setting are purely descriptive. That being the case, the squared sample correlation between the actual and predicted values, $r_{y,\hat{y}}^2 = corr^2(y, \hat{y}) = corr^2\left(y, x'\hat{\beta}\right)$, would likely be a useful descriptor. Note, though, that this is not a proportion of variation explained, as is R^2; it is a measure of the agreement of the model predictions with the actual data.

5.3 Feasible GLS

When Ω is **unknown**, but has **specific structures,** we can estimate β using feasible generalized least squares (FGLS). A typical problem involves a set of parameters α such that $\Omega = \Omega(\alpha)$. A typical example is autocorrelation: $\Omega(\rho) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho^2 & \cdots & \rho^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$. Also, another usually-assumed heteroskedasticity is formed as: $\sigma_i^2 = \sigma^2 z_i^{\theta}$.

Procedure of Feasible GLS works as:

- Estimate $\hat{\alpha}$, a consistent estimator of α; (We will consider later how such an estimator might be obtained.)
- Obtain $\hat{\Omega} = \Omega(\hat{\alpha})$ and estimate β using $\hat{\beta} = \left(X'\hat{\Omega}^{-1}X\right)^{-1} X'\hat{\Omega}^{-1}y$, when satisfying

$$\begin{cases} \text{plim } \left[\frac{1}{n}X'\hat{\Omega}^{-1}X - \frac{1}{n}X'\Omega^{-1}X\right] = \mathbf{0} \\ \text{plim } \left[\frac{1}{\sqrt{n}}X'\hat{\Omega}^{-1}\varepsilon - \frac{1}{\sqrt{n}}X'\Omega^{-1}\varepsilon\right] = \mathbf{0} \end{cases}$$

The first of these equations states that if the weighted sum of squares matrix based on the true Ω converges to a positive definite matrix, then the one based on $\hat{\Omega}$ converges to the same matrix. We are assuming that this is true. In the second condition, if the transformed regressors are well behaved, then the right-hand-side sum will have a limiting normal distribution.

Efficiency of the FGLS Estimator An asymptotically efficient FGLS estimator does not require that we have an efficient estimator of α; only a consistent one is required to achieve full efficiency for the FGLS estimator.

5.3.1 Weighted Least Squares

In the most general case, given that $\sigma_i^2 = \sigma^2 w_i$, we have

$$\sigma^2\Omega = \sigma^2 \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix}$$

$$\Longrightarrow \Omega^{-1} = \begin{bmatrix} \frac{1}{\omega_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\omega_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\omega_n} \end{bmatrix}$$

$$P = P' = \begin{bmatrix} \frac{1}{\sqrt{\omega_1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\omega_2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\omega_n}} \end{bmatrix}$$

Therefore, the GLS estimator is obtained by regressing

$$Py = \begin{bmatrix} \frac{y_1}{\sqrt{\omega_1}} \\ \frac{y_2}{\sqrt{\omega_2}} \\ \vdots \\ \frac{y_n}{\sqrt{\omega_n}} \end{bmatrix} \text{ on } PX = \begin{bmatrix} \frac{x_1'}{\sqrt{\omega_1}} \\ \frac{x_2'}{\sqrt{\omega_2}} \\ \vdots \\ \frac{x_n'}{\sqrt{\omega_n}} \end{bmatrix}$$

Hence, the **weighted least square (WLS)** estimator is

$$\hat{\beta} = \left[\sum_{i=1}^{n} w_i x_i x_i'\right]^{-1} \left[\sum_{i=1}^{n} w_i x_i y_i\right], \text{ where } w_i = \frac{1}{\omega_i}$$

The logic of the computation is that observations with smaller variances receive a larger weight in the computations of the sums and therefore have greater influence in the estimates obtained.

5.3.1.1 WLS with Known Ω A common specification is that the variance is proportional to one of the regressors or its square. Say if $\sigma_i^2 = \sigma^2 x_{ik}^2$, then the transformed regression model for GLS is

$$\frac{y}{x_k} = \beta_k + \beta_1 \left(\frac{x_1}{x_k} \right) + \beta_2 \left(\frac{x_2}{x_k} \right) + \cdots + \frac{\varepsilon}{x_k}$$

However, it is rarely possible to be certain about the nature of the heteroskedasticity in a regression model. In one respect, this problem is minor, since the WLS estimator $\hat{\beta} = \left[\sum_{i=1}^n w_i x_i x_i'\right]^{-1} \left[\sum_{i=1}^n w_i x_i y_i\right]$ is consistent regardless of the weights used, as long as the weights are uncorrelated with the disturbances.

5.3.1.2 WLS with Unknown Ω Two-Step estimators:

1. Obtain estimates $\hat{\sigma}_i$ using $e_i = y_i - x_i' b$ based on OLS regression;

   The OLS estimator of β, although inefficient, is still consistent. As such, statistics computed using the OLS residuals, $e_i = y_i - x_i' b$, will have the same asymptotic properties as those computed using the true disturbances, $\varepsilon_i = y_i - x_i' \beta$.

2. Obtain $\hat{\hat{\beta}}$ using

$$\hat{\hat{\beta}} = \left[\sum_{i=1}^n \left(\frac{1}{\hat{\sigma}_i^2} \right) x_i x_i' \right] \left[\sum_{i=1}^n \left(\frac{1}{\sigma_i^2} x_i y_i \right) \right]$$

Remarks: The two-step estimator may be iterated by recomputing the residuals after computing the FGLS estimates and then reentering the computation.

5.4 White heteroskedasticity consistent estimator

We treat $\sigma^2 \Omega$ as a whole parameter, and assume $\mathrm{tr}\Omega = n$, which is just a normalization. $\sigma^2 \Omega$ contains $1 + \cdots + n = \dfrac{n(n+1)}{2}$ unknown parameters, which is almost impossible to estimate with a moderately large sample. We define

$$Q^* = \text{plim}\ \frac{X' \left(\sigma^2 \Omega\right) X}{n} = \text{plim}\ \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \sigma_{ij}^2 x_i x_j'$$

The least square estimator b is a consistent estimator of β, which implies that the least squares e_i are "pointwise" consistent estimators of their population counterparts ε_i. The general approach is to use X and e to devise an estimator of Q for the heteroskedasticity case, i.e., $\sigma_{ij} = 0$ when $i \neq j$.

White shows that, under very general conditions, the estimator $S_0 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i x_i'$ has plim $S_0 = Q^*$. Hence, White heteroskedasticity consistent estimator is

$$\text{Est.Asy.Var}\left[b\right] = \frac{1}{n}\left(\frac{X'X}{n}\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i x_i'\right]\left(\frac{X'X}{n}\right)^{-1}$$

which can be used to estimate the asymptotic covariance matrix of b.

This result implies that without actually specifying the type of heteroskedasticity, we can still make appropriate inferences based on the least squares estimator. This implication is especially useful if we are unsure of the precise nature of the heteroskedasticity (which is probably most of the time). However, the estimator is too optimistic when the sample is small.

5.5 White Test for Heteroskedasticity

Tests for heteroskedasticity are based on the following strategy. Ordinary least squares is a consistent estimator of β even in the presence of heteroskedasticity. As such, the ordinary least squares residuals will mimic, albeit imperfectly because of sampling variability, the heteroskedasticity of the true disturbances. Therefore, tests designed to detect heteroskedasticity will, in general, be applied to the ordinary least squares residuals.

$$\begin{aligned} &H_0 : \sigma_i^2 = \sigma^2, \forall i = 1, 2, \cdots, n \\ &H_1 : \text{Not } H_0 \end{aligned}$$

The intuition of White test for heteroskedasticity is that, if there is no heteroskedasticity, then $V = s^2(X'X)^{-1}$ will give a consistent estimator of $\text{Var}\left[b|X\right] = \sigma^2(X'X)^{-1}\left(X'\Omega X\right)\left(X'X\right)^{-1}$, which as we have seen can be estimated using White's heteroskedasticity consistent estimator.

Operationally speaking, White test is an F test to test the hypothesis that $\gamma_1 = \gamma_2 = 0$ in the following regression:

$$e_i^2 = \gamma_0 + x_i'\gamma_i + \left(x_i \otimes x_i\right)' \gamma_2 + v_i$$

Remarks: The White test is extremely general. To carry it out, we need not make any specific assumptions about the nature of the heteroscedasticity. Although this characteristic is a virtue, it is, at the same time, a potentially serious shortcoming. The test may reveal heteroskedasticity, but it may instead simply identify some other specification error (such as the omission of x^2 from a simple regression). Except in the context of a specific problem, little can be said about the power of White's test; it may be very low against some alternatives. In addition, unlike some of the other tests we shall discuss, the White test is **nonconstructive**. If we reject the null hypothesis, then the result of the test gives no indication of what to do next.

5.6 Clustering and the Moulton Factor

A bivariate model,

$$y_{ig} = \beta_0 + \beta_1 x_g + \varepsilon_{ig}$$

where $\varepsilon_{ig} = v_g + \eta_{ig}$. We assume that v and η are independent, $v_g \sim N\left(0, \sigma_v^2\right)$ and $\eta_{ig} \sim N\left(0, \sigma_\eta^2\right)$.

Therefore,

$$\text{Cov}\left[\varepsilon_{ig}, \varepsilon_{jg}\right] = \text{Cov}\left[v_g + \eta_{ig}, v_g + \eta_{jg}\right] = \sigma_v^2$$

As $\text{Var}\left[\varepsilon_{ig}\right] = \sigma_\varepsilon^2 = \sigma_v^2 + \sigma_\eta^2$, and denote $\rho = \dfrac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2} = \dfrac{\sigma_v^2}{\sigma_\varepsilon^2}$, so we have

$$\text{Cov}\left[\varepsilon_{ig}, \varepsilon_{jg}\right] = \sigma_v^2 = \rho\sigma_\varepsilon^2 > 0$$

The conclusion is that, if we do not cluster the standard errors to the right level, the variance will be mistakenly cut down by a Moulton factor:

$$\frac{V\left(\hat{\beta}_1\right)}{V_c\left(\hat{\beta}_1\right)} = 1 + (n-1)\rho$$

From the setting, we can first clarify the variables of interest:

$$y_g = \begin{bmatrix} y_{1g} \\ y_{2g} \\ \vdots \\ y_{n_g g} \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{bmatrix}$$

$$\varepsilon_g = \begin{bmatrix} \varepsilon_{1g} \\ \varepsilon_{2g} \\ \vdots \\ \varepsilon_{n_g g} \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_G \end{bmatrix}$$

$$X = \begin{bmatrix} l_1 x_1 \\ l_2 x_2 \\ \vdots \\ l_G x_G \end{bmatrix}$$

where l_g is a column vector of n_g-many 1s, and G is the number of groups.

$$\mathrm{E}\left[\varepsilon\varepsilon'|X\right] = \Omega = \begin{bmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_G \end{bmatrix}$$

$$\begin{aligned} \Omega_g &= \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \\ &= \sigma_\varepsilon^2 \left[(1-\rho)\,\mathbf{I} + \rho \cdot \left(l_g l_g'\right)\right] \end{aligned}$$

where $\left(l_g l_g'\right)$ is a $g \times g$ matrix full of 1s, and $\rho = \dfrac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2}$.

As we know, the heteroskedastic covariance matrix is $(X'X)^{-1}\,(X'\Omega X)\,(X'X)^{-1}$, and the homoskedastic covariance matrix is $\sigma^2(X'X)^{-1}$, so we compute the following two matrices $X'X$ and $X'\Omega X$:

$$\begin{aligned}
X'X &= \begin{bmatrix} x_1' l_1' & x_2' l_2' & \cdots & x_G' l_G' \end{bmatrix} \begin{bmatrix} l_1 x_1 \\ l_2 x_2 \\ \vdots \\ l_G x_G \end{bmatrix} \\
&= \sum_{g=1}^{G} n_g x_g' x_g \\
X'\Omega X &= \begin{bmatrix} x_1' l_1' & x_2' l_2' & \cdots & x_G' l_G' \end{bmatrix} \begin{bmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_G \end{bmatrix} \begin{bmatrix} l_1 x_1 \\ l_2 x_2 \\ \vdots \\ l_G x_G \end{bmatrix} \\
&= \sum_{g=1}^{G} x_g' l_g' \Omega_g l_g x_g \\
&= \sum_{g=1}^{G} x_g' l_g' \left\{ \sigma_\varepsilon^2 \left[(1-\rho) \mathbf{I} + \rho l_g l_g' \right] \right\} l_g x_g \\
&= \sigma_\varepsilon^2 \sum_{g=1}^{G} \left\{ x_g' \left[(1-\rho) l_g' l_g + \rho l_g' l_g l_g' l_g \right] \right\} x_g \\
&= \sigma_\varepsilon^2 \sum_{g=1}^{G} \left\{ x_g' \left[(1-\rho) n_g + \rho n_g^2 \right] \right\} x_g \\
&= \sigma_\varepsilon^2 \sum_{g=1}^{G} n_g \left[1 + (n_g - 1) \rho \right] x_g' x_g \\
&:= \sigma_\varepsilon^2 \sum_{g=1}^{G} n_g \tau_g x_g' x_g
\end{aligned}$$

If the group sizes are equal, that is, $n_g = n$ for all $g = 1, \cdots, G$; then $\tau := \tau_g = 1 + (n-1)\rho$. In this case,

$$\begin{aligned}
V\left(\hat{\beta}\right) &= \sigma_\varepsilon^2 \tau \left(\sum_g n x_g' x_g \right)^{-1} \left(\sum_g n x_g' x_g \right) \left(\sum_g n x_g' x_g \right)^{-1} \\
&= \sigma_\varepsilon^2 \tau \left(\sum_g n x_g' x_g \right)^{-1} \\
&= \tau V_c \left(\hat{\beta}\right)
\end{aligned}$$

So we jointly have

$$\frac{V\left(\hat{\beta}\right)}{V_c\left(\hat{\beta}\right)} = \tau = 1 + (n-1)\,\rho$$

which we called the Moulton factor.

The basic framework of a regression model for panel data is

$$\begin{aligned} y_{it} &= x_{it}'\beta + z_i'\alpha + \varepsilon_{it} \\ &= x_{it}'\beta + c_i + \varepsilon_{it} \end{aligned}$$

There are K regressors in x_{it}, not including a constant term. The heterogeneity, or individual effect is $z_i'\alpha$ where z_i contains a constant term and a set of individual or group-specific variables which are constant over time t. The heterogeneity can be either observed (e.g., sex and education) or unobserved (e.g., family specific characteristics, individual heterogeneity in skill or preference). If z_i is observed for all individuals, then the entire model can be treated as an ordinary linear model and fit by least squares. The complications arise when c_i is unobserved, which will be the case in most applications.

The main objective of the analysis will be consistent and efficient estimation of the partial effects,

$$\beta = \frac{\partial \mathrm{E}\left[y_{it}|x_{it}\right]}{\partial x_{it}}$$

Whether this is possible depends on the assumptions about the unobserved effects. We begin with a strict exogeneity assumption for the independent variables. Strict exogeneity requires that

$$\mathrm{E}\left[\varepsilon_{it}|x_{i1}, x_{i2}, \cdots, c_i\right] = \mathrm{E}\left[\varepsilon_{it}|X_i, c_i\right] = 0$$

That is, the current disturbance is uncorrelated with the independent variables in every period, past, present, and future. The crucial aspect of the model concerns the heterogeneity. A particularly convenient assumption would be mean independence, that is, the unobserved variables are uncorrelated with the included variables:

$$\mathrm{E}\left[c_i|x_{i1}, x_{i2}, \cdots\right] = \alpha$$

If the missing variable(s) are uncorrelated with the included variables, then, as we shall see, they may be included in the disturbance of the model. This is the assumption that underlies the random effects model, as we will explore later. It is, however, a particularly strong assumption—it would be unlikely in the labor market and health care examples mentioned previously. The alternative would be

The unobserved variables are correlated with the included variables if

$$\mathrm{E}\left[c_i | x_{i1}, x_{i2}, \cdots\right] = h\left(x_{i1}, x_{i2}, \cdots\right) = h\left(X_i\right)$$

This formulation is more general, but at the same time, considerably more complicated, the more so since it may require yet further assumptions about the nature of the function.