Heteroskedasticity, or non-spherical errors, means that the variance of errors is not constant and that two errors might be correlated with each other. That is,

$$\text{Var}(\varepsilon_i) \neq \sigma^2 \quad \text{and} \quad \text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$$

or

$$E[\varepsilon\varepsilon^T] \neq \sigma^2 I$$

In matrix form, we no longer have a constant diagonal matrix. We can denote this by $\text{Var}(\varepsilon | X) = \sigma^2 \Psi$, where σ^2 is still the true variance and Ψ is some positive-definite matrix ($v^T \Psi v > 0$)

Is the OLS estimator still unbiased? Prove it.

What is affected then? The variance and hence the SE. Compute OLS variance

$$\text{Var}(\hat{\beta}) = \text{Var}\left((X^TX)^{-1}X^Ty\right)$$

$$\vdots$$

$$= (X^TX)^{-1}X^T E[\varepsilon\varepsilon^T] X (X^TX)^{-1}$$

$$= (X^TX)^{-1}X^T \sigma^2\Psi X (X^TX)^{-1}$$

which is clearly different from $\text{Var}(\hat{\beta})$ under homoskedasticity

As a consequence, the t and F - tests will be biased and inference will be wrong. Unfortunately, software packages do not account for this

automatically. Often these assume the GMAs hold.

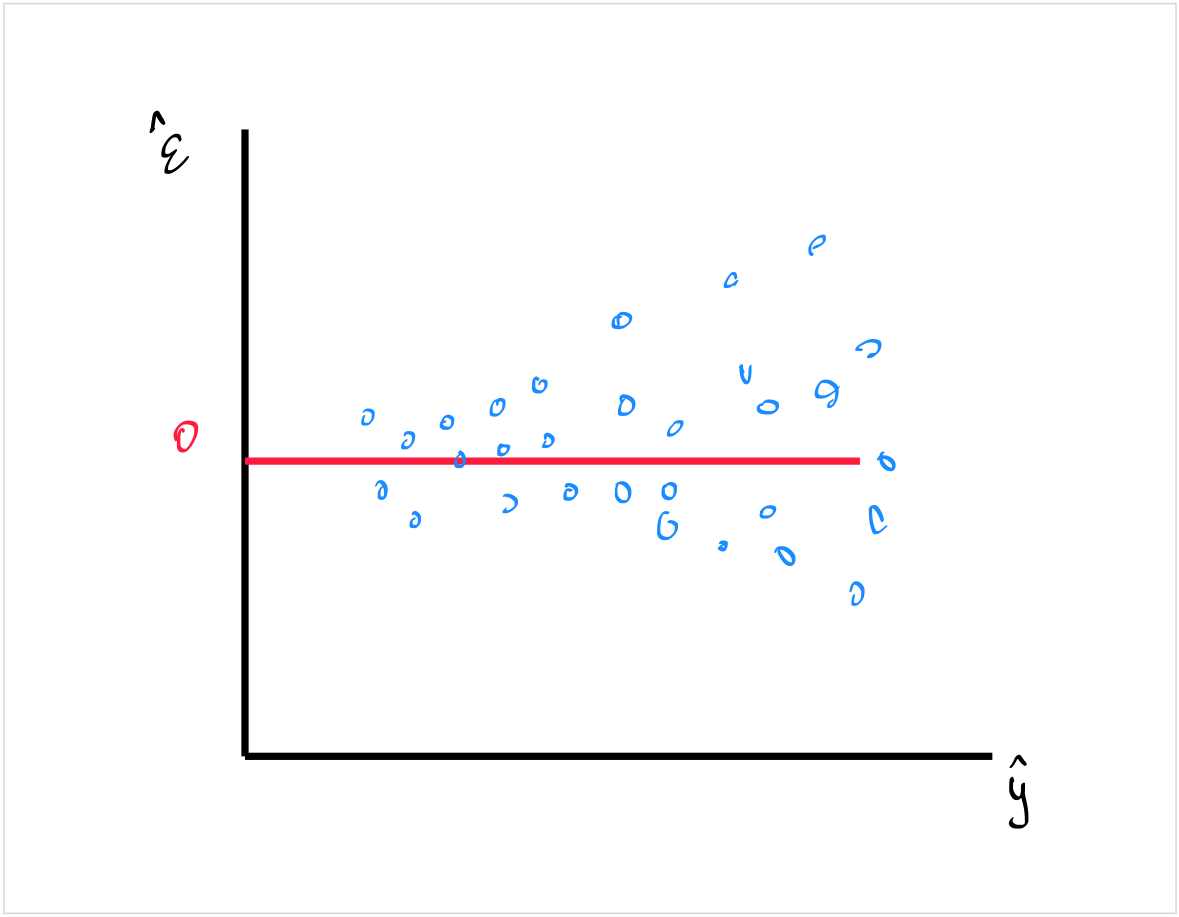
While the potential sources of multicollinearity and mis specification are fairly obvious, the ways in which non-spherical errors may creep into your cross-sectional dataset are less intuitive.

Suppose you are modeling income based on education. We can intuitively agree that the effect will depend on the quantiles of education (say high school, undergraduate, and graduate). But, will the range of errors be equally distributed within each group? Debate it for a bit.

My intuition suggests that the more education you have, the more opportunities you will have. Hence, someone with only HS education will only have a limited number of employment options and hence the range of incomes they could earn is narrower (i.e., more constraint). Now, if you have a PhD your options are more diverse. Perhaps you land a passion research role that doesn't pay super well, a gov. job with decent salary, or a quant job that pays millions. Although education will not be the only relevant factor, we expect that on average the fit of a regression that doesn't account for these natural clusters will perform poorly. This is an

example of heteroskedasticity. The best way to diagnose this is a fitted vs Residuals plot. It may look something

Like a "cone".



To claim evidence of homoskedasticity we want this plot to look like a random cloud of points.

If my way of thinking about this makes sense to you, then you are basically looking for the range of your regressor. If the range is big (large difference between min and max value) then it is more likely that you'll encounter hetero skedasticity.

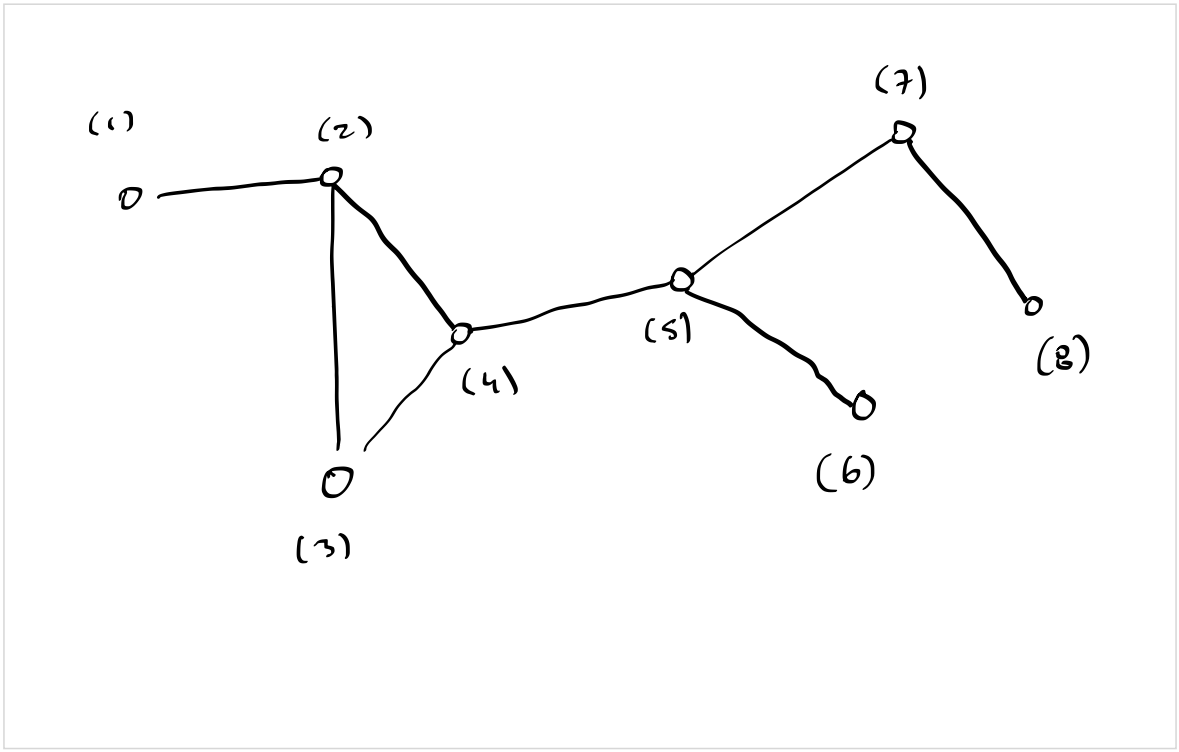
Autocorrelation is super common in time-series models with lagged vars. Intuitively, it makes that yesterday's errors will be correlated with today's errors.

But, how does this arise in cross-sectional scenarios where there is no time component?

Think of spatial relationships: states, neighborhoods, or trade networks. For example, if a house gets sold for a very high price then the sorrounding houses will also appreciate. Not necessarily because it makes them better, but simply because of this proximity effect. Consequently, the residuals of houses' value in that neighborhood will be correlated because they share this common factor (i.e., some house selling for a high price).

Bank networks or supply chains have

similar properties. The spatial factor here is proximity in the network.



If Banks 4 or 5 fail, then the effects will diffuse throughout the entire bank network because they are highly connected. The residuals will capture these effects as they

diffuse and hence they would be correlated with each other.

Generally, there are four common cases:

1) Proportional het: $Var(\varepsilon_i) = \sigma^2 \cdot x_i$
2) Grouped het: $Var(\varepsilon_i) = \begin{cases} \sigma^2 \text{ if } i \text{ in group 1} \\ s^2 \text{ if } i \text{ in group 2} \end{cases}$
3) Multiplicative het: $Var(\varepsilon_i) = \sigma^2 \cdot e^{z_i' \alpha}$
4) Unknown

o oo

Strategies

So, what can we do to fix non-spherical

errors? We have a few strategies. In general, we either transform the variables to "flatten" the errors or we fix the OLS standard errors after estimation. Of course, there is always the option of re-specifying the model entirely.

Active strategy (transformations)

Suppose $y = X\beta + \varepsilon$ does not have spherical errors. So $\mathbb{E}[\varepsilon\varepsilon^T] = \sigma^2 \Psi$, where Ψ is a positive-definite matrix.

The idea is to find a transformation that turns Ψ into an identity matrix. That is, we want to find a matrix that when multiplied by Ψ gives us

I. Logically, this matrix will be the inverse Ψ^{-1}. Recall the definition of an inverse

$$\Psi\Psi^{-1} = I$$

(Theorem) If Ψ is positive definite then there exists a matrix $P \in \mathbb{R}^{n \times n}$, which is square and non-singular, such that $\Psi^{-1} = PP^T$

$$\Psi^{-1} = P^TP \Rightarrow \Psi = (P^TP)^{-1} = P^{-1}(P^T)^{-1}$$

It follows that $P\Psi P^T = PP^{-1}(P^T)^{-1}P^T = I_n$ Putting Ψ in the middle is a convenient choice to ensure we get the identity. Hence, $P\Psi P^T$ is precisely the transform-ation we want to apply.

Exogeneity is assumed to hold and, because the transformation is deterministic, we can assume P is not random. That is,

$$E[\varepsilon | X] = 0 \quad \text{and} \quad E[P\varepsilon | X] = P\,E[\varepsilon | X] = 0$$

Therefore, the variance of the transformed version is

$$\begin{aligned} Var(P\varepsilon | X) &= P\,Var(\varepsilon | X)\,P^T \\ &= P\,\sigma^2 \Psi\,P^T = \sigma^2\,P\,\Psi P^T \\ &= \sigma^2 I_n \end{aligned}$$

Recall that a constant comes out squared from the variance operator. Pre multiplying by P and post

multiplying by P^T is the same concept in matrix form.

So, while ε is heteroskedastic $P\varepsilon$ is not. We can use P to transform our regression model.

$$Py = PX\beta + P\varepsilon \Rightarrow y^* = X^*\beta + \varepsilon^*$$

Under this model, the estimator $\hat{\beta}^*$ is BLUE and has the same interpretation as an untransformed OLS model.

$$\begin{aligned}\hat{\beta}^* &= (X^* X^{*T})^{-1} X^{*T} y^* \\ &= (X^T P^T P X)^{-1} X^T P^T P y \\ &= (X^T \Psi^{-1} X)^{-1} X^T \Psi^{-1} y\end{aligned}$$

This is called the Generalized Least Squares (GLS) estimator. There is one problem though, we don't know the variance-covariance matrix a priori. So we need to estimate it using the sample data. Suppose $\hat{\Psi}$ is this estimator and so $\hat{\Psi} \xrightarrow{p} \Psi$.

Then, we have $\hat{\beta}^* = (X^T \hat{\Psi}^{-1} X)^{-1} X^T \Psi^{-1} y$

which is known as the Feasible GLS.

$$\text{Var}(\hat{\beta}^*) = \sigma^2 (X^{*T} X^*)^{-1}$$

$$= \sigma^2 (X^T P^T P X)^{-1}$$

$$= \sigma^2 (X^T \Psi^{-1} X)^{-1}$$

$$\Rightarrow \hat{\text{Var}}(\hat{\beta}^*) = \hat{\sigma}^2 (X^T \hat{\Psi}^{-1} X)^{-1}$$

↳ Sample variance

$$\sigma^2 = \frac{1}{n-k} \sum \hat{\varepsilon}_i^{*2}$$

Note that $\sum \hat{\varepsilon}_i^{*2} = \varepsilon^{*T}\varepsilon^{*}$

$= (y^* - X^*\hat{\beta}^*)^T (y^* - X^*\hat{\beta}^*)$

$$Var(\hat{\beta}^*) = \frac{(y^* - X^*\hat{\beta}^*)^T (y^* - X^*\hat{\beta}^*)(X^T\hat{\psi}^{-1}X)^{-1}}{n-k}$$

$$= \frac{1}{n-k} (y^{*T} - \hat{\beta}^{*T}X^{*T})(y^* - X^*\hat{\beta}^*)(X^T\psi^{-1}X)^{-1}$$

Let's take this step by step...

$$\varepsilon^* = (y^* - X^*\hat{\beta}^*) = (Py - PX(X^T\psi^{-1}X)^{-1}X^T\psi^{-1}y)$$

$$= P[y - X(X^T\psi^{-1}X)^{-1}X^T\psi^{-1}y] = P[y - X\hat{\beta}^*]$$

$$\varepsilon^{*T} = (P[y - X(X^T\psi^{-1}X)^{-1}X^T\psi^{-1}y])^T = [y - X\hat{\beta}^*]^T P^T$$

$$\varepsilon^{*T}\varepsilon^{*} = (y - X\hat{\beta}^{*})^{T} P^{T} P (y - X\hat{\beta}^{*})$$

$$= (y - X\hat{\beta}^{*})^{T} \hat{\psi}^{-1} (y - X\hat{\beta}^{*})$$

Putting everything together...

$$Var(\hat{\beta}^{*}) = \frac{(y - X\hat{\beta}^{*})^{T} \hat{\psi}^{-1} (y - X\hat{\beta}^{*}) (X^{T}\hat{\psi}^{-1}X)^{-1}}{n - k}$$

Let's compare it to the heteroskedastic model's variance.

OLS_{H} FGLS

$$\sigma^{2}(X^{T}X)^{-1}X^{T}\psi X(X^{T}X)^{-1} \geq \sigma^{2}(X^{T}\psi X)^{-1}$$

$$Var(\hat{\beta}_{H}) \geq Var(\hat{\beta}^{*})$$

Think of FGLS as a weighting protocol. We assign less weight to observations

with high variance and more weight to those with low variance. In contrast, OLS weights all variables equally. So, intuitively, the FGLS optimizes for lower variance by selecting optimal weights.

Geometrically, OLS minimizes the SSR (euclidean distance) and FGLS minimizes the weighted SSR (mahalanobis distance). Thus, under heteroskedastic errors, the FGLS objective function is much better aligned with the true error structure.

Active Strategy

Suppose we have a model $y = X\beta + \varepsilon$ with heteroskedastic but no auto

correlated errors. That is, $Var(\varepsilon|X) = \sigma^2\Psi$

where $\Psi = \begin{bmatrix} h_1^2 & & 0 \\ & \ddots & \\ 0 & & h_n^2 \end{bmatrix} \Rightarrow \Psi^{-1} = \begin{bmatrix} 1/h_1^2 & & 0 \\ & \ddots & \\ 0 & & 1/h_n^2 \end{bmatrix}$

Here, h_i is the heteroskedasticity function that gives the variance for error ε_i. Each error may have a different value, hence a function. Common specifications for h_i are: $z_i \propto \frac{1}{z}$

$$h_i = x_i \;,\; h_i^2 = \sqrt{x_i} \;,\; h_i = e$$

Let's start by transforming the model

$$Py = PX\beta + P\varepsilon$$

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

Recall $\Psi^{-1} = P^TP$, so $Py = \begin{bmatrix} \frac{y_1}{h_1} \\ \vdots \\ \frac{y_n}{h_n} \end{bmatrix}$

So we have the following model

$$\frac{y_i}{h_i} = \left(\frac{x_i}{h_i}\right)^T \beta + \frac{\varepsilon_i}{h_i}$$

diagonal

$$Var\left(\varepsilon_i^*\right) = Var\left(\frac{\varepsilon_i}{h_i}\right) = Var\left(P\varepsilon\right) = P\ Var(\varepsilon)\ P^T$$

$$= P\ \sigma^2 \Psi\ P^T = \sigma^2 I_n$$

homoskedastic

$$\hat{\beta}^* = \left(X^T \Psi^{-1} X\right)^{-1} X^T \Psi^{-1} y \quad , \quad \hat{\Psi}^{-1} = P^T P = \begin{bmatrix} \frac{1}{h_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{h_n^2} \end{bmatrix}$$

Because Ψ has a specific form now, the name for this GLS is Weighted Least Squares (WLS). Where h_i are the weights. In our current example, the largest weights are assigned to the observations with the most precise information (i.e., lowest variance).

We know that $Var\left(\hat{\beta}^*\right) = \sigma^2 \left(X^T \Psi^{-1} X\right)^{-1}$

At this point, we either assume a functional form for h_i that depends on the data or we try to estimate it. That is, we would try to find $\hat{h}_i$ st $\hat{h}_i \xrightarrow{p} h_i$.
In matrix form, we want to find $\hat{\psi} \xrightarrow{p} \psi$ so that we can compute the FGLS $\hat{\beta}^*$ with

$$\widehat{Var}(\hat{\beta}^*) = \hat{\sigma}^2 (X^T \hat{\psi}^{-1} X)^{-1}$$

where

$$\hat{\sigma}^2 = \frac{\varepsilon^{*T} \varepsilon^*}{n-k}$$

Once we have this*, we can run t or F tests for inference

* Variance-Covariance matrix

If the errors are normal, then the FGLS has approximates a normal distribution

In small samples $\hat{\beta}^* \sim \mathcal{N}(\beta, \text{Var}(\hat{\beta}^*))$

In large samples (CLT) $(\hat{\beta}^* - \beta) \xrightarrow{d} \mathcal{N}(0, \text{Var}(\hat{\beta}^*))$

Passive Strategy

Sometimes we can't find the heteroskedasticity function or we prefer not to re specify our model. In these situations we can employ a "passive" solution by correcting the biased SE of the original OLS estimation. But it won't

fix efficiency problems completely.

In 1980, econometrician Halbert White proved that the squared residuals of a heteroskedastic OLS estimate can be used to estimate the variance-covariance matrix that is, if we have heteroskedastic but no auto correlated errors

$$Var(\varepsilon_i) = \sigma_i^2 I_n$$

We could use $\hat{\varepsilon}_i^2 I_n$ to estimate $Var(\varepsilon_i)$ and therefore $Var(\hat{\beta})$. Even when the form of heteroskedasticity (i.e., h_i) is unknown.

Basically, under het. we have

$$Var(\hat{\beta}) = (X^T X)^{-1} X^T \sigma_i^2 I_n X (X^T X)^{-1}$$

Since we don't know σ_i^2, replace with $\hat{\varepsilon}_i^2$

$$Var(\hat{\beta}_{HC}) = (X^TX)^{-1} X^T \hat{\varepsilon}_i^2 I_n X (X^TX)^{-1}$$

which yield the so-called Heteroskedasticity-Consistent (HC) variance estimator.

$$Var(\hat{\beta}_{HC}) \xrightarrow{P} Var(\hat{\beta})$$

The procedure can be broken down into 6 steps for robust inference

1) Run OLS to get $\hat{\beta}$ and $\hat{\varepsilon}$
2) Compute $\hat{\varepsilon}_i^2$ or $\hat{\varepsilon}^T\varepsilon$
3) Create matrix $\hat{\Omega} = diag(\hat{\varepsilon}_1^2, \hat{\varepsilon}_2^2, \ldots, \hat{\varepsilon}_n^2)$

4) Compute $Var(\hat{\beta}_{HC}) = (X^TX)^{-1}X^T\hat{\Omega}X(X^TX)^{-1}$

5) Get standard errors $\hat{\varepsilon}_i$ (i.e., $\sqrt{\hat{\Omega}}$)

6) Compute t-stats with these SE.

There is a caveat tho. This procedure requires a large sample ($n > 200$) to get good estimates from the residuals. If the sample small we will still overestimate the true SE (although it's still better than doing nothing...)

A few improvements have been published since 1980.

HC1: Deegrees of freedom adjustment

$$Var(\hat{\beta}_{HC1}) = \frac{n}{n-k}(X^TX)^{-1}X^T\hat{\Omega}X(X^TX)^{-1}$$

better for moderate sample sizes ($n > 100$) if there are no unusual leverage points (outliers wrt to the independent variable)

Other estimators of the HC family (HC2, ...) introduce different ways of dealing with high leverage points in smaller samples. I encourage you look into these if you encounter heteroskedasticity in your final project.