ECON2228.05 Econometric Methods

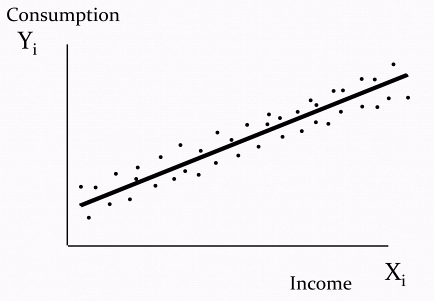
Module 1
Heteroskedasticity

Alberto Cappello

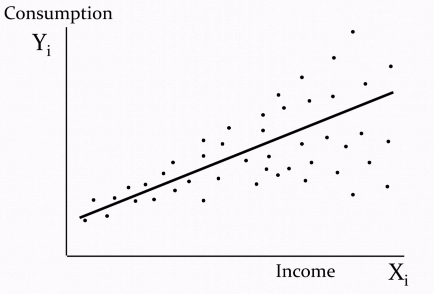
Department of Economics, Boston College

Spring 2025

Motivating Example: Income and Consumption Variability



(a) Homoskedasticity



(b) Heteroskedasticity

Motivating Example: Income and Consumption Variability

Scenario: You want to analyze the relationship between:

- Y: Consumption (\$)
- X: Income (\$)

Observation:

- For low-income households, consumption tend to be small and less variable.
- For high-income households, consumption vary widely.

Model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Heteroskedasticity: the variance of the error term depends on X_i, violating the assumption of constant variance.

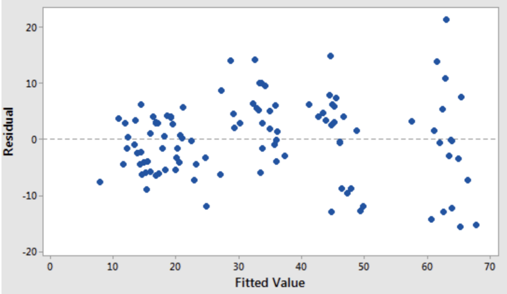
$$\text{Var}(\epsilon_i | X_i) = \sigma_i^2$$

Consequences

- This will mean that your SEs are larger (not smaller) than they should be. The result is that you have less statistical power than you should.
- The reason is that the residuals scatter more widely about the regression line in one proportion of the range than another. As a result, those data contain less information about the location of the conditional mean (the vertical position of the regression line at that point) than the model assumes. They should be down-weighted (and the region with lower residual variance should be up-weighted), but if each point gets full weight anyway, then the model thinks 'these data vary a lot, we just aren't sure where the line is supposed to be.
- In the presence of heteroscedasticity, the standard errors of OLS estimators are *not* provided by the usual OLS formulas. If we persist in using the usual OLS formulas, the t- and F-tests based on them can be highly misleading, resulting in erroneous conclusions.
- Our estimates are no longer efficient and thus not BLUE.

Testing for Heteroskedasticity

- Since the coefficients are unbiased, the residual e_i is an unbiased estimate of ε_i.
- We have
$$\sigma_i^2 = \mathrm{Var}(\varepsilon_i \mid X) = E\left[\varepsilon_i^2 \mid X\right].$$
- We use e_i^2 as a proxy for ε_i^2.
- Visual Test: Check if e_i^2 is related to predictors that vary across observations.



Breusch-Pagan Test

- Run the auxiliary regression:
$$e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \eta_i.$$
- Obtain R_a^2 from this regression.
- The test statistic is
$$N \times R_a^2,$$
which is distributed as χ_K^2, where K is the number of predictors.
- Alternatively, an F-test for $\alpha_1 = \alpha_2 = 0$ gives similar results (assuming normal errors).

White Test

- White's auxiliary regression is:

$$e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{1i}^2 + \alpha_4 x_{2i}^2 + \alpha_5 x_{1i} x_{2i} + \eta_i.$$

- The test statistic is still $N \times R_a^2$, but the degrees-of-freedom increase (to 5 in this example).
- A simplified version uses the regression:

$$e_i^2 = \alpha_0 + \alpha_1 \hat{y}_i + \alpha_2 \hat{y}_i^2 + \eta_i.$$

- This version always has 2 degrees-of-freedom, regardless of the number of predictors in $\hat{y}_i$.

Resolving Heteroskedasticity

(i) Generalized Least Squared (GLS)
(ii) Feasible Generalized Least Squared (FGLS)

Notice:

- The aim of GLS and FGLS is to restore a homoskedastic model and achieve BLUE estimates.
- However, these methods change the estimation procedure and yield different coefficient estimates compared to OLS.
- If OLS is used with incorrect (typically too small) standard errors, inference is compromised.
- An alternative is to keep OLS estimates but adjust their standard errors to account for heteroskedasticity.

Generalized Least Squares (GLS)

Multiply each term by

$$w_i = \frac{1}{\sigma_i} \quad \text{where} \quad \sigma_i \quad \text{are known}$$

Rewriting the original model using w_i:

$$Y_i = \beta_0 w_i + \beta_1 (X_{2i} w_i) + \beta_2 (X_{2i} w_i) + (u_i \, w_i).$$

We can define new variables:

$$Y_i^* = w_i \, Y_i, \quad X_{ji}^* = w_i \, X_{ji}, \quad u_i^* = w_i \, u_i.$$

Hence, the transformed equation becomes

$$Y_i^* = \beta_0^* + \beta_1^* X_{1i}^* + \beta_2^* X_{2i}^* + u_i^*,$$

which satisfies the homoskedasticity assumption for the error term u_i^*.

Feasible Generalized Least Squared (FGLS)

Notice that σ_i is unknown and needs to be estimated. Therefore WLS is not feasible in practice. To implement Feasible Generalized Least Squared (FGLS):

(i) Run OLS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,$$

(ii) Since $\sigma_i^2 = \mathrm{Var}(\varepsilon_i \mid X) = E\left[\varepsilon_i^2 \mid X\right]$, σ_i^2 can be estimated with the squared residuals ${\hat{u}_i}^2$ where $\hat{u}_i = Y_i - \hat{Y}_i$. Hence, run OLS on

$$\log(\hat{u}_i^2) = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + \eta_i,$$

(iii) Let $\tilde{u}_i$ denote the fitted values of this regression. Apply GLS using $w_i = \sqrt{exp(\tilde{u}_i)}$ to estimate the parameters of interest $(\beta_0, \beta_1, \beta_2)$. Denote the estimated parameters as

$$\hat{\beta}_{FGLS}$$

Property: FGLS is asymptotically efficient and consistent.