Heterocedasticidad

Gabriel V. Montes-Rojas

Supuestos de Gauss-Markov

- Supuesto 5: Homocedasticidad $Var(u|\mathbf{x}) = \sigma^2$. (homo: igual, cedasticidad: varianza)

Teorema Gauss-Markov: Bajo los Supuestos 1-5, los estimadores MCO $(\hat{\beta}_1, ..., \hat{\beta}_K)$ *son los mejores estimadores lineales de* $\beta_1, ..., \beta_K$. *Note: MEJOR significa mínima varianza.*

La **heterocedasticidad** *(hetero: distinta, cedasticidad: varianza) se define como* $Var(u_i|\mathbf{x}_i) = \sigma_i^2$ *for* $i = 1, 2, ..., N$. *Significa que la varianza condicional del error no es constante.*

Heteroscedasticidad

- El principal problema con la heterocedasticidad es que $\widehat{Var(\hat{\beta}|x)}$ no es más válido cuando se estima bajo los supuestos de Gauss-Markov. ¿Por qué?
- Supongamos un modelo de regresión simple $y_i = \beta_1 + \beta_2 x_i + u_i, i = 1, 2, ..., N$:

$$Var(\hat{\beta}_2|x) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- Ahora consideremos una expresión más general:

$$Var(\hat{\beta}_2|x) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2\sigma_i^2}{(\sum_{i=1}^{N}(x_i - \bar{x})^2)^2}$$

  donde $\sigma_i^2 = Var(u_i|x_i) = \sigma^2(x_i)$
- Estas dos expresiones no tienen por qué coincidir. Entonces, **la inferencia puede ser inválida porque se usan los errores estándar incorrectos.** (Estudiar en qué condiciones son iguales. Solución: σ_i^2 no correlacionado con x_i.)
- Sin embargo, notemos que los estimadores MCO siguen siendo **insesgados** y **consistentes**. Hacer la prueba de insesgadez, y fijarse que en ningún momento se usa el supuesto de heterocedasticidad.

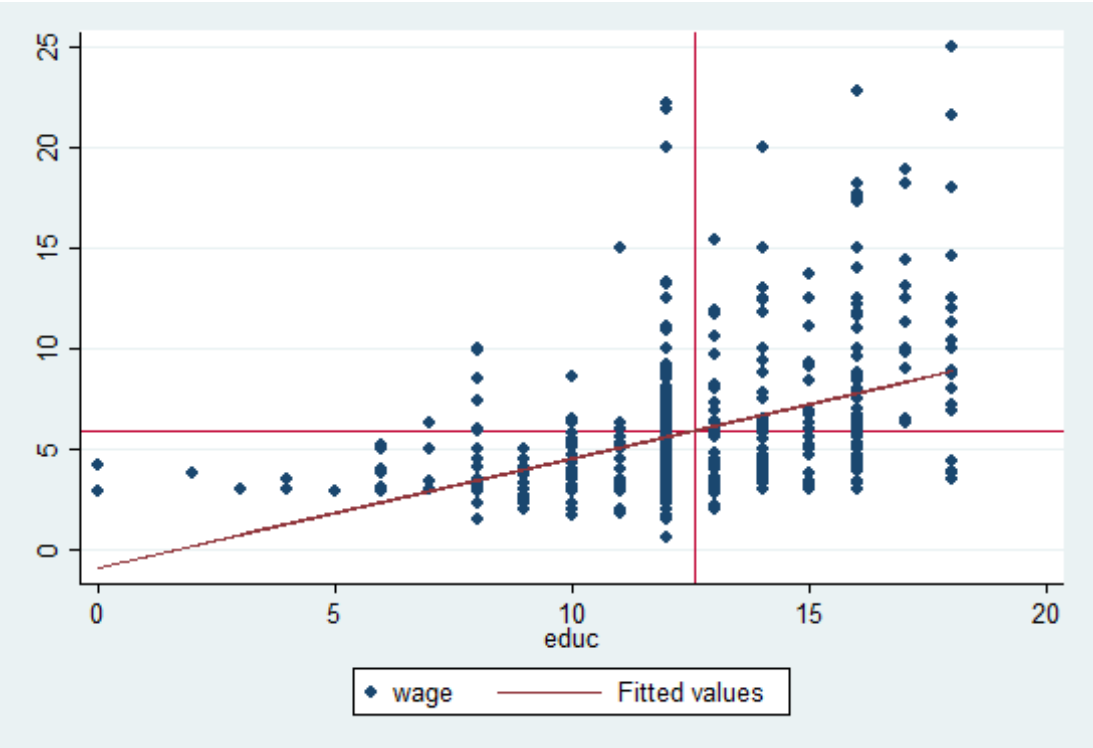
Ejemplo: *use http://fmwww.bc.edu/ec-p/data/wooldridge/wage1, clear*

reg wage educ (para correr la regresión, $wage_i = \beta_1 + \beta_2 educ_i + u_i$)

predict wage_hat (para predecir los salarios, $\widehat{wage} = \hat{\beta}_1 + \hat{\beta}_2 educ$)

scatter wage educ || line wage_hat educ, xline(12.57) yline(5.90)

(hace un gráfico con la nube de puntos y la línea de regresión)



Heteroscedasticidad

- En general, cuando hay heteroscedasticidad, los errores estándar de MCO están subestimados.
- Comparemos la varianza calculada asumiendo homoscedasticidad

$$V_1 = \frac{\bar{\sigma}^2(\sum_{i=1}^{N}(x_i - \bar{x})^2)}{(\sum_{i=1}^{N}(x_i - \bar{x})^2)^2}$$

donde $\bar{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2(x_i)$ y con la varianza (real) con heteroscedasticidad

$$V_2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2\sigma_i^2(x_i)}{(\sum_{i=1}^{N}(x_i - \bar{x})^2)^2}$$

Ahora dividiendo uno sobre otro:

$$V_1 / V_2 = \frac{\bar{\sigma}^2(\sum_{i=1}^{N}(x_i - \bar{x})^2)}{\sum_{i=1}^{N}(x_i - \bar{x})^2\sigma_i^2(x_i)}$$

.

Heteroscedasticidad

Ejemplo 1: x_i es $N(0,1)$, u_i es $N(0,x_i^2)$, $y_i = x_i + u_i$.
Ejemplo 2: x_i es $N(0,1)$, u_i es $N(0,i)$, $y_i = x_i + u_i$.
Ejemplo 3: x_i es $N(0,1)$, u_i es $N(0,1/x_i^2)$, $y_i = x_i + u_i$.

Varianza de MCO bajo homoscedasticidad

Teorema: Bajo los Supuestos 1-5,

$$Var(\hat{\beta}_j|\mathbf{x}) = \frac{\sigma^2}{SCT_j(1-R_j^2)},\ j = 2, ..., K$$

donde $SCT_j = \sum_{i=1}^{N}(x_{ji} - \bar{x}_j)^2$ es la variación total en x_j y R_j^2 es el R-cuadrado de una regresión de x_j en todas las otras variables (inluyendo el intercepto) $\{1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_K\}$.

En notación matricial es más simple $Var(\hat{\boldsymbol{\beta}}|\mathbf{x}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
Prueba:
$Var(\hat{\boldsymbol{\beta}}|\mathbf{x}) = Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var[\boldsymbol{y}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} =$
$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
Acá la clave es que $Var[\boldsymbol{y}|\mathbf{X}] = Var[\boldsymbol{u}|\mathbf{X}] = E[\boldsymbol{u}\boldsymbol{u}'|\mathbf{X}] = \sigma^2\boldsymbol{I}_N$.

Varianza de MCO bajo heteroscedasticidad

$Var(\hat{\boldsymbol{\beta}}|\mathbf{x}) = Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var[\boldsymbol{y}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var[\boldsymbol{u}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$
donde

$$\Omega := Var[\boldsymbol{u}|\mathbf{x}] = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{bmatrix}$$

Esta es la típica fórmula sandwich.

Contraste de heterocedasticidad

- Consideremos un contraste con hipótesis nula: $H_0 : Var(u|\boldsymbol{x}) = \sigma^2$, que es equivalente a $H_0 : E(u^2|\boldsymbol{x}) = \sigma^2$ (¿Por qué?) La hipótesis alternativa es $H_A : Var(u_i|\boldsymbol{x}_i) = \sigma_i^2 \neq Var(u_j|\boldsymbol{x}_j) = \sigma_j^2$ para al menos un par $i \neq j$.
- Una forma de contrastar por esto es chequear si u^2, el error al cuadrado, está relacionado con x. Para ello consideremos la siguiente regresión:

$$u^2 = \delta_1 \, x_1 + ... + \delta_k \, x_K + v$$

y contrastemos

$$H_0 : \delta_2 = ... = \delta_K = 0$$

$$H_A : \delta_j \neq 0, \quad \textit{al menos un } j.$$

- ¿Por qué δ_1 es omitido? (recordemos $x_1 = 1$ es la constante)

Contraste de heterocedasticidad

- Pero tenemos un problema importante.... No observamos u.
- Sin embargo, el modelo de regresión lo puede reemplazar:

$$\hat{u}^2 = \delta_1 \, x_1 + ... + \delta_k \, x_K + v$$

y contrastamos

$$H_0 : \delta_2 = ... = \delta_k = 0$$

- Estos contrastes son los de Breusch y Pagan (1980) y Koenker (1983). Breusch y Pagan (1980) asumen normalidad de los errores, o sea $u \sim N(0, \sigma^2)$. Sin embargo, esto no funciona si los errores no son normales, y para eso esta el contraste de Koenker (1983).

$$F = \frac{R^2_{\hat{u}^2} / K - 1}{(1 - R^2_{\hat{u}^2}) / (n - k - 1)} \overset{a}{\sim} F_{K-1,N-K}$$

$$LM = N \cdot R^2_{\hat{u}^2} \overset{a}{\sim} \chi^2_{K-1}$$

donde $\overset{a}{\sim}$ significa "asintóticamente distribuido"

STATA

- `reg y x2 ... xK` (regresión principal)
- `predict u, resid` (para obtener los residuos)
- `gen u2=u^2` (el cuadrado de los mismos)
- `reg u2 x2... xK` (regresión auxiliar)
- Test F:
  `test x2 ... xK`
- Test LM:

```
scalar LM=e(N)*e(r2)
scalar pvalueLM=chi2tail($K-1,LM)
display "Robust LM statistic :  " LM
display "p-value :  " pvalueLM
```

Inferencia robusta con heterocedasticidad

- Supongamos un modelo de regresión simple $y_i = \beta_1 + \beta_2 x_i + u_i, i = 1, 2, ..., N$.
- Un estimador de

$$Var(\hat{\beta}_2) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}$$

es

$$\widehat{Var(\hat{\beta}_2)} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2},$$

$\sqrt{Var(\hat{\beta}_2)}$ se llama el error estándar robusto a heterocedasticidad dado por el paper de White (1980). Es válido porque $\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2 \hat{u}_i^2}{SST_x^2} \stackrel{a}{=} \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2 u_i^2}{SST_x^2}$.

Normalidad asintótica

Si tenemos heteroscedasticidad:

- Entonces tenemos

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} Normal(\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1})$$

- $\boldsymbol{A} = E[\boldsymbol{x}'\boldsymbol{x}]$, que se puede estimar con $\hat{\boldsymbol{A}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i'\boldsymbol{x}_i$.
- En este caso necesitamos un estimador consistente de $\boldsymbol{B} = E[\boldsymbol{x}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{x}]$: $\hat{\boldsymbol{B}} \equiv N^{-1}\sum_{i}^{N} \hat{u}_i^2 \boldsymbol{x}_i'\boldsymbol{x}_i$.
- Así, $AVar(\hat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}$. Sin embargo todavia tenemos que proponer un estimador de la varianza, $\widehat{AVar(\hat{\boldsymbol{\beta}})} = \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{A}}^{-1}$.
- La inferencia con este estimador es el estimador robusto de White.

Inferencia robusta con heterocedasticidad

- Si la heterocedasticidad fuera conocida en su forma funcional:

$$Var(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$$

donde $h(.) > 0$ es la forma funcional de la heterocedasticidad.

Notemos que

$$Var(u/\sqrt{h(\boldsymbol{x})}|\boldsymbol{x}) = \sigma^2$$

entonces $u/\sqrt{h(\boldsymbol{x})}$ es homocedástico. Entonces tenemos que la siguiente regresión satisface los supuestos de Gauss-Markov:

$$y_i/\sqrt{h_i} = \beta_1/\sqrt{h_i} + \beta_2 x_2/\sqrt{h_i} + ... + \beta_K x_K/\sqrt{h_i} + u_i/\sqrt{h_i}$$

STATA

- `reg y x2 ... xK, robust` (corrección de los errores estándar de White (1980))
- `gen h=h(x)` (Usar la función h(.), que debe ser conocida de antemano)
- `reg y x2 ... xK [aw=1/h]` (MCO con pesos muestrales)

Ejemplos en la web

- `http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge12.html`
- `https://www.stata.com/manuals13/p_robust.pdf`