Maestría en Ciencia y Análisis de Datos- Universidad Mayor de San Andrés

Modelos lineales y modelos lineales generalizados

Rolando Gonzales Martinez, PhD

Fellow postdoctoral Marie Skłodowska-Curie

Universidad de Groningen (Países Bajos)

Investigador (researcher)

Iniciativa de Pobreza y Desarrollo Humano de la Universidad de Oxford (UK)

Contenido del curso

(2) Diagnóstico y Evaluación de Modelos Lineales

- Diagnóstico de residuos.
- Multicolinealidad.
- Otras patologías que pueden surgir en los modelos lineales.
- Métricas de evaluación de ajuste.
- Laboratorio: Evaluación de modelos lineales

Teorema Gauss-Markov

En un modelo lineal de regresión múltiple (MLRM), si se cumplen las siguientes condiciones:

$$y = X\beta + \epsilon$$
$$E(\epsilon) = 0$$
$$Var(\epsilon) = \sigma^2 I$$
$$Cov(\epsilon_i, \epsilon_j) = 0$$
$$Cov(X, \epsilon) = 0$$

1. **Linealidad**: El modelo es lineal en los parámetros
2. Los errores tienen media cero
3. **Homocedasticidad**: Los errores tienen varianza constante
4. **No autocorrelación**: Los errores no están correlacionados entre sí
5. **Independencia entre X y ϵ**: Los errores no están correlacionados con las variables explicativas

Entonces, el estimador de mínimos cuadrados ordinarios es el mejor estimador lineal insesgado (MELI, BLUE). Esto significa que:

- **Insesgado**: $E(\hat{\beta}) = \beta$
- **Varianza mínima**: Entre todos los estimadores lineales insesgados, el estimador MCO tiene la varianza más baja.

Normalidad de los residuos

- La normalidad de los residuos no es parte del Teorema Gauss-Markov, por lo que no es necesaria para que los estimadores MCO/OLS sean MELI/BLUE.
- Sin embargo la normalidad de residuos es relevante para la docimasia de hipótesis y para la construcción de intervalos de confianza precisos.
- Si los residuos son normales los estimadores MCO coinciden con los estimadores máximo verosímiles y son eficientes en el sentido de tener varianza mínima.

Diagnóstico de residuos

Diferentes pruebas estadísticas pueden aplicarse a los residuos para testear normalidad y los supuestos del teorema Gauss-Markov respecto a los residuos de la regresión:

	Test	Hipótesis nula
Normalidad	Jarque-Bera	Normalidad
Linealidad	Harvey-Collier	Linealidad
Homocedasticidad (heteroscedasticidad)	Breusch–Pagan	Homocedasticidad
	Goldfeld-Quandt	Homocedasticidad
	White	Homocedasticidad

Test Jarque-Bera

- El estadístico de Jarque-Bera se distribuye asintóticamente como una distribución chi-cuadrado con dos grados de libertad
- La hipótesis nula es una hipótesis conjunta de que la asimetría y el exceso de curtosis son nulos (asimetría = 0 y curtosis = 3).

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

Linealidad: test Harvey-Collier

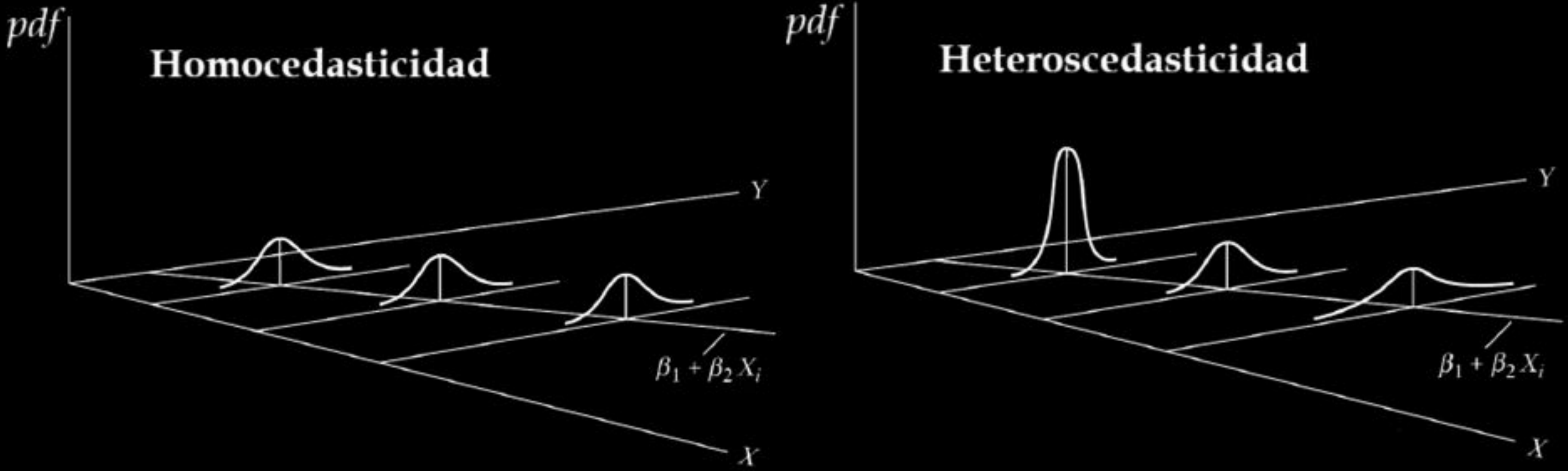
$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i$$

$$\hat{u}_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}} \qquad \hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

$$HC = \frac{\frac{1}{n} \sum_{i=1}^{n} \hat{u}_i}{\hat{\sigma}/\sqrt{n}} \sim t\,(n - k - 1)$$

Homocedasticidad y Heterocedasticidad

$$Y_i = \beta_1 + \beta_2 X_i$$



Afecta eficiencia y validez de la inferencia

Homocedasticidad: Test Breusch–Pagan

$$y_i = X_i\beta + \varepsilon_i, \quad i = 1, \ldots, n$$

$$g_i = \hat{\varepsilon}_i^2/\hat{\sigma}^2, \quad \hat{\sigma}^2 = \sum \hat{\varepsilon}_i^2/n$$

$$g_i = \gamma_1 + \gamma_2 z_{2i} + \cdots + \gamma_p z_{pi} + \eta_i$$

$$\gamma_2 = \cdots = \gamma_p = 0$$

$$LM = \frac{n \cdot R^2}{2}$$

$$F = \frac{\left(SSR_{\text{reducido}} - SSR_{\text{completo}}\right)/k}{SSR_{\text{completo}}/(n-k-1)}$$

Homocedasticidad: Test Goldfeld-Quandt

Ordenar los datos y dividir en 2 grupos:

Para el primer grupo:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i$$

Para el segundo grupo:

$$y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_k X_{kj} + \epsilon_j$$

$$RSS_1 = \sum_{i \in \text{Grupo 1}} \hat{\epsilon}_i^2$$
$$RSS_2 = \sum_{j \in \text{Grupo 2}} \hat{\epsilon}_j^2$$

$$F = \frac{\frac{RSS_2}{df_2}}{\frac{RSS_1}{df_1}}$$

Homocedasticidad: Test de White

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

$$\hat{\epsilon}^2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_k X_k + \alpha_{k+1} X_1^2 + \alpha_{k+2} X_2^2 + \ldots + \alpha_{2k} X_k^2 + \alpha_{2k+1} X_1 X_2 + \ldots + v$$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$$\hat{\epsilon}^2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1^2 + \alpha_4 X_2^2 + \alpha_5 X_1 X_2 + v$$

$$LM = n \cdot R^2 \sim \chi^2_{(d)}$$

Heterocesdasticidad: soluciones

- Transformaciones de variables, WLS
- Matriz varianza-covarianza robusta:
    - HC0: Matriz varianza-covarianza robusta clásica de White. No ajusta los grados de libertad. Adecuada para muestras grandes.
    - HC1: Ajuste de MacKinnon y White (1985) para muestras pequeñas.
    - HC2, HC3: Ajustes alternativos para muestras pequeñas.

Heterocesdasticidad: soluciones

$$\text{HC0} = X'X^{-1}X'\text{diag}(u_i^2)XX'X^{-1}$$

$$\text{HC1} = \left(\frac{n}{n-k}\right)\text{HC0}$$

$$\text{HC2} = X'X^{-1}X'\text{diag}\left(\frac{u_i^2}{1-h_i}\right)XX'X^{-1}$$

$$H = X(X'X)^{-1}X'$$

$$\text{HC3} = X'X^{-1}X'\text{diag}\left(\frac{u_i^2}{(1-h_i)^2}\right)XX'X^{-1}$$