Supuestos Regresión Lineal: Heteroscedasticidad

Clase 28

Nicolás Mejía M.
n.mejia10@uniandes.edu.co

Probabilidad y Estadística II
Departamento de Ingeniería Industrial
Universidad de Los Andes, Bogotá, Colombia

2020-19

Outline

Definición

Cuando planteamos el modelo de regresión lineal,

$$Y = \beta_0 + \beta_1 X + \cdots + \beta_k X_k + \epsilon$$

hicimos el supuesto de que los errores aleatorios:

$$\epsilon \sim N(0, \sigma^2)$$

es decir, la dispersión de las observaciones sobre la recta de regresión es la misma a lo largo de toda al recta. Es decir, se tiene una situación de HOMCEDASTICIDAD.

Pero esto no es verdad siempre.

Definición

Ejemplo

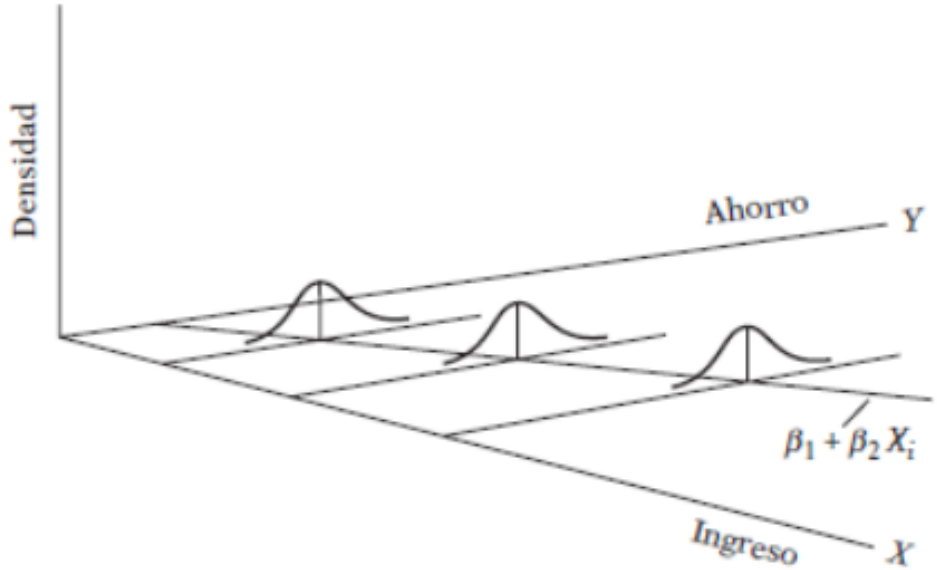
Suponga que quiere explicar el ingreso de un egresado de ingeniería industrial de la universidad (Y), en función de sus años de experiencia(X).

- Para los primeros años de experiencia se tendrán salarios relativamente homogéneos entre los individuos pues son casi indistinguibles.
- A medida que avanza la experiencia, las diferencias entre salarios se hace más notoria por factores como crecimiento profesional, otros estudios, etc.

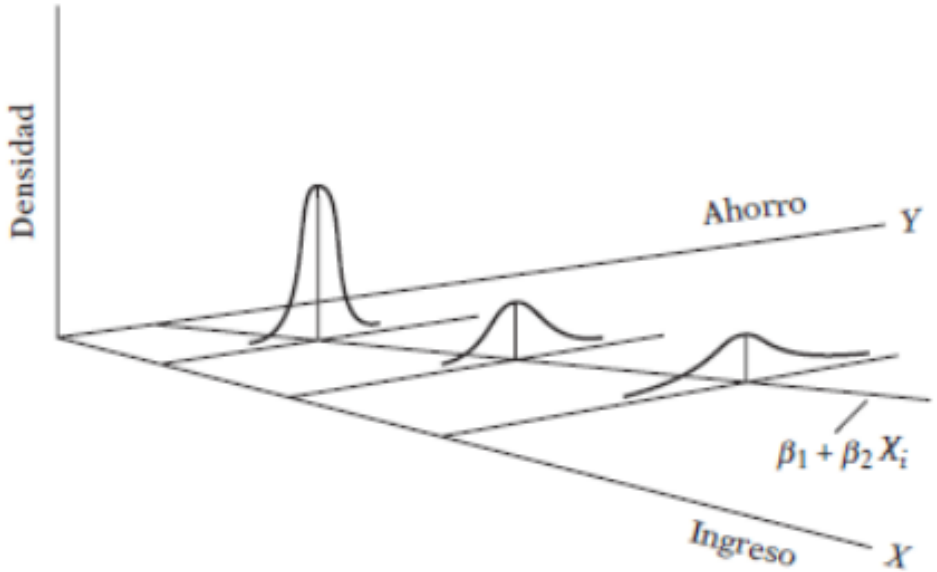
Realmente no hay motivo por el cual se deba tener una varianza constante, y el supuesto puede ser terriblemente equivocado.

Definición

Perturbaciones homoscedásticas.



Perturbaciones heteroscedásticas.



Definición

De esta forma realmente se tiene una situación de la forma:

$$Y = \beta_0 + \beta_1 X + \cdots + \beta_k X_k + \epsilon_i$$

donde ahora los errores aleatorios:

$$\epsilon_i \sim N(0, \sigma_i^2)$$

es decir, la dispersión de las observaciones sobre la recta CAMBIA a lo largo de toda al recta.

En este caso decimos que hay HETEROCEDASTICIDAD

Definición

Definiciones

- Homocedasticidad: Decimos que existe homocedasticidad si la varianza es homogénea en los tratamientos, es decir si σ^2 es constante.
- Heterocedasticidad: Decimos que existe heterocedasticidad si la varianza NO es homogénea en los tratamientos, es decir si σ^2 es cambiante (i.e $\sigma^2 = \sigma^2_{ij}$).

¿Por qué decimos que la heterocedasticidad es un problema? Para esto veamos en donde aparece σ^2 en nuestros cálculos

Definición

Los supuestos originales eran:

$$E(\epsilon) = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

$$Var(\epsilon) = \begin{bmatrix} Var(\epsilon_1) & Cov(\epsilon_1, \epsilon_2) & \cdots & Cov(\epsilon_1, \epsilon_n) \\ Cov(\epsilon_1, \epsilon_2) & Var(\epsilon_2) & \cdots & Cov(\epsilon_2, \epsilon_n) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(\epsilon_1, \epsilon_n) & Cov(\epsilon_2, \epsilon_n) & \cdots & Var(\epsilon_n) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

Definición

Bajo la presencia de heteroscedasticidad se tendría:

$$E(\epsilon) = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

$$Var(\epsilon) = \begin{bmatrix} Var(\epsilon_1) & Cov(\epsilon_1, \epsilon_2) & \cdots & Cov(\epsilon_1, \epsilon_n) \\ Cov(\epsilon_1, \epsilon_2) & Var(\epsilon_2) & \cdots & Cov(\epsilon_2, \epsilon_n) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(\epsilon_1, \epsilon_n) & Cov(\epsilon_2, \epsilon_n) & \cdots & Var(\epsilon_n) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} = \sigma^2 \mathbf{W} \neq \sigma^2 \mathbf{I}$$

Outline

Implicaciones

Varianza de coeficientes

$$Var(\hat{\beta}) = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

La expresión encontrada para la varianza deja de ser valida!

Prueba F

El estadístico $F = \frac{MSR}{MSE}$ ya no tiene la distribución F que habíamos determinado!. Para esto necesitaríamos conocer la estructura de varianzas por observación, cosa que no tenemos.

Estimador de la Varianza

De forma análoga, es claro entonces que el mejor estimador de la varianza ya NO es el MSE debido a que no existe una única varianza!

Implicaciones

Inferencia Estadística

Ahora bien, si el MSE no tiene sentido, entonces nuestros procedimientos de inferencia para combinaciones lineales tampoco:

$$\hat{\theta} = \sum_{j=0}^{k} c_j \hat{\beta}_j = \mathbf{c}^T \hat{\beta}_j \rightarrow \hat{Var}(\hat{\beta}) = \mathbf{c}^T MSE \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{c}$$

$$\text{Pruebas de hipótesis } t = \frac{\hat{\theta} - \theta_0}{\sqrt{\hat{Var}(\hat{\beta})}} \sim t_{gIE}$$

$$\text{Intervalos de confianza } IC_{1-\alpha/2}(\theta) = \hat{\theta} \pm t_{1-\alpha/2,gIE} \sqrt{\hat{Var}(\hat{\beta})}$$

Implicaciones

- Las pruebas de significancia individual y global no son creíbles: la varianza puede estar sub o sobrestimandose.
- Los intervalos de confianza tienen amplitudes erroneas! Ya no se puede asegurar su verdadero nivel de confianza.
- En resumen, se pierde la posibilidad de hacer las conclusiones con el nivel de significancia que se afirma.
- Los estimadores siguen siendo insesgados, pero dejan de ser eficientes!

Outline

Identificación

Para saber si hay problemas de Heterocedasticidad, hay estrategias gráficas y analíticas. La mayoria se basan en los residuales. La heterocedasticidad está asociada a la varianza de los errores aleatorios ϵ, y el mejor representante de estos valores son los residuales:

$$e_i = Y_i - \hat{Y}_i$$

teniendo en cuenta que $\sigma^2 = Var(\epsilon_i) = E(\epsilon_i^2) - E(\epsilon_i)^2 = E(\epsilon_i^2)$, un estimador de la varianza de la i-esima observación está dado por:

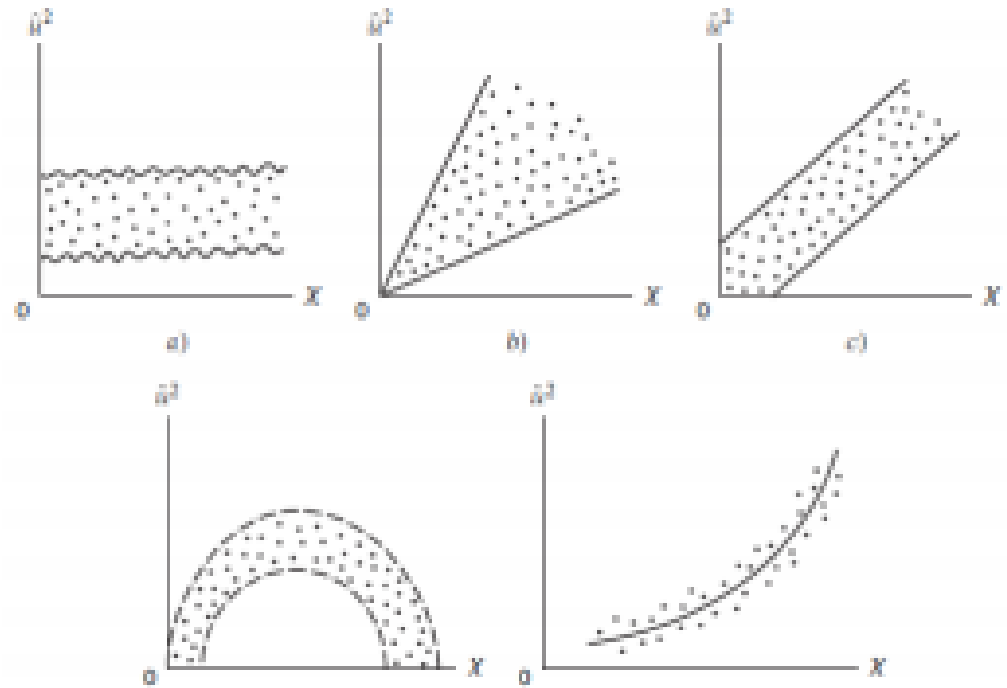
$$e_i^2 = (Y_i - \hat{Y}_i)^2$$

Si hay homocedasticidad, los ϵ_i^2 deberían ser homogenos entre ellos, pero en caso de heterocedasticidad, estos mostrarán patrones.

Identificación

Métodos Gráficos

La varianza debe ser constante, luego si se grafican los e_i^2 vs X_j, Y Y $\hat{Y}$, estos deben mantener un comportamiento homogéneo. De lo contrario, hay problemas de heterocedasticidad:



Identificación

Métodos analíticos

Las pruebas estadísticas NO pueden probar la presencia de heterocedasticidad de forma general, sino que prueban si esta se da de una forma en especifico:

Suponga que se quiere ver si la varianza de los errores se da de la forma:

$$Var(\epsilon_i) = \sigma_i^2 = \sigma^2 f(X_{1i}, \cdots, X_{Ki}, E(Y_i))$$

entonces estime la regresión:

$$e_i^2 = \alpha_o + \alpha_1 f(X_{1i}, \cdots, X_{ki}, E(Y_i)) + \eta_i$$

Si es globalmente significativa, hay evidencia estadística de que existe heterocedasticidad. En caso de no ser significativa, no hay motivo para creer que existe heterocedasticidad de la FORMA PLANTEADA.

Identificación

Métodos analíticos

Formalmente se tendría lo siguiente:

$$H_0 : \text{No hay Heterocedasticidad de la forma } f(X_1, \cdots, X_k, E(Y))$$

$$H_1 : \text{Si hay Heterocedasticidad de la forma } f(X_1, \cdots, X_k, E(Y))$$

Estime la regresión auxiliar $e_i^2 = \alpha_o + \alpha_1 f(X_{1i}, \cdots, X_{ki}, E(Y_i)) + \eta_i$, calcule su R^2_{aux} y halle el estadístico de prueba:

$$nR^2_{aux} \sim \chi^2_m$$

y utilice como región de rechazo la cola superior $\chi^2_{1-\alpha,m}$ donde m corresponde a los grados de libertad de la regresión auxiliar.

Identificación/Test de White

Test de White

Suponga que usted quiere probar si existen problemas de heterocedasticidad en el modelo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$.

White sugiere utilizar como modelo auxiliar la regresión en la que se incluyen los términos lineales, cuadráticos y además los términos cruzados:

$$e_i^2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_1^2 + \alpha_5 X_2^2 + \alpha_6 X_3^2 + \alpha_7 X_1 X_2 + \alpha_8 X_1 X_3 + \alpha_9 X_2 X_3 + \eta$$

Esto permite capturar gran cantidad de posibilidades en las que se puede presentar la heterocedasticidad.

En nuestro ejemplo, el estadístico de prueba nR^2_{aux} tendría $m = 9$ grados de libertad.

Identificación/Test de Breusch-Pagan

Test de Breusch-Pagan

Suponga que usted quiere probar si existen problemas de heterocedasticidad en el modelo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$.

Breusch-Pagan, en su versión tradicional, sugiere utilizar como modelo auxiliar la regresión en la que se incluye $\hat{Y}$, como estimación de $E(Y)$

$$e_i^2 = \alpha_0 + \alpha_1 \hat{Y}_i + \eta_i$$

Observe que no se incluye ningún valor de los Xs. En nuestro ejemplo, el estadístico de prueba $BP = nR^2_{aux}$ tendría $m = 1$ grado de libertad.

El test también puede aplicarse de forma individual para cada variable.

Identificación/Test de Breusch-Pagan

Test de Breusch-Pagan

En el caso en que se desee identificar si la heterocedasticidad se produce por una variable en específico, considere las siguientes regresiones auxiliares:

$$e_i^2 = \alpha_0 + \alpha_1 X_{1i} + \eta_i \rightarrow BP_1 = nR^2_{aux1}$$

$$e_i^2 = \alpha_0 + \alpha_2 X_{2i} + \eta_i \rightarrow BP_2 = nR^2_{aux2}$$

$$e_i^2 = \alpha_0 + \alpha_3 X_{3i} + \eta_i \rightarrow BP_3 = nR^2_{aux3}$$

En este caso, se tendría un estadístico de BP para cada variable, y cada uno con $m = 1$ grado de libertad. Las variables que en la prueba estadística rechacen la hipótesis nula, se pueden considerar como variables problemáticas que causan la heterocedasticidad.

Outline

Corrección

Para corregir la heterocedasticidad, es necesario identificar en primer lugar como se presenta. Es decir, es necesario identificar la función f tal que:

$$\sigma_i^2 = \sigma^2 f(X_{1i}, \cdots, X_{ki}, E(Y_i)) = \sigma^2 f$$

Esto se hace por medio de las pruebas anteriores. En especial por medio de la pruebas de Breusch-Pagan para cada variable, y el test de White. Dedique tiempo para esta tarea.

Con eso determinado, el proceso de corrección es directo!

Correción

Cree las siguientes nuevas variables:

$$\tilde{Y} = \frac{Y}{\sqrt{f}}$$

$$\tilde{X_0} = \frac{1}{\sqrt{f}}$$

$$\tilde{X_j} = \frac{X_j}{\sqrt{f}} \ \forall j = 1, \cdots, k$$

$$\tilde{\epsilon} = \frac{\epsilon}{\sqrt{f}}$$

y estime la regresión en términos de las nuevas variables:

$$\tilde{Y} = \beta_0 \tilde{X_0} + \beta_1 \tilde{X_1} + \cdots + \beta_k \tilde{X_k} + \tilde{\epsilon}$$

Correción

Si el trabajo está bien hecho, esta regresión ya no tiene problemas de heterocedasticidad:

$$Var(\tilde{\epsilon}) = Var\left(\frac{\epsilon}{\sqrt{f}}\right) = \frac{Var(\epsilon)}{f} = \frac{\sigma^2 f}{f} = \sigma^2$$

luego la estimación de estos coeficientes si es acertada, y por tanto se pueden hacer todos los procedimientos tradicionales sin problemas.

No olvide revertir la transformación cuando trabaje con la regresión! OJO con los coeficientes

Correción

Otras metodologías

- Mínimos cuadrados ponderados: Consiste en estimar los coeficientes minimizando una versión ponderada del SSE:

$$\sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{\sigma_i} \right)^2$$

  Eligiendo σ_i estratégicamente esto sería equivalente a lo que ya hicimos.

- Mínimos cuadrados robustos: Este estimador trata de hallar la verdadera distribución $\hat{\beta}$ al estimar la varianza de cada uno de los errores de forma individual. Desgraciadamente, el estimador NO es eficiente y se tienen intervalos innecesariamente grandes.