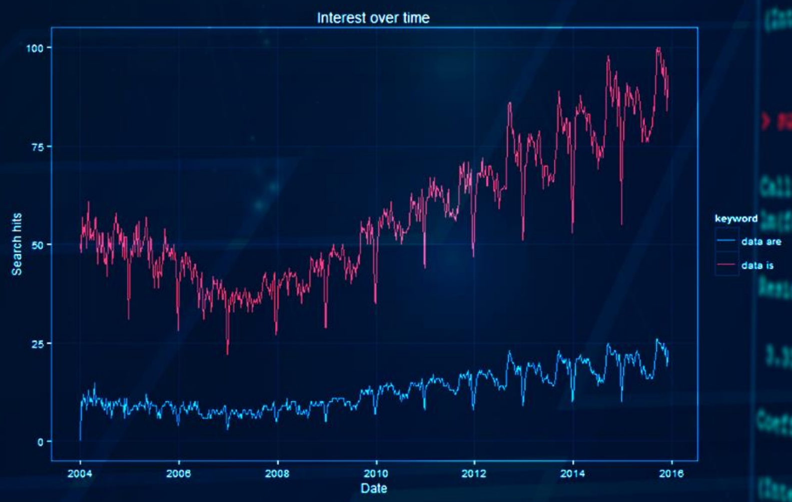
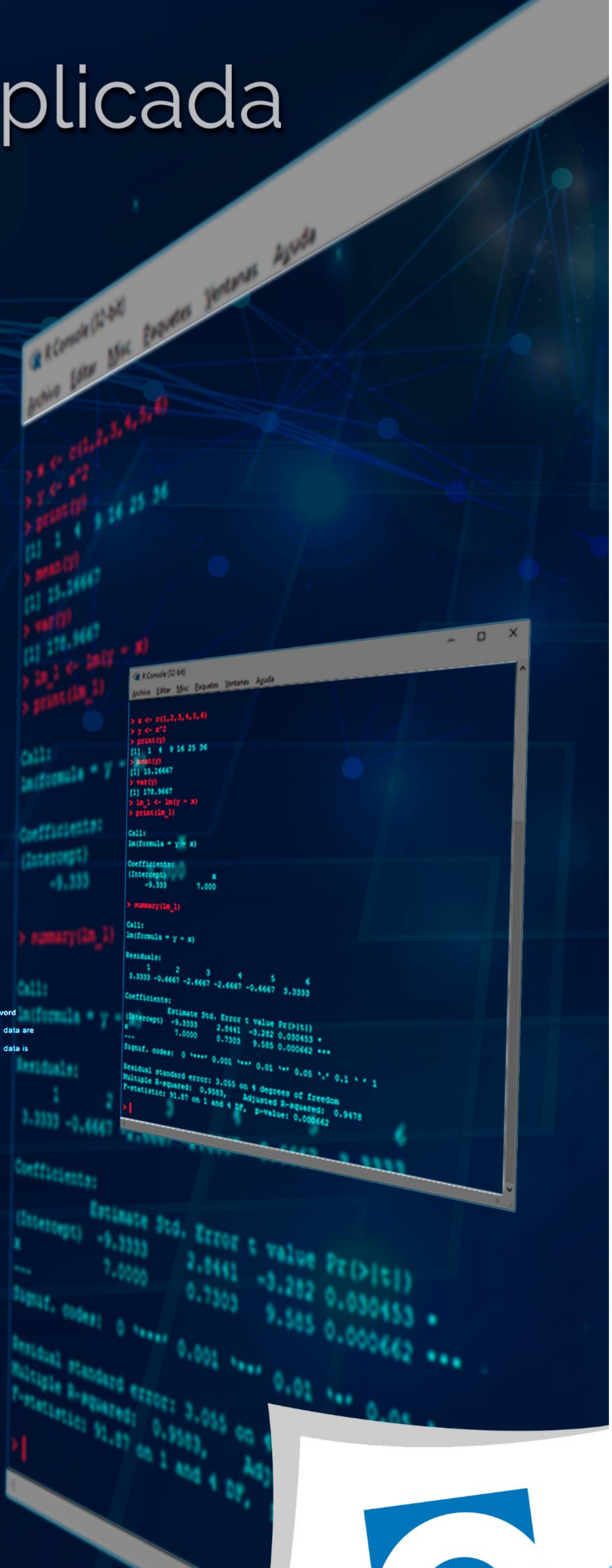
Econometría Aplicada con R





INSTITUTO CIENTÍFICO DEL PACÍFICO

SESIÓN 9: Heteroscedasticidad

Contenido

Introducción

Un supuesto importante del modelo clásico de regresión lineal es que las perturbaciones que aparecen en la función de regresión poblacional son homoscedásticas; es decir, que todas tienen la misma varianza.

Generalmente este problema se presenta cuando se trabaja con datos de corte transversal, pues se suele tomar como muestra a unidades económicas de diferentes tamaños.

En la siguiente sesión se verá la teoría relacionada a la violación del supuesto de homoscedasticidad, la detección de la misma y las soluciones que se pueden plantear para corregir este problema.

Heteroscedasticidad

El modelo básico de regresión lineal tiene como supuesto que la varianza del término de error se constante:

$$Var(u_i) = \sigma^2$$

Se debe tener en cuenta dicho supuesto ya que cuando se habla de estimación de parámetros en el modelo de regresión lineal, estos deben ser o mostrar una relación estable entre los predictores y la variable dependiente, por lo que los valores muestrales de la variable Y deben mostrarse igualmente dispersos ante variaciones de los regresores.

Causas de la heteroscedastidad

Existen diversas situaciones por la cuales se produce la violación al supuesto de homoscedasticidad en el modelo de regresión lineal, entre las principales se pueden enumerar las siguientes:

Valores dispersos

Esta es la situación más frecuente por la que se presenta el problema de heteroscedasticidad en el modelo de regresión. Se da porque frecuentemente las observaciones de los datos en el modelo no suelen tener un comportamiento homogéneo.

Omisión de variables relevantes

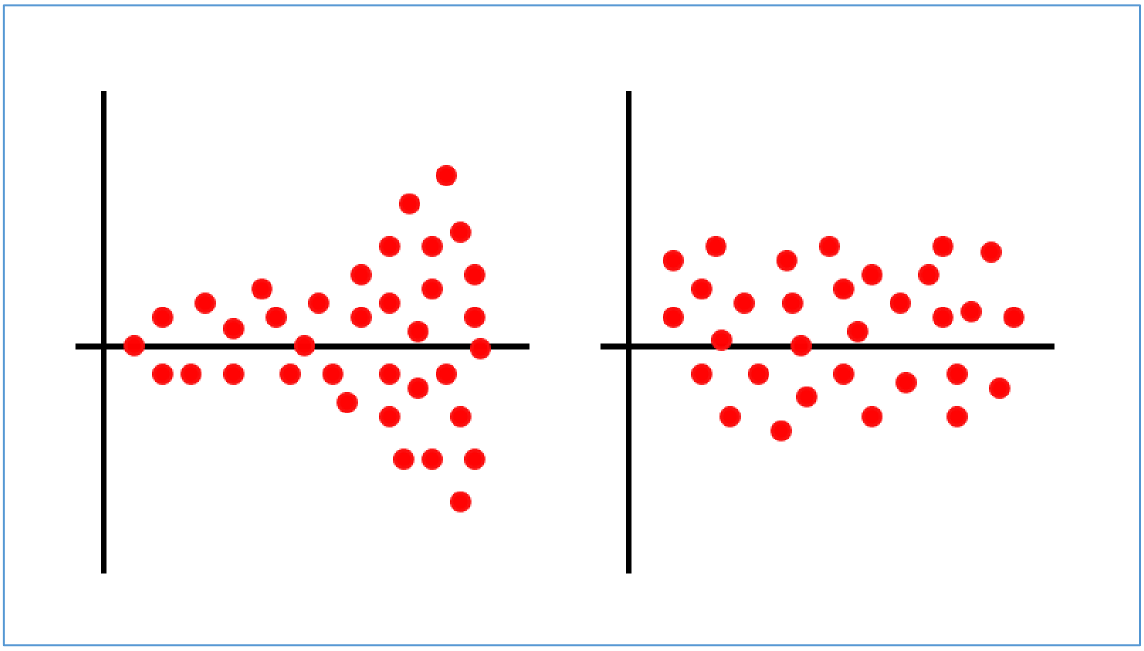
Si se omite una variable en la especificación, dicha variable quedará parcialmente recogida en el comportamiento de los errores aleatorios, pudiendo generar heteroscedasticidad en el modelo.

Cambio de estructura

Un cambio de estructura en el modelo va a hacer que el modelo esté mal ajustado, esto generará una varianza no constante en los términos de error del modelo estimado.

Forma funcional incorrecta

Cuando se especifica un modelo de forma incorrecta respecto a su forma funcional, se puede provocar que la calidad del ajuste de la regresión es muy baja, porque no sólo tendrá errores mayores, si no también errores muy dispersos.



La imagen muestra cómo se distribuye el error, a la izquierda se puede notar que la varianza no es constante (heteroscedasticidad), y a la derecha se nota que si lo es (homoscedasticidad).

Consecuencias de la heteroscedasticidad

Existen diversas consecuencias por la estimación de un modelo de regresión ignorando la presencia de homoscedasticidad. Las principales son las siguientes:

- Existirá un error en el cálculo del estimador en la matriz de varianzas y covarianzas de los estimadores de mínimos cuadrados.
- El estimador del modelo perderá eficiencia.

Detección de la heteroscedasticidad

Existen diversos métodos para detectar el problema de la heteroscedasticidad, se puede recurrir a los contrastes o a métodos gráficos.

Métodos gráficos

Entre los métodos gráficos se tiene:

Gráfica del error a través las observaciones

Un simple gráfico en donde el error del modelo recorre las observaciones puede mostrar la distribución de las perturbaciones en el modelo, esto sirve para observar si al trascurrir las observaciones se da lugar a un aumento o disminución del error, lo que puede ser un indicativo de heteroscedasticidad.

En R, se deberá generar el vector de los residuos del modelo una vez estimado el mismo, luego de esto se hará un gráfico de líneas teniendo en cuenta a los residuos y las observaciones.

```
modelo = lm(gashog2d ~ inghog2d + mieperho,
            data = sumaria_2015)
```

Una vez especificado el modelo se deben crear los residuos del mismo, esto se hará con la opción **residuals** en el modelo.

```
resid = modelo$residuals
```

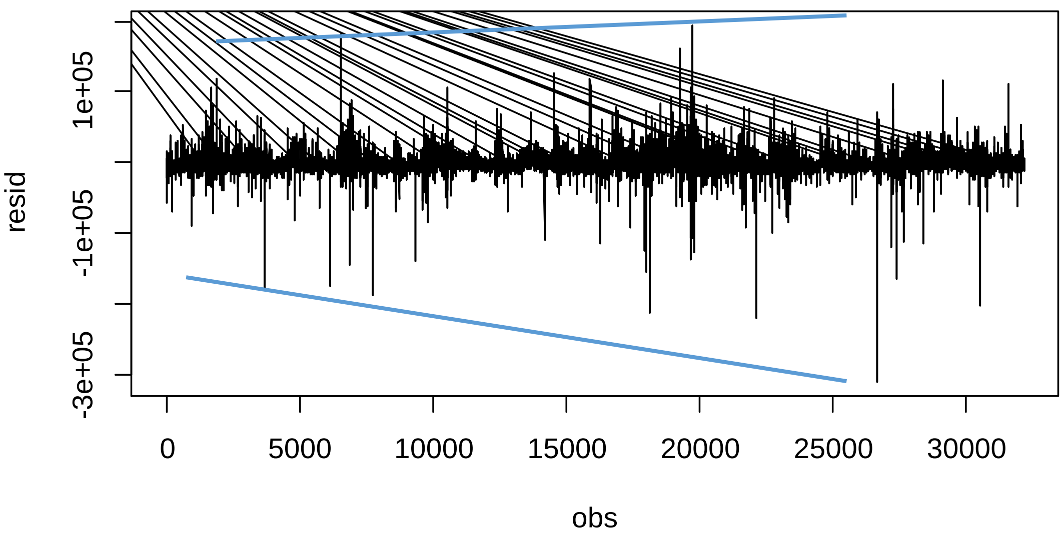
De acuerdo al número de observaciones se generará un vector columna que contenga a los números de cada observación:

```
obs = c(1:32188)
```

Luego de haber generado los dos elementos se hará un gráfico de líneas, para ellos se usará el comando **plot()**, y en una de las opciones del gráfico se especificará que se quiere un gráfico de líneas.

```
plot(obs, resid, type="l")
```

Esto mostrará el siguiente gráfico:



En este gráfico se podrá observar que existe cierta tendencia al incremento en la dispersión de los residuos a lo largo de las observaciones.

Gráfico del error al cuadrado y de la variable endógena

Esta representación de los valores cuadráticos del término de error y de la variable endógena puede revelar la existencia de un patrón sistemático en la varianza de la perturbación. En este tipo de gráfico se tendrá una idea preliminar de la existencia de heteroscedasticidad en el modelo.

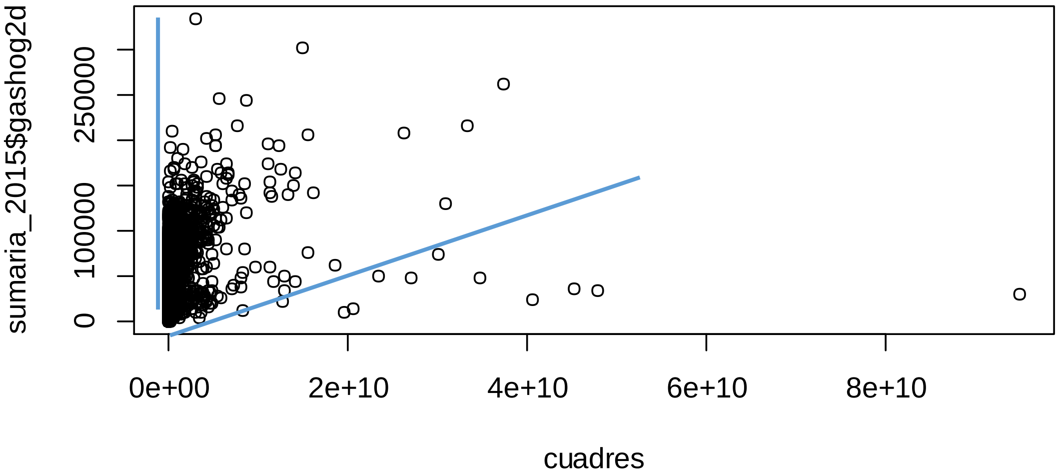
En R se deberá generar en base a los residuos del modelo, el cuadrado de los mismos.

```
cuadres = resid^2 # Se debe tener en cuenta que los residuos ya fueron generados de forma anterior.
```

Luego de eso se procede a realizar el gráfico de dispersión entre el cuadrado de los residuos y los valores observados de la variable dependiente en el modelo.

```
plot (cuadres, sumaria_2015$gashog2d)
```

Esta sentencia mostrará el siguiente gráfico:



Este gráfico muestra la existencia de cierto patrón, aunque muy disperso, por lo que no dice mucho acerca de la existencia de la heteroscedasticidad en el modelo de regresión.

Contrastes de heteroscedasticidad

Existen diversos contrastes o pruebas que permiten saber en base a una prueba de hipótesis si hay o no heteroscedasticidad.

Test de Breusch-Pagan

El test de Breusch-Pagan (1979) es una prueba muy popular para evaluar la heteroscedasticidad en el modelo de regresión. Este método analiza si la varianza de los residuos de un modelo de regresión depende de los valores de las variables independientes.

El test viene dado por la siguiente prueba:

$$H_0: Homoscedasticidad$$

$$H_1: Heteroscedasticidad$$

Suponiendo que se tiene el siguiente modelo:

$$y = \beta_0 + \beta_1 x + u_i$$

Se debe estimar los residuos $\hat{u}_i$, de acuerdo a lo estimado se debe estimar el siguiente modelo de regresión:

$$\hat{u}_i^2 = \gamma_0 + \gamma_1 x + v_i$$

Si el test, que vendrá dado por la prueba de significancia global F confirma que las variables independientes son significativas, entonces se puede rechazar la hipótesis nula de homoscedasticidad.

En R existe un comando que permite evaluar de forma directa la heteroscedasticidad en el modelo, la función es **bptest()**, que está en el paquete **lmtest**.

```
> bptest(modelo)

	studentized Breusch-Pagan test

data:  modelo
BP = 6799.3, df = 2, p-value < 2.2e-16
```

Los resultados del test indican que la hipótesis nula se rechaza y por lo tanto se entiende que el modelo presenta heteroscedasticidad.

Test de White

Esta es otra prueba de heteroscedasticidad. Donde la hipótesis a evaluarse viene dada por:

$$H_0: Homoscedasticidad$$

$$H_1: Heteroscedasticidad$$

Suponiendo el modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u_i$$

Para elaborar la prueba de White se siguen los siguientes pasos:

1. Se debe estimar el modelo y calcular en base a este los errores del mismo.
2. Se debe efectuar una regresión auxiliar, con los errores estimados como variable dependiente, de la siguiente forma:

$$\hat{u}_i = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1^2 + \alpha_4 x_2^2 + \alpha_5 x_1 x_2 + v_i$$

3. En base a eso se tendrá el R-Cuadrado de la regresión y construir el estadístico de prueba:

$$n \times R_{aux}^2 \sim \chi_{5g.l.}^2$$

Los grados de libertad de esta prueba van a depender del número de parámetros de la regresión auxiliar.

La regla de decisión está dada por la siguiente:

- Si $(n \times R_{aux}^2) > \chi_{kg.l.}^2$, se rechaza H_0.
- Si $(n \times R_{aux}^2) < \chi_{kg.l.}^2$, se acepta H_0.

Para realizar la prueba de White en R, se tienen que seguir los pasos generando y estimando el modelo, ya que no existe una función que de forma directa realice el test White en este programa.

Contraste de Goldfeld-Quandt

Esta prueba sirve para ver si hay o no heteroscedasticidad. Viene dada por la siguiente hipótesis:

$$H_0: Homoscedasticidad$$

$$H_1: Heteroscedasticidad$$

Suponiendo que se tiene el siguiente modelo:

$$y = \beta_0 + \beta_1 x + u_i$$

Donde se cree que la heteroscedasticidad es ocasionada por la variable explicativa x_i. Los pasos a seguir para contrastar la hipótesis de la existencia del problema en el modelo son los siguientes:

1. Ordenar la muestra en función de x_i, a partir del valor más bajo de x_i.
2. Eliminar c observaciones centrales, de tal manera que las $n - c$ observaciones restantes queden divididas en dos grupos.
3. Al interior de cada observación realizar un MCO y capturar la suma de cuadrados de los residuos en cada una de las submuestras (SCR_1 y SCR_2). SCR_1 representa los valores más bajos de x_i y SCR_2 representa a los valores más grandes de x_i. Cada una de estas tiene $(n - c - 2k)/2$ grados de libertad.

Se capturará el estadístico λ definido como:

$$\lambda = \frac{SCR_2/gl_2}{SCR_1/gl_1}$$

$$\lambda \sim F[(gl_1, gl_2); \alpha]$$

Donde la regla de decisión será la siguiente:

- Si $\lambda > F \rightarrow$ Se rechaza H_0.
- Si $\lambda < F \rightarrow$ Se acepta H_0.

Para realizar el contraste de Goldfeld-Quandt en R se debe hacer uso de la función **gqtest()**, del paquete **lmtest**.

```
gqtest(modelo, order.by = ~ gashog2d, data = sumaria_2015)
```

La sentencia anterior indica que se quiere hacer una prueba de Goldfeld-Quandt para la variable **gashog2d**.

Los resultados mostrarán lo siguiente:

```
> gqtest(modelo, order.by = ~ gashog2d, data = sumaria_2015)

	Goldfeld-Quandt test

data:  modelo
GQ = 15.109, df1 = 16091, df2 = 16091, p-value < 2.2e-16
```

Estos resultados muestran que la hipótesis nula es rechazada y la alternativa que afirma que la heteroscedasticidad es generada por la variable **gashog2d** se afirma.

Solución a la heteroscedasticidad

Mínimos cuadrados generalizados

Suponiendo el modelo:

$$y = \beta_0 + \beta_1 x_1 + u_i$$

Si se conocen las varianzas heteroscedásticas σ^2 para transformar el modelo (ponderarlo) y de esa manera disminuir la dispersión se divide a cada uno de los elementos del modelo entre σ_i , el modelo quedaría así:

$$\frac{y}{\sigma_i} = \beta_0\left(\frac{1}{\sigma_i}\right) + \beta_2\left(\frac{x_1}{\sigma_i}\right) + \frac{u_i}{\sigma_i}$$

La varianza se hace homoscedástica. De esta manera los estimadores de Mínimos Cuadrados Generalizados son los mejores estimadores lineales insesgados.

Bibliografía

Aparicio, J., Martinez, M., & Morales, J. (2004). Modelos Lineales Aplicados en R. Dto. Estadística, Matemáticas e Informática. Universidad Miguel Hernandez. <http://umh3067.edu.umh.es/wp-content/uploads/sites/240/2013/02/Modelos-Lineales-Aplicados-en-R.pdf>

De Arce, R., Mahía, R., & Definición, I. (2001). Conceptos básicos sobre la heterocedasticidad en el modelo básico de regresión lineal tratamiento con EViews. Universidad Autónoma de Madrid. <https://www.uam.es/personal_pdi/economicas/rarce/pdf/heterocedasticidad.pdf>

Kleiber, C. & Zeileis, A. (2008). *Applied econometrics with R*. Springer Science & Business Media.

Quintana, L. & Mendoza, A. (2016). *Econometria aplicada usando R*. Universidad Nacional Autónoma de México.