Test de Goldfeld Quandt y Transformaciones Estabilizadoras de la Varianza

Lin Chiu Chen Yang[1],
Suvieta Moyehuara Carlos[2]

Facultad de Fieecs
Universidad Nacional de Ingenieía

June 12, 2024

Homocedasticidad

Homocedasticidad vs Heterocedasticidad

En estadística, se suele llamar "cedasticidad" a la distribución de los errores. En un modelo lineal de forma $y_i = x_i\beta + \varepsilon_i$ se dice que es homocedástico, si la varianza de los errores ε_i es la misma (o es constante) en todas las muestras. De lo contrario, el modelo será heterocedástico

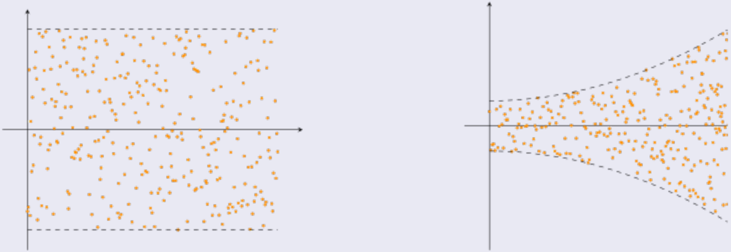


Figure: Homo(left) vs Hetero(right)

Consecuencias de Heterocedasticidad

- El teorema de Gauss-Markov no se cumpliria
- los estimadores MCO no son los mejores estimadores lineales insesgados (BLUE) y su varianza no es la mínima
- El estimador MCO ya no es asintóticamente eficiente
- La heterocedasticidad puede conducir a resultados de regresión inconsistentes e ineficientes.
- Errores estándar incorrectos
- Intervalos de confianza incorrectos
- Pruebas de hipótesis inválidas

Test de contraste para heterocedasticidad

Para contrastar la existencia de heterocedasticidad nuestra hipótesis es:

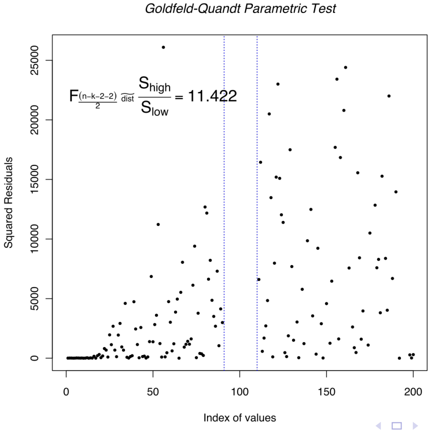
$$H_0 : ausenciadeheterocedasticidad.$$

$$H_a : existenciadeheterocedasticidad.$$

Existe una gran variedad de test de contraste de heterocedasticidad que se diferencian entre sí en su potencia para detectarla. Entre ellas hay el test de Goldfeld y Quandt, el test de Breusch y Pagan o test de White

Test de Goldfeld y Quandt (1965)

Desarrollada por Stephen Goldfeld y Richard E. Quandt publicado en un artuculo 1965. Consiste en dividir los datos en dos subgrupos analizarlos por separado y comparalos si el modelo cumple el supuesto de homocedasticidad entonces deberian ser significativamente similares.



Pasos para realizar la prueba Goldfeld-Quandt

1. Ordenar las observaciones de todas las variables del modelo en la muestra según un ordenamiento de los valores de X_i de menor a mayor.
2. Dividir en dos bloques de tamaño N_1 y N_2 dejando fuera p observaciones centrales para hacer más independientes los dos grupos.
3. Estimar por MCO el modelo de regresión separadamente para cada grupo de observaciones. Guardar la Suma de Cuadrados Residual (SCR) de cada regresión.
4. Construir el estadístico de prueba

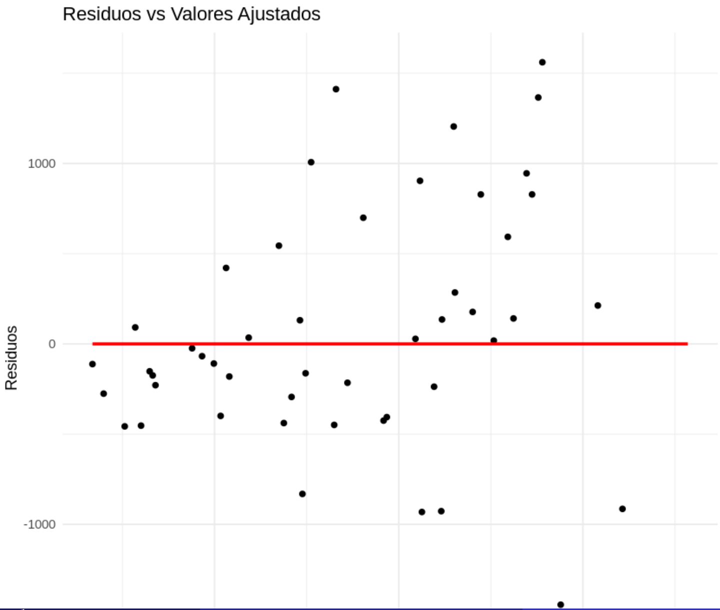
$$GQ = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{\hat{\mu}_2^{'}\hat{\mu}_2}{\hat{\mu}_1^{'}\hat{\mu}_1}\frac{N_1 - K}{N_2 - K} \sim F(N_2 - K, N_1 - K)$$

5. Si existe homocedasticidad las varianzas han de ser iguales; rechazaremos H_0, a un nivel de significancia α si: $GQ > F(N_1 - K, N_2 - K)|\alpha$

Aplicación del test de Goldfeld y Quandt

```
data1 <- read.csv("data2.csv", sep = ";")

model <- lm(revenue ~ costs, data = data1)

data1$residuals <- residuals(model)
data1$fitted <- fitted(model)

ggplot(data1, aes(x = fitted, y = residuals)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red"
  labs(title = "Residuos vs Valores Ajustados",
       x = "Valores Ajustados",
       y = "Residuos") +
  theme_minimal()

gqtest(model, order.by = ~ costs, data = data1, fracti
```

Resultados



Resultados

```
[ ]  gqtest(model, order.by = ~ costs, data = data1, fraction = 10)

        Goldfeld-Quandt test

data:  model
GQ = 8.5916, df1 = 18, df2 = 18, p-value = 1.614e-05
alternative hypothesis: variance increases from segment 1 to 2
```

para un $\alpha = 5\%$ el p-valor es inferior a 0.05 podemos concluir que existe heterocedasticidad

Transformaciones Estabilizadoras de la Varianza

Las transformaciones estabilizadoras de la varianza son técnicas utilizadas en el análisis de regresión para corregir la heterocedasticidad, una condición en la que la varianza de los errores no es constante a lo largo de las observaciones. Estas transformaciones modifican la variable dependiente para estabilizar la varianza de los errores, mejorando así la validez de los modelos de regresión.

Métodos más comunes de transformaciones de estabilización de la varianza:

- **Transformación Logarítmica (Log):**

  $$y' = \log(y)$$

  Utilizada cuando la varianza es proporcional a la media o cuando los datos muestran una distribución sesgada positivamente.

- **Transformación de Raíz Cuadrada (Square Root):**

  $$y' = \sqrt{y}$$

  Útil cuando la varianza es proporcional a la media.

- **Transformación Inversa (Recíproca) (Inverse):**

  $$y' = \frac{1}{y}$$

  Se utiliza cuando los valores grandes tienen mucha varianza.

Métodos más comunes de transformaciones de estabilización de la varianza:

1. **Transformación de Arcoseno (Arcsine):**

$$y' = \arcsin(\sqrt{y})$$

Principalmente usada para proporciones o datos binomiales.

2. **Transformación de Potencia (Box-Cox):**

$$y' = \frac{y^{\lambda} - 1}{\lambda} \quad (\text{para } \lambda \neq 0)$$

$$y' = \log(y) \quad (\text{para } \lambda = 0)$$

La constante λ es determinada para que la varianza sea lo más constante posible.

3. **Transformación de Raíz Cuadrada Inversa:**

$$y' = \frac{1}{\sqrt{y}}$$

Aplicable cuando los valores grandes tienen varianzas significativamente mayores.

Medidas de corrección de la heterocedasticidad-Logaritmos

- La transformación logarítmica es de las mas utilizadas, ya que al tomar logaritmos las series se suavizan y los problemas de heterocedasticidad tienden a desaparecer.
- Esta transformación consiste simplemente en tomar logaritmos en una o varias variables del modelo.
- Es importante tener en cuenta que con este tipo de transformaciones la interpretación de los parámetros del modelo cambia, pasando a ser en términos porcentuales.

Modelo	Variable dependiente	Variable independiente	Interpretación del parámetro
Nivel-nivel	y	x	$\Delta y = \beta_i \Delta x$
Nivel-log	y	Log(x)	$\Delta y = (\beta_i/100)\%\Delta x$
Log-nivel	Log(y)	x	$\Delta\% y = (100\beta_i)\Delta x$
Log-log	Log(y)	Log(x)	$\Delta\% y = \beta_i \%\Delta x$

Medidas de corrección de la heterocedasticidad-Logaritmos

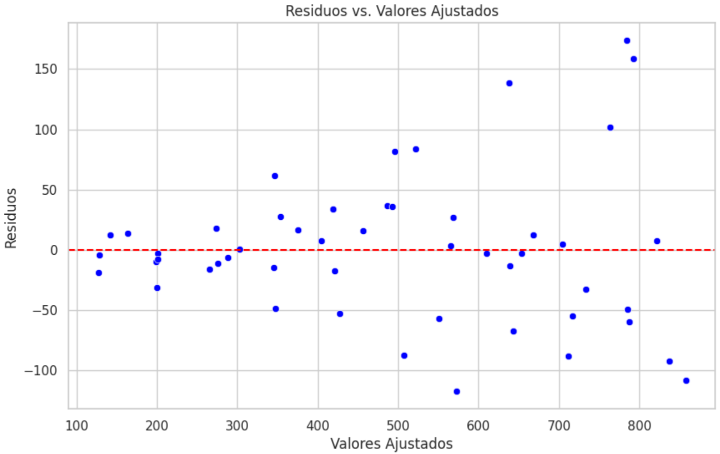
Interpretación de los parámetros en presencia de logaritmos:

- $\Delta y = \beta_i \Delta x$: Cuando X aumenta en una unidad, Y aumenta en β unidades.
- $\Delta y = (\beta_i/100)\%\Delta x$: Cuando X aumenta en un 1%, Y aumenta en $\beta/100$ unidades.
- $\Delta\% y = (100\beta_i)\Delta x$: Cuando X aumenta en una unidad, Y aumenta en $100 \cdot \beta\%$.
- $\Delta\% y = \beta_i\%\Delta x$: Cuando X aumenta en un 1%, Y aumenta en $\beta\%$.

Analizando la Heterocedasticidad del modelo clasico

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              GASTOOCIO   R-squared:                       0.927
Model:                            OLS   Adj. R-squared:                  0.925
Method:                 Least Squares   F-statistic:                     607.4
Date:                Sat, 08 Jun 2024   Prob (F-statistic):           6.73e-29
Time:                        23:48:32   Log-Likelihood:                -276.17
No. Observations:                  50   AIC:                             556.3
Df Residuals:                      48   BIC:                             560.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         53.9422     19.791      2.726      0.009      14.149      93.735
INGRESOS       0.0812      0.003     24.646      0.000       0.075       0.088
==============================================================================
Omnibus:                        7.542   Durbin-Watson:                   2.013
Prob(Omnibus):                  0.023   Jarque-Bera (JB):                6.634
Skew:                           0.731   Prob(JB):                       0.0363
Kurtosis:                       4.023   Cond. No.                     1.36e+04
==============================================================================
```

Analizando la Heterocedasticidad del modelo clasico



Analizando la Heterocedasticidad del modelo clasico

Haciendo la prueba de Goldfeld-Quandt para las hipotesis:

$$H_0 : ausenciadeheterocedasticidad.$$

$$H_a : existenciadeheterocedasticidad.$$

```python
gq_test = het_goldfeldquandt(y, X, split=0.2)
print(f'Estadístico F: {gq_test[0]}')
print(f'P-valor: {gq_test[1]}')
# Definir un nivel de significancia
alpha = 0.05

# Interpretar el resultado de la prueba
heteroscedasticidad = gq_test[1] < alpha

# Imprimir la conclusión del test
print(f'Conclusión del test (True indica heterocedasticidad): {heteroscedasticidad}')
```

```
Estadístico F: 17.393517187649394
P-valor: 1.542842343092295e-10
Conclusión del test (True indica heterocedasticidad): True
```

Se rechaza la hipotesis nula. para un p valor<0.05

Realizando la transformacion logaritmo

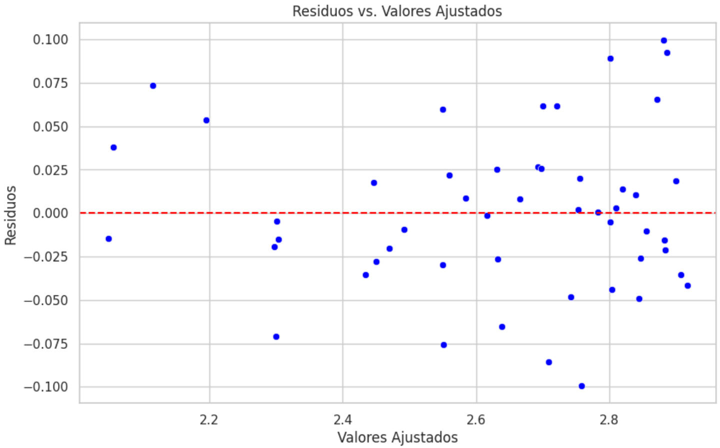
Interpretaríamos el parámetro beta estimado como la variación porcentual que se produce en la variable dependiente ante cambios porcentuales en la variable independiente, es decir: Cuando los ingresos aumentan en un 1% el ocio aumenta en un 0,83%

	GASTOOCIO	INGRESOS	log_gastos	log_ingresos
0	108.050000	900.0	2.033625	2.954243
1	123.980005	918.0	2.093352	2.962843
2	154.069995	1080.0	2.187718	3.033424
3	177.380005	1350.0	2.248905	3.130334
4	189.650000	1791.0	2.277953	3.253096

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             log_gastos   R-squared:                       0.963
Model:                            OLS   Adj. R-squared:                  0.963
Method:                 Least Squares   F-statistic:                     1264.
Date:                Sun, 09 Jun 2024   Prob (F-statistic):           3.88e-36
Time:                        05:59:26   Log-Likelihood:                 83.443
No. Observations:                  50   AIC:                            -162.9
Df Residuals:                      48   BIC:                            -159.1
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.4152      0.086     -4.825      0.000      -0.588      -0.242
log_ingresos     0.8339      0.023     35.548      0.000       0.787       0.881
==============================================================================
Omnibus:                        0.340   Durbin-Watson:                   2.116
Prob(Omnibus):                  0.843   Jarque-Bera (JB):                0.432
Skew:                           0.180   Prob(JB):                        0.806
Kurtosis:                       2.721   Cond. No.                         51.5
==============================================================================
```

Realizando la transformacion logaritmo

Analizando el grafico:



Analizando la Heterocedasticidad del modelo clasico

Haciendo la prueba de Goldfeld-Quandt para las hipotesis:

$$H_0 : \text{ausencia de heterocedasticidad.}$$

$$H_a : \text{existencia de heterocedasticidad.}$$

```
gq_test = het_goldfeldquandt(LOGY, LOGX, split=0.2)
print(f'Estadístico F: {gq_test[0]}')
print(f'P-valor: {gq_test[1]}')
alpha = 0.05

# Interpretar el resultado de la prueba
heteroscedasticidad = gq_test[1] < alpha

# Imprimir la conclusión del test
print(f'Conclusión del test (True indica heterocedasticidad): {heteroscedasticidad}')
```

```
Estadístico F: 1.3242604806629705
P-valor: 0.2613661935881414
Conclusión del test (True indica heterocedasticidad): False
```

Se acepta la hipotesis nula. para un p valor>0.05. Como vemos se ha resuelto el problema de la heterocedasticidad.