Análisis de la Heteroscedasticidad y Autocorrelación

José Ángel Carretero Montes
Ismael Sallami Moreno
Fernando José Gracia Choin

Diciembre 2024

1 Análisis de la Heteroscedasticidad y Autocorrelación

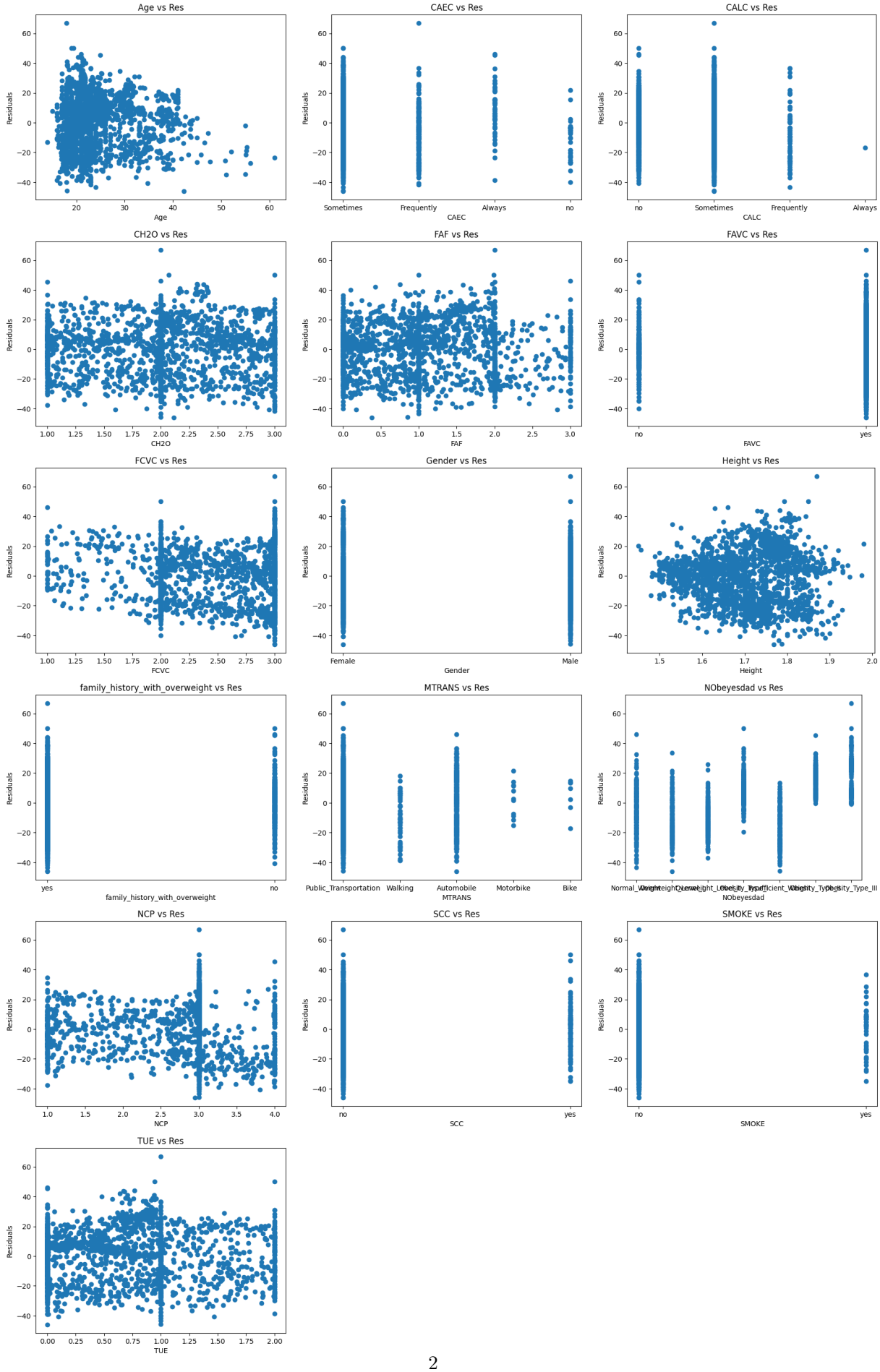
1.1 Análisis de la Heteroscedasticidad

En econometría, un modelo de regresión presenta Heteroscedasticidad cuando la varianza del término de error ε_i no es constante para todas las observaciones. Es decir,

$$\text{Var}[\varepsilon_i] = \sigma_i^2 \quad \forall i.$$

En un modelo homocedástico, $\sigma_i^2 = \sigma^2$, lo cual simplifica los procedimientos de estimación. Sin embargo, cuando hay Heteroscedasticidad, los estimadores de Mínimos Cuadrados Ordinarios (MCO) dejan de ser eficientes, aunque siguen siendo insesgados.

1.1.1 Gráficos



Los gráficos presentados muestran la relación entre las variables predictoras (independientes) utilizadas en el modelo de regresión y los **residuos** obtenidos del ajuste del modelo. Estos gráficos son conocidos como *scatter plots* y permiten evaluar visualmente la validez del modelo, identificando posibles problemas o ajustes necesarios. A continuación, se describe qué representan y cómo interpretarlos.

- **Eje X (Variables independientes):** Cada gráfico representa una de las variables predictoras del modelo, tales como:
  - Características personales: *Gender* (Género), *Age* (Edad), *Height* (Altura).
  - Hábitos: *SMOKE* (Fumar), *FAVC* (Frecuencia de consumo de alimentos calóricos).
  - Otras variables: *MTRANS* (Medio de transporte), *SCC* (Control de consumo de calorías), etc.
- **Eje Y (Residuos):** Los residuos (*residuals*) son las diferencias entre los valores observados y los valores predichos por el modelo de regresión. Representan el error del modelo para cada observación.

El objetivo de estos gráficos es diagnosticar el comportamiento del modelo de regresión y verificar si cumple con ciertos supuestos clave:

1. **Ausencia de patrones en los residuos:**
   - Si los puntos están distribuidos aleatoriamente sin formar patrones claros, esto indica que no hay relación no explicada entre la variable independiente y los residuos.
   - Esto sugiere que el modelo está bien ajustado a los datos.
2. **Presencia de patrones:**
   - Si se observan patrones como formas parabólicas, tendencias lineales o bandas, esto puede indicar:
     - Relaciones no lineales entre las variables independientes y la variable de respuesta.
     - Variables importantes omitidas en el modelo.
     - Problemas de *Heteroscedasticidad*, que es lo que estamos estudiando en este apartado.
3. **Homocedasticidad:**
   - La dispersión uniforme de los puntos en el eje Y sugiere que los errores tienen varianza constante.
   - Si los residuos parecen expandirse o contraerse a medida que cambia el valor de la variable independiente, esto indica *Heteroscedasticidad*.

Por ejemplo:

- En el gráfico de ***Gender vs Residuos***, se evalúa si los residuos muestran algún patrón específico según el género de los individuos.
- En el gráfico de ***Age vs Residuos***, se analiza si la edad tiene alguna influencia no capturada por el modelo.
- En el gráfico de ***Height vs Residuos***, se observa si la altura está relacionada con errores sistemáticos en las predicciones.

Al observar los gráficos, se pueden hacer las siguientes observaciones:
1. La mayoría de las variables muestran residuos distribuidos aleatoriamente, lo que sugiere un buen ajuste y ausencia de Heteroscedasticidad en gran parte del modelo.
2. Sin embargo, en la variable `Age`, los residuos muestran un patrón creciente en su dispersión, lo que indica la presencia de Heteroscedasticidad.
3. Para otras variables, como `Height`, `Gender`, `family_history_with_overweight`, y `FCVC`, los residuos se distribuyen sin patrones claros, lo cual es un indicio positivo.

1.2 Pruebas de Detección de Heteroscedasticidad

Existen diversos métodos analíticos y gráficos para detectar la Heteroscedasticidad. Nos enfocaremos en los siguientes contrastes estadísticos:

1.2.1 Test de Goldfeld-Quandt

El test de Goldfeld-Quandt evalúa la homocedasticidad dividiendo la muestra en dos subgrupos y comparando las sumas de cuadrados residuales (SCR). Los pasos son:

1. Ordenar los datos según el predictor sospechoso de causar Heteroscedasticidad.
2. Omitir los valores centrales para garantizar una mejor separación entre los dos grupos.
3. Estimar el modelo por MCO en ambos subgrupos y calcular las SCR: SCR_1 y SCR_2.
4. Calcular el estadístico F:
$$F = \frac{\text{SCR}_2}{\text{SCR}_1},$$
donde n_1 y n_2 son los tamaños de las muestras en los subgrupos, y k es el número de regresores.

Si $F > F_{n_1-k,n_2-k,1-\alpha}$, se rechaza la hipótesis nula de homocedasticidad. En función del código del colab de este trabajo:

```
GQ = sms.het_goldfeldquandt(y, sm.add_constant(datos["Age"]))
```

El resultado:
$$\text{GQ} = (1.2406, p\text{-valor} = 0.000237).$$
Indica Heteroscedasticidad dado que el p-valor es menor a 0.05.

1.2.2 Test de Breusch-Pagan

Este test evalúa la relación entre los residuos al cuadrado y las variables explicativas del modelo. Los pasos son:

1. Estimar el modelo inicial y calcular los residuos $\hat{e}_i$.
2. Ajustar el modelo auxiliar:
$$\hat{e}_i^2 = \delta_0 + \delta_1 X_{i1} + \cdots + \delta_k X_{ik} + \nu_i.$$
3. Calcular el estadístico LM:
$$LM = nR^2 \sim \chi_k^2.$$
Siendo n el tamaño de la muestra y R^2 el coeficiente de determinación
4. Estadístico LM: es una medida utilizada en econometría para realizar pruebas de hipótesis, en particular, en la detección de Heteroscedasticidad y otros problemas en modelos de regresión. Aunque en las diapositivas de la Asignatura no se menciona explícitamente con este nombre, forma parte implícita de los cálculos de algunos contrastes como el de Breusch-Pagan.

En el código:

```
BP = sms.het_breuschpagan(results.resid, results.model.exog)
```

El resultado:
$$\text{BP} = (410.86, p\text{-valor} = 1.92 \times 10^{-77}).$$
Esto confirma Heteroscedasticidad, dado el p-valor extremadamente bajo.

1.2.3 Test de White

El test de White evalúa Heteroscedasticidad generalizada, incluyendo términos no lineales e interacciones de las variables explicativas en el modelo auxiliar:
$$\hat{e}_i^2 = \delta_0 + \delta_1 X_{i1} + \cdots + \delta_k X_{ik} + \delta_{k+1} X_{i1}^2 + \cdots + \nu_i.$$

En el código:

```
W = sms.het_white(results.resid, results.model.exog)
```

El resultado:
$$\text{White} = (938.445, p\text{-valor} = 1.40 \times 10^{-122}).$$
Nuevamente, el p-valor indica Heteroscedasticidad.

1.2.4 Test de Glejser

Este test examina si los residuos absolutos están relacionados con una transformación de las variables explicativas:

$$|\hat{e}_i| = \delta_0 + \delta_1 z_i^h + \nu_i,$$

donde z_i es una variable explicativa y h toma valores como $\pm 1, \pm 2, \pm 0.5$.

El procedimiento implica estimar para distintos valores de h y contrastar $H_0 : \delta_1 = 0$. Los resultados del ejemplo indican que para $h = \{-2; -1; -0.5; 0.5; 1; 2\}$, hay evidencia significativa de Heteroscedasticidad.

1.3 Interpretación de Resultados

- El test de Goldfeld-Quandt muestra varianzas crecientes en los datos.
- Breusch-Pagan y White confirman una fuerte evidencia de Heteroscedasticidad con p-valores extremadamente bajos.
- Glejser identifica relaciones entre los errores absolutos y transformaciones específicas del predictor.

Estos resultados sugieren que el modelo presenta Heteroscedasticidad y, por lo tanto, los estimadores MCO no son eficientes.

1.4 Corrección de la Heteroscedasticidad mediante Mínimos Cuadrados Ponderados (MCP)

El problema de la Heteroscedasticidad se presenta cuando la varianza de los errores no es constante, lo que viola uno de los supuestos clave de los Mínimos Cuadrados Ordinarios (MCO). Esto afecta la eficiencia de los estimadores, aunque sigan siendo insesgados. Para abordar este problema, se emplean los Mínimos Cuadrados Ponderados (MCP), que ajustan el modelo asignando un peso inversamente proporcional a la varianza de los errores de cada observación.

1.4.1 Transformación del Modelo

El objetivo de MCP es transformar el modelo original:

$$y = X\beta + u, \quad \text{con } \text{Var}(u) = \sigma^2 \Omega,$$

donde Ω es una matriz diagonal que contiene las varianzas de los errores heterocedásticos. Esta transformación se realiza aplicando una matriz de ponderación $P = \text{diag}(\frac{1}{\sqrt{w_i}})$, donde w_i es la ponderación para cada observación. El modelo transformado es:

$$Py = PX\beta + Pu.$$

Bajo este modelo, los errores transformados cumplen:

$$\text{Var}(Pu) = \sigma^2 I,$$

lo que asegura que los errores son homocedásticos, permitiendo aplicar MCO de manera válida.

1.4.2 Aplicación al Modelo Analizado

En el modelo ajustado, los pesos utilizados para las observaciones homogenizan la varianza de los errores. Los resultados del ajuste mediante MCP muestran:

- Un R^2 ajustado de 0.581, indicando un buen poder explicativo del modelo.
- p-valores significativos para la mayoría de las variables independientes, excepto SMOKE y NCP, lo que sugiere que estas variables no contribuyen significativamente al modelo.
- Una correcta especificación de la matriz de varianzas-covarianzas de los errores, como se indica en las notas del modelo.

1.4.3 Ventajas de MCP

Los Mínimos Cuadrados Ponderados solucionan el problema de Heteroscedasticidad al:

1. Garantizar que los errores transformados tienen una varianza constante (homocedasticidad).
2. Proveer estimadores eficientes, es decir, con varianza mínima entre los estimadores lineales insesgados.
3. Validar la inferencia estadística en los coeficientes estimados (intervalos de confianza y pruebas de hipótesis).

1.4.4 Diagnóstico Post-Ajuste

Aunque los MCP corrigen la Heteroscedasticidad, el análisis de los residuos indica ciertos aspectos a considerar:

- **Prueba Omnibus:** Una ligera desviación de normalidad en los errores ($p = 0.000$).
- **Durbin-Watson ($DW = 0.745$):** Posible autocorrelación positiva en los errores, que podría requerir un análisis adicional.

1.4.5 Mínimos Cuadrados Ordinarios

```
WLS Regression Results
=======================================================
Dep. Variable:              Weight       R-squared:              0.584
Model:                         WLS       Adj. R-squared:         0.581
Method:             Least Squares       F-statistic:           183.5
Date:               Tue, 26 Nov 2024    Prob (F-statistic):     0.00
Time:                   07:43:40        Log-Likelihood:        -9001.8
No. Observations:           2111        AIC:                  1.804e+04
Df Residuals:               2094        BIC:                  1.813e+04
Df Model:                     16
Covariance Type:        nonrobust
=======================================================
                     coef    std err   t       P>|t|   [0.025   0.975]
-------------------------------------------------------
const             -216.3333   10.357  -20.888    0.000 -236.644 -196.022
Gender              -5.9958    1.021   -5.872    0.000   -7.998   -3.993
Age                  0.8450    0.086    9.789    0.000    0.676    1.014
Height             128.8866    5.913   21.797    0.000  117.290  140.483
family_history      16.3267    1.111   14.690    0.000   14.147   18.506
FAVC                 7.6898    1.245    6.175    0.000    5.248   10.132
FCVC                 8.8173    0.748   11.795    0.000    7.351   10.283
NCP                  0.9702    0.508    1.912    0.056   -0.025    1.966
CAEC                 7.7443    0.860    9.000    0.000    6.057    9.432
SMOKE               -0.0856    2.764   -0.031    0.975   -5.506    5.335
CH2O                 1.4428    0.650    2.220    0.027    0.168    2.717
SCC                 -6.7508    1.827   -3.695    0.000  -10.333   -3.168
FAF                 -2.6776    0.487   -5.497    0.000   -3.633   -1.722
TUE                 -1.5889    0.651   -2.441    0.015   -2.866   -0.312
CALC                -3.7291    0.774   -4.820    0.000   -5.246   -2.212
MTRANS               3.9056    0.392    9.953    0.000    3.136    4.675
NObeyesdad           2.6400    0.222   11.917    0.000    2.206    3.074
=======================================================
Omnibus:              27.439   Durbin-Watson:              0.745
Prob(Omnibus):         0.000   Jarque-Bera (JB):          16.804
Skew:                 -0.003   Prob(JB):                0.000224
Kurtosis:              2.563   Cond. No.                  789.000
=======================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

1.5 Análisis de la Autocorrelación en Datos Transversales

En el contexto de análisis de datos transversales, no tiene sentido realizar un análisis de autocorrelación entre variables de manera similar a como se haría en series temporales. Esto se debe a varias razones fundamentales:

1. **Ausencia de Dependencia Temporal:** En un conjunto de datos transversales, las observaciones corresponden a unidades diferentes (como individuos, empresas o países) en un único punto en el tiempo. A diferencia de las series temporales, en las que se estudian las observaciones de una misma unidad a lo largo del tiempo, en los datos transversales no existe una estructura temporal que justifique la existencia de autocorrelaciones entre las observaciones en diferentes períodos.

2. **Independencia de las Unidades de Análisis:** Las observaciones en un conjunto de datos transversal suelen ser independientes entre sí. Esto significa que las relaciones entre las variables en diferentes unidades no necesariamente tienen un patrón temporal o causal. En otras palabras, no se puede asumir que el comportamiento de una unidad en un período determinado influencie o dependa del comportamiento de otra en el mismo período.

3. **Limitación en la Interpretación de Autocorrelaciones:** En series temporales, la autocorrelación puede ayudar a identificar patrones en la evolución de una misma variable a lo largo del tiempo. Sin embargo, en datos transversales, las autocorrelaciones reflejan solo asociaciones entre las variables en un punto en el tiempo sin considerar el cambio o la dinámica temporal. Por lo tanto, la interpretación de la autocorrelación en datos transversales puede ser más superficial y no reflejar relaciones causales.

4. **Estructura de los Datos:** La estructura de los datos transversales no permite que las autocorrelaciones entre variables reflejen una interacción temporal o dinámica. Cualquier análisis de autocorrelación realizado en datos transversales debe ser interpretado con cautela, ya que las relaciones entre variables pueden ser simplemente una coincidencia en un punto específico del tiempo, sin que se pueda generalizar a lo largo del tiempo.