Estadística II - Modelos lineales:

Ficha nº6 Heterocedasticidad en el modelo de regresión lineal

---

Material didáctico para el curso de Estadística II del Programa de Doctorado en Ciencia Social

Tabaré Fernández Aguerre

junio 2004

IX. Heterocedasticidad

✎ Por definición, el supuesto hecho al ajustar por mínimos cuadrados ordinarios un modelo de regresión es que el término de error tiene varianza constante. Formalmente:

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

✎ Dado que la notación ocupa aquí un papel esencial, conviene mostrar que la varianza representada por la letra griega sigma no tiene subíndice, mostrando así que la varianza es homocedástica.

✎ Para examinar el supuesto, se propone el siguiente método:

- Naturaleza y origen de la heterocedasticidad
- Consecuencias de la violación del supuesto
- Formas de detectar heterocedasticidad
- Medidas remediales

1. Naturaleza y origen de la heterocedasticidad

✎ En consecuencia, la presencia de varianza heterocedástica se define de la siguiente forma (obsérvese el subíndice debajo de sigma cuadrada):

$$Var(e_i) = \sigma^2{}_i$$

✎ Si recordamos que el término de error del modelo es, la diferencia entre el valor observado para la i-ésima unidad y el valor estimado para la i-ésima unidad mediante el modelo de regresión, se puede introducir la varianza condicional.

✎ En síntesis, la heterocedasticidad es una característica del modelo por la que las varianzas condicionales del error no son constantes.

- Obsérvese de forma muy atenta que la idea de varianza condicional refiere al modelo completo: el problema aparece cuando la combinación lineal de todos los regresores del modelo generan errores cuyas varianzas condicionales no son constantes.

- Esto es importante porque luego se observarán situaciones en las que por alguna razón teórica y o empírica, se ha supuesto que una variable es la generadora de la heterocedasticidad. Sin embargo, esta situación es una simplificación a veces didáctica o en todo caso, la situación más favorable para luego emprender medidas correctivas.

✎ En los textos de estadística y de econometría se registran algunas de las causas por las que puede aparecer este problema. Haciendo una síntesis de Gujarati (2004), Chaterjee et. Al (2000), de Arce (2001), se proponen las siguientes razones:

- La **heterocedasticidad aparece porque ha omitido una variable relevante** para explicar el comportamiento de Y. Con frecuencia, entonces, lo que parece ser un problema de heterocedasticidad puede deberse al hecho de que alguna de las variables importantes han sido omitidas del modelo. Como se verá, este problema de especificación conduce a dos problemas: los estimadores pueden ser sesgados (por variables omitidas) y las varianzas no son eficientes. De aquí se derivan dos claras directivas:

  - ☞ Realizar el análisis de heterocedasticidad cuando el modelo está completo, con todas las variables relevantes incluidas.

  - ☞ Si en el proceso de ajustar de regresión ha dejado afuera del modelo alguna variable relevante, **inclúyala** antes de continuar con el examen

de este supuesto.

- En ocasiones, la **naturaleza teórica del problema indica heterocedasticidad**.

  - Cuando se está interesado en regresar el ahorro sobre el nivel de ingresos de un hogar, la teoría económica indica que el incremento del ingreso puede entenderse también con un incremento en las alternativas discresionales de utilización de este ingreso marginal. Por tanto, también se incrementará la varianza en el nivel de ahorro observado para estos hogares con altos ingresos respecto de los hogares con más bajos ingresos.

  - El campo del análisis organizacional proporciona otro ejemplo paradigmático. Sabido es que el número de supervisores o de personas que desempeñan roles intermedios en la estructura organizacional es una función del tamaño de la organización. Las pequeñas empresas pueden disponer de uno o dos supervisores o incluso no contar con ellos. Sin embargo, a medida en que se incrementa el tamaño de la organización, el proceso de burocratización en la organización conlleva a incrementar el número de supervisores. Pero las organizaciones grandes pueden optar por diversos parámetros de diseño en materia de supervisión, observándose en consecuencia una mayor varianza en el número de supervisores.

- Todo análisis que se basa en **modelos de aprendizajes de errores** (por ejemplo en el error que se comete al responder un examen) se puede sostener que los errores de respuesta con menores. Este aspecto afecta principalmente a la psicometría y a toda aplicación en el campo de las evaluaciones realizadas durante varios momentos en el tiempo.

- La relación entre una variable X y la variable dependiente Y se ha especificado con una **forma funcional incorrecta**. Este es el caso cuando se supone que la relación es lineal y en realidad sigue la forma de una parábola y en consecuencia habría de ajustarse añadiendo un término cuadrádico para X.

- Otra fuente de heterocedasticidad proviene de la asimetría en la distribución de una o más variables regresoras incluidas en el modelo de regresión. Los ejemplos más frecuentes son el ingreso, variables que miden el bienestar, los años de escolaridad formal, la edad, etc. En tal caso se

sugiere utilizar una transformación de las variables para corregir dichas asimetrías.

- Sin embargo, la **heterocedasticidad aparece también con una incorrecta transformación de los datos**. Por ejemplo, se ha tomado la decisión de transformar la variable dependiente siguiendo alguna función (v.g. un logaritmo) cuando no era apropiada; o se ha decidido transformar una o algunas variables independientes cuando esto no era necesario teórica o empíricamente (v.g. tomando primeras diferencias o rezagando).

- La heterogeneidad puede ser el resultado de **observaciones "atípicas"**, "fuera de regla", "influyentes" o también denominadas "puntos aberrantes" en la distribución conjunta para el modelo.

  - ☞ Si no se ha hecho un análisis de los "outlayers" y de los puntos influyentes, conviene realizarlo antes de proseguir con el análisis de heterocedasticidad.

- La heterocedasticidad aparece porque en el análisis se ha desconocido que existen al menos dos niveles de análisis. Prácticamente esto sucede toda vez que se ha procedido a agregar información de un nivel de análisis inferior a otro superior. En la bibliografía este problema es tratado dentro de la noción más amplia de sesgos de agregación y recibe particular atención en el análisis organizacional.

  - ☞ Una variable registrada a nivel de los individuos, como por ejemplo, el nivel de aprendizajes de un alumno, los días que un paciente ha estado internado o el salario recibido por un empleado, han sido promediados a nivel de una organización (escuela, hospital o empresa).

  - ☞ La recaudación anual que hace una escuela por concepto de contribuciones voluntarias de los padres (lo que es conocido en Argentina como "cooperadora" y en Uruguay como "comisión de fomento") se ha promediado entre los distritos escolares para realizar un análisis de la asignación estatal de los recursos entre distritos.

  - ☞ La inversión anual en innovación y desarrollo que realiza una empresa ha sido promediada a nivel del país, para hacer un análisis comparativo entre países.

- La **variable dependiente es una proporción o un porcentaje**. En tal caso, por definición la varianza de una proporción está interconstruida con su promedio, por lo que varía conforme a ésta.

$$Var(Y)=P*(1-P)$$

- A medida que **mejoran las técnicas de recolección** de información, es probable que la varianza de los errores se reduzcan.

✎ En síntesis, la heterocedasticidad puede surgir del tipo de problema analizado; de la forma en como se especificó el modelo; de la forma en cómo se recolectó la información y de las decisiones tomadas en el tratamiento de los datos. Principalmente se presentará en estudios transversales, aunque también está presente en estudios que involucran observaciones en el tiempo.

2. Consecuencias de la heterocedasticidad

✎ En presencia de heterocedasticidad, las estimaciones hechas mediante mínimos cuadrados (MCO) siguen siendo insesgadas pero ya no son de varianza mínima.

- Por definición, para computar el error estándar de cada coeficiente de regresión para los regresores (β_1, β_2, ... β_p,), requiere en su numerador la varianza homocedástica de los residuos (σ^2).

$$ee_{b_p}=\sqrt{\frac{\hat{\sigma}^2}{S_{Xp}{}^2*(1-R_{Xp}{}^2)}}$$

- En consecuencia, si la varianza condicional no es constante, la utilización de la fórmula MCO para estimar los errores de los coeficientes puede ser que sub-estime o sobre-estime la verdadera varianza. Estrictamente no se conoce cuál es la verdadera varianza.

- Por lo general, los textos de estadística suelen afirmar que la varianza computada por MCO es menor que la verdadera, por lo que al utilizar aquellas se está dando una falsa imagen de precisión en la estimación.

✎ Ahora bien, si no es posible estimar con certeza los errores estándares para cada uno de los coeficientes de regresión, esto supone que tampoco podrá computarse el valor del estadístico t .

- Es claro que si la fórmula del estadístico requiere en su denominador el valor del error estándar del coeficiente, y si este no es posible de calcularlo, entonces **los valores de *t* que entrega el SPSS ® no pueden utilizarse**.

✎ Finalmente, no es posible tomar decisiones respecto de las hipótesis nulas formuladas para los coeficientes de regresión porque no se cuentan con estadísticos t.

3. Detección de heterocedasticidad

✎ La detección de heterocedasticidad supone que ya se ha ajustado el modelo de regresión lineal, por ejemplo:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_p X_{pi} + e_i$$
$$Var(e_i) = S^2{}_i$$

✎ Para probar si un modelo presenta heterocedasticidad, se disponen de dos grandes tipos de pruebas: las gráficas, también denominadas no-formales,

y las pruebas formales.

- Para hacer una prueba gráfica, por ejemplo, se coloca en las abscisas uno a uno los regresores X_p incluidos en el modelo "sospechoso". En las ordenadas se puede colocar alternativamente: la variable Y y los residuos estandarizados. Este tipo de gráficos permitirá observar si los errores varían con alguna o con varios regresores.

- Las pruebas gráficas entran indicios sobre la existencia del problema en el caso más simple de todos: cuando la heterocedasticidad está originada en una variable.

- Sin embargo, el supuesto refiere a la varianza condicional para el modelo complejo ajustado. En consecuencia, si la heterocedasticidad se origina en la combinación lineal de todas o de algunas de las variables incluidas en el modelo, las pruebas gráficas serán insuficientes; no podrán detectarla. Por esta razón se recomienda realizar pruebas formales para su detección.

✎ La **prueba de Goldfeld-Quant** resulta una de las más directas de realizar. Opera comparando la magnitud de la suma cuadrada de los errores de regresiones (SCE) ajustadas con subconjuntos de la muestra original de los datos. La hipótesis nula es que existe homocedasticidad: si es correcta entonces el conjunto de modelos ajustados tendrán los mismos residuos cuadrados.

- La lógica es que la heterocedasticidad introducida por un regresor genera una varianza condicional que sigue la siguiente función, donde el investigador propone la hipótesis de que la varianza condicional es proporcional al cuadrado de una variable X:

$$\sigma^2{}_i = \sigma^2 * X^2{}_{pi}$$

- Tiene dos supuestos: i) se supone que existe una variable métrica (interval o de razón) X_p incluida en el modelo que genera varianzas heterocedásticas; ii) se supone que los residuos del modelo se distribuyen normalmente.

- El **primer paso** de la prueba consiste en ordenar la matriz de datos de acuerdo a los valores de X, de menor a mayor por ejemplo.

- El **segundo paso** implica determinar un número de observaciones centrales de la matriz de datos, en adelante "C", que será "filtrado" de los análisis subsiguientes. Quedarán dos sub-poblaciones de datos, denominadas "n_1" y "n_2". Idealmente, conviene que: $n_1 = n_2$. Por lo que se deberá manipular C para cumplir con este requisito.

- El **tercer paso** consiste en ajustar para cada sub-población una regresión MCO con el mismo p número de parámetros que los utilizados para la regresión original. Se obtendrá para cada una la SCE y los grados de libertad, GdL, siendo estos últimos:

$$GdL_1 = n_1 - p$$

- En el **cuarto paso** se calculará la razón siguiente:

$$\lambda = \frac{SCE_2 \,/\, GdL_2}{SCE_1 \,/\, GdL_1}$$

- "Si se ha supuesto que los errores, e_i están distribuidos normalmente (lo cual usalmente se hace) y si el supuesto de homocedasticidad es válido, entonces puede mostrarse que λ sigue la distribución F con un número de grados de libertad en el numerador y en el denominador iguales a (N - C - 2 p)/2" . (Gujarati, 2004: 393).

- Si en una prueba el valor observado de F es superior al valor crítico de F, entonces se puede tomar la decisión de rechazar la hipótesis nula en virtud de que hay indicios de heterocedasticidad.

✎ La prueba de **Breuch-Pagan-Godfrey** es más flexible que la anterior en la medida en que ya no supone en forma restringida, que la

heterocedasticidad proviene de una única variable. Supone gruesamente que si la hipótesis nula de homocedasticidad es correcta, los cuadrados de los residuos no serán explicados por una función lineal de las variables independientes del modelo.

- Dado el siguiente modelo:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_p X_{pi} + e_i$$

- Se supone, como hipótesis alternativa, que la varianza heterocedástica del modelo, σ^2_i, es una función de un sub-conjunto de "m" regresores de los p regresores totales incluidos en el modelo original:

$$\sigma_i^2 = a_0 + a_1 Z_{1i} + a_2 Z_{2i} + \ldots + a_m Z_{mi}$$

- Como **primer paso** de la prueba, se estima el modelo de regresión original y se obtienen los residuos estimados e_i.

- En un **segundo paso** se obtiene el estimado la varianza tal que $\sigma^2 = \sum e_i / n$.

- En un **tercer paso**, construyáse la variable dependiente para el modelo auxiliar de la prueba, tal que:

$$r_i = e_i / \sigma^2$$

- El **cuarto paso** consiste en el ajuste del modelo auxiliar donde r_i será la variable dependiente y como independientes se seleccionará un sub-conjunto de "m" variables de las "p" variables del modelo original. El término residual de esta nueva ecuación será "v".

$$r_i = a_0 + a_1 Z_{1i} + a_2 Z_{2i} + \ldots + a_m Z_{mi} + v_i$$

- En el **quinto paso**, se retiene de esta regresión auxiliar la suma de los cuadrados del modelo de regresión (SCM) y se define:

$$\theta = \frac{1}{2} SCM_r$$

- Suponiéndose que los errores del modelo original, e_i, se distribuyen normalmente, se puede mostrar que si hay homocedasticidad y si el tamaño "n" de la muestra aumenta indefinidamente, entonces:

$$\Theta \sim \chi^2_{m-1}$$

- Es decir Θ sigue una distribución de χ^2 con (m-1) grados de libertad. Por consiguiente, si en una prueba el valor de Θ observado excede al valor crítico de χ^2 al nivel de significación seleccionado, se puede rechazar la hipótesis de homocedasticidad.

✎ La **prueba de White** constituye una forma general de identificar un posible problema de heterocedasticidad, sin que se hagan supuestos sobre la incidencia de una variable en particular o sobre la distribución de los residuos.

- En primer lugar, supóngase que se ajusta el siguiente modelo con tres variables regresoras relevantes (es decir, se supone correctamente especificado):

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + e_i$$

- En segundo lugar, obténgase para este modelo los residuos y se los eleva al cuadrado para evitar los signos.

- En **tercer lugar** ajústese el siguiente modelo de regresión auxiliar para la prueba de White:

$$e_i^2 = a_0 + a_1 X_{1i} + a_2 X_{2i} + a_3 X_{3i} +$$
$$a_4 X_{1i}^2 + a_5 X_{2i}^2 + a_6 X_{3i}^2 +$$
$$a_7 X_{1i} * X_{2i} + a_8 X_{1i} * X_{3i} + a_9 X_2 * X_3 +$$
$$\nu_i$$

- Este modelo auxiliar contiene en la primera fila los mismos tres regresores incluidos en el modelo original. En la segunda fila, cada uno de los tres regresores ha sido incluido pero elevándolo al cuadrado. En la tercera fila se han incluido las interacciones tomadas de a dos entre los regresores. Finalmente, en la cuarta fila se encuentra un término residual (designado con la letra griega nu).

- En **cuarto lugar**, regístrese el número de casos con que se realiza este análisis y el coeficiente de determinación obtenido.

- En quinto lugar obténgase el estadístico siguiente:

[1] $$n^* R^2 \sim_{asi} \chi^2_{gdl}$$

- Bajo la hipótesis nula de que el modelo es homocedástico, puede demostrarse que el tamaño de la muestra multiplicado por el coeficiente de determinación del modelo auxiliar sigue asintóticamente (asi) la distribución ji-cuadrada con grados de libertad (gdl) igual a p-número de regresores (parámetros menos la constante) del modelo auxiliar. En el caso del ejemplo, los grados de libertad son 9 porque se han incluido 9 regresores en el modelo (las tres variables originales, sus cuadrados y sus interacciones).

- En sexto lugar, compárese el estadístico obtenido para el modelo auxiliar con el valor crítico de ji-cuadrada definida según el nivel de significación

deseado y los grados de libertad. Si el valor observado supera al valor crítico, se puede tomar la decisión de rechazar la hipótesis nula y sostener que existen inicidios de heterocedasticidad en el modelo original.

- Ahora bien, conviene transcribir una nota de precaución hecha por Gujarati :

  > "En los casos en que el estadístico de White es significativo estadísticamente, la heterocedasticidad puede no ser necesariamente la causa, sino los errores de especificación. En otras palabras, la prueba de White puede ser una prueba de heterocedasticidad (pura) o de error de especificación o de ambos. Se ha argumentado que si no estuviesen los productos cruzados entre las variables en el procedimiento de White, entonces constituye una prueba de heterocedasticidad pura." (Gujarati 2004, 399-400).

- En síntesis, la prueba de White no es concluyente respecto de la existencia de heterocedasticidad. Para establecer tal conclusión, el modelo requiere estar correctamente especificado. Si no lo está, la significación puede deberse a la ausencia de términos cuadrados o de interacciones no consideradas en el modelo original.

4. Correcciones para modelos heterocedásticos

La heterocedascitidad puede ser un problema inoportuno a resolver automáticamente en el camino del ajuste final de un modelo o una oportunidad de realizar un hallazgo sustantivo. Por lo general, en los análisis predominan más los primeros que los segundos.

Conviene comenzar con algunos criterios generales que se deben tener en cuenta.

- Las medidas remediales disponibles para lidiar con el problema de la heterocedasticidad dependen de la causa que la ha generado. Conviene que antes de tomar decisiones, se haya identificado la o las causas de heterocedasticidad presentes.

- La etapa de detección a través de gráficos o a través de pruebas puede proporcionar un indicio firme para luego formular una medida correctiva. Tal es el caso cuando se ha identificado que la varianza heterocedástica es una función de **sólo uno** de los regresores.

- Algunas de las formas más corrientes de resolver heterocedasticidad no pueden aplicarse utizando el SPSS ®, sino que es necesario utilizar otros paquetes estadísticos como el E-VIEWS ® o el STATA ® .

- Si la heterocesdacticidad ha sido provocada por la vía de la agregación, la forma más razonable de resolverla es asumir la existencia de dos niveles de análisis (v.g. individuos y organización) y modelizar esta relación mediante el uso de modelos estadísticos multinivel del tipo HLM ® o MLWIN ®.

- Dado que la heterocedasticidad puede ser el resultado de un sesgo de especificación, el uso de la prueba White podría conducir a introducir términos de interacción o potencias en la ecuación, antes de continuar con alguna medida remedial.

✎ Frente a la heterocedasticidad, cuatro son las medidas que se puede adoptar:

- No hacer nada e informar las estimaciones tal como son generados por MCO.

- Aceptar las estimaciones hechas por MCO, pero corregir una por una las varianzas de los coeficientes de regresión, bajo determinados supuestos muy específicos.

- Corregir la heterocedasticidad mediante Mínimos Cuadrados Ponderados (MCP). Chaterjee et al (2000: 181-199) y Gujarati (2004: 400-407) presentan varias medidas ad-hoc y estrategias distintas, dependiendo cada de cuál sea la función que explica la varianza heterocedástica.

- Por lo general, cuando se ha decidido resolver el problema se vuelve a ajustar el modelo solicitando “errores estándares robustos” o “errores

estándares de White" entre las opciones que proveen los paquetes econométricos como el E-VIEWS ® o el STATA ® . Esto da lugar a un ajuste por Mínimos Cuadrados Generalizados (MCG, o GLS por sus siglas en inglés).

SUPUESTO	PRUEBA	VIOLACIÓN	CONSECUENCIAS	TRATAMIENTOS
Homo-cedasticidad Var ($\varepsilon_{i/X}$) = σ^2	1) Naturaleza teórica del problema 2)Diagrama de dispersión simple de Y contra cada X. 3) Diagrama de dispersión de error estandarizado frente al Y estimado 4) Prueba de Goldfeld-Quant (idem Chow). 5) Breusch-Pagan-Godfrey (BPG) 6) Prueba de White	Var ($\varepsilon_{i/X}$) = $\sigma^2{}_j$ 1) Es teóricamente esperable una relación decreciente de los errores con el tiempo en base al aprendizaje 2) Se han usado fuentes secundarias cuya producción está explicada por una variable que también explica la variabilidad de Y 3) Se excluyó una variable relevante. 4) con el paso del tiempo se mejoran las mediciones de Y. 5) Presencia de casos totalmente desviados La varianza de ε depende de los valores de una X_i incluida en el modelo.	1) los coeficientes son insesgados pero no son eficientes, **si todos los restantes supuestos se cumplen.** 2) Los estimadores son ineficientes (no tienen varianza mínima). 3) La varianza del coeficiente b_k en presencia de heterocedasticidad, no se sabe si es mayor o menor que la verdadera varianza. 3) Es probable que las pruebas F y t den resultados imprecisos.	Dependen del paquete estadístico que se está empleando. 1) Si todos los supuestos se comprueban y se conoce σ^2 podría corregirse la fórmula de la variranza del coeficiente por: $\text{var}(\hat{\beta}_k) = \frac{\sum \left[(X_i - \overline{X})^2 \sigma^2 \right]}{\left(\sum \left[(X_i - \overline{X})^2 \right] \right)^2}$ 2) En el SPSS puede ajustarse un modelo en base a Mínimos Cuadrados Ponderados (MCP o WLS en inglés) 3) En otros paquetes, tal como el E-Views y el Stata, correr el modelo con correción de White errores estándares robustos) .