REAL INSTITUTO Y OBSERVATORIO DE LA ARMADA EN SAN FERNANDO
BOLETÍN ROA N.º 4/2024

Trabajo de Fin de Máster en Técnicas de Ayuda a la Decisión

Técnicas avanzadas para el estudio de modelos de regrasión. Método de Mínimos Cuadrados Ponderados

MINISTERIO DE DEFENSA

REAL INSTITUTO Y OBSERVATORIO DE LA ARMADA EN SAN FERNANDO

BOLETIN ROA Nº 4/2024

Trabajo de Fin de Master en Técnicas de Ayuda a la Decisión

Técnicas avanzadas para el estudio de modelos de regresión. Método de Mínimos Cuadrados Ponderados.

MINISTERIO DE DEFENSA

Catálogo de Publicaciones de Defensa
https://publicaciones.defensa.gob.es

Catálogo de Publicaciones de la Administración General del Estado
https://cpage.mpr.gob.es

publicaciones.defensa.gob.es
cpage.mpr.gob.es

Edita:

MINISTERIO DE DEFENSA
SUBSECRETARÍA DE DEFENSA
SECRETARÍA GENERAL TÉCNICA

Paseo de la Castellana 109, 28046 Madrid

NIPO 083-24-208-1 (edición impresa)
ISBN 978-84-9091-932-3 (edición impresa)
Depósito legal M 14838-2024
Boletín ROA, ISSN 1131-5040 (edición impresa)

NIPO 083-24-207-6 (edición en línea)

Fecha de edición: junio de 2024
Maqueta e imprime: Imprenta Ministerio de Defensa

Foto portada:

Fachada del Edificio Principal del Real Instituto y Observatorio de la Armada en San Fernando (siglo XVIII).

Técnicas avanzadas para el estudio de modelos de regresión. Método de Mínimos Cuadrados Ponderados.

Trabajo Final de Máster

Autor/a: D. Juan Manuel González Sánchez

Director/a: Dr. Tomás Baenas Tormo

Tutor/a: D. José Manuel Pertusa Arronis

Máster Universitario en Técnicas de Ayuda a la Decisión

Especialidad: Estadística

Curso: 2022/2023 – convocatoria: julio

TÉCNICAS PARA EL ESTUDIO DE MODELOS DE REGRESIÓN. MÉTODO DE MÍNIMOS CUADRADOS PONDERADOS.

RESUMEN:

El método de Mínimos Cuadrados Ponderados es un procedimiento estadístico que permite adaptar el método de Mínimos Cuadrados Ordinarios en condiciones de heterocedasticidad de los datos. A partir de un modelo de regresión lineal simple se transformarán las variables para obtener un modelo de regresión homocedástico donde se minimizará la varianza residual con la introducción de pesos que permitirán que los estimadores resulten insesgados y eficientes. Para detectar la heterocedasticidad se implementarán diferentes tests de hipótesis específicos para el análisis de la homocedasticidad. El software empleado en el presente trabajo será: Excel, RStudio, SPSS y GRETL.

ABSTRACT:

The Weighted Least squares method is a statistical procedure that allows adapting the Ordinary Least Squares method under conditions of heteroscedasticity of the data. Starting from a simple linear regression model, variables will be transformed to obtain a homoscedastic regression model where the residual variance will be minimized with the introduction of weights that will enable the estimators to be unbiased and efficient. To detect heteroscedasticity, several hypothesis tests for homoscedasticity analysis will be performed. The software used in this work will be: Excel, RStudio, SPSS and GRETL.

A Eloísa, mis hijos y mis padres, por todo el tiempo que no les he podido dedicar, así como por el cariño, aliento y sobre todo, paciencia, que me han podido brindar durante este año y medio

Agradecimientos

A mis compañeros de la Sección de Hora, en especial, al Capitán de Fragata Héctor Esteban Pinillos sin cuya amistad, apoyo, mediación y consideración no se podría haber materializado este proyecto. A mi compañero de Máster, Teniente (EA) Manel Vargas Bono, sincero colega y amigo, que gracias a su tenacidad e inconformismo considero han sido cruciales para afrontar el curso de máster. Al Director del Real Instituto y Observatorio de la Armada, CN. Antonio Ángel Pazos García, por su consideración y confianza depositada al permitir que pudiera solicitar el curso de Diploma Militar en Estadística. Al Jefe de la Sección de Técnicas de Apoyo a la Decisión (SETAD) del EMA, CF Eduardo Gómez Quijano, al tutor de prácticas, CF José Manuel Pertusa Arronis, y resto de personal que componen la SETAD, quienes han contribuido con su acogida y asesoramiento a la confección del presente trabajo. A mi compañero de prácticas en la SETAD, Teniente (GC) Luis Enrique Arjona Sánchez, por su leal compañía, voluntad de vencer, "catarsis" permanente y perspectiva práctica del curso. A mi Director de TFM, Dr. Tomás Baenas Tormo, sin cuya experiencia, dedicación, apoyo y paciencia no hubiera sido posible este trabajo.

Contenido

Tablas vi

Capítulo 1. Introducción 1

1.1. Objetivos 2

1.2. Estructura 2

1.3. Metodología 3

Capítulo 2. La regresión lineal simple 5

2.1. El método de Mínimos Cuadrados Ordinarios (MCO) 5

2.1.1. El modelo de regresión lineal simple 6

2.1.2. El coeficiente de correlación lineal 10

2.2. Inferencias estadísticas sobre un parámetro poblacional 12

2.2.1. Contraste de la regresión lineal simple 14

2.2.2. Contraste para el coeficiente de correlación lineal 15

2.2.3. Contraste para el coeficiente de regresión lineal 16

2.3. Heterocedasticidad 18

2.3.1. Causas de la heterocedasticidad 21

2.3.2. Detección gráfica de la heterocedasticidad 23

2.3.2.1 Gráficos de residuos 23

2.3.2.2 Gráficos de dispersión 25

2.4. Contrastes de heterocedasticidad 27

2.4.1. El contraste de *Goldfeld* y *Quandt* (1965) 27

2.4.2. El contraste de *Breusch-Pagan-Godfrey* (1979) 29

2.4.3. El contraste de *Koenker-Basset* (1979) 31

2.4.4. El contraste de *White* (1980) 32

Capítulo 3. El método de Mínimos Cuadrados Ponderados 35

3.1. Introducción 35

3.1.1. Propiedades de los estimadores de MCO bajo heterocedasticidad 35
3.2. El método de mínimos cuadrados ponderados 37
3.2.1. Ejemplo comparativo $\sigma i2$ conocida y desconocida con MCP 41
3.2.2. El método de MCP con el modelo de regresión logística simple 44
Capítulo 4. Ejercicio práctico de aplicación 49
4.1. Motivación del estudio 50
4.2. Características técnicas de la aeronave 51
4.3. Validación del modelo de regresión. Aplicación de contrastes de heterocedasticidad 53
4.3.1. Goldfeld y Quandt 53
4.3.2. Breusch-Pagan-Godfrey 54
4.3.3. Koenker-Bassett 55
4.3.4. White 56
4.4. Análisis de los datos 56
4.4.1. Análisis de los datos mediante el método MCO 57
4.4.2. Análisis de los datos mediante el método MCP con $\sigma i2$ desconocida 63
4.4.3. Comparativa estimadores MCO frente a estimadores MCP 71
Capítulo 5. Conclusiones y líneas futuras 73
Apéndice A. Datos de horas de vuelo y consumo de combustible (2012-2022) 77
Apéndice B. Cálculo de Contrastes de Heterocedasticidad 81
B.1. Resultados test Goldfeld y Quandt 82
B.2. Resultados test Goldfeld y Quandt 83
B.3. Resultados test Breusch-Pagan-Godfrey 84
B.4. Resultados test Koenker-Basset 85
B.5. Resultados test de White 86

Referencias.....88

Ilustraciones

ILUSTRACIÓN 1. ERROR ALEATORIO Y ERROR RESIDUAL DE UN VALOR $\boldsymbol{xi}$ EN UNA ECUACIÓN DE REGRESIÓN. FUENTE: GUJARATI (2004) 8

ILUSTRACIÓN 2. PATRONES DE CORRELACIÓN EN FUNCIÓN DEL VALOR DE $\boldsymbol{r}$.FUENTE: (WWW.STATLECT.COM, S. F.) 11

ILUSTRACIÓN 3. DISTRIBUCIÓN NORMAL HOMOCEDÁSTICA CON LA VARIABLE EXPLICATIVA $\boldsymbol{x}$. FUENTE: WOOLDRIDGE (2010) 13

ILUSTRACIÓN 4. DISTRIBUCIÓN NORMAL HOMOCEDÁSTICA VS HETEROCEDÁSTICA. FUENTE: GUJARATI (2004) 20

ILUSTRACIÓN 5. DISMINUCIÓN ERRORES CONFORME SE INCREMENTA EL TIEMPO DE PRÁCTICAS. FUENTE: GUJARATI (2004) 21

ILUSTRACIÓN 6. BOX-PLOT CON PRESENCIA DE VALORES ATÍPICOS. FUENTE: ELABORACIÓN PROPIA CON SPSS 22

ILUSTRACIÓN 7. TASA DE EMPLEO FRENTE AL PRODUCTO INTERIOR BRUTO (PIB) FUENTE: ELABORACIÓN PROPIA CON SPSS 23

ILUSTRACIÓN 8. GRÁFICO DE RESIDUOS POR GRUPOS. FUENTE: GALLEGO (2008) 24

ILUSTRACIÓN 9. PATRONES HIPOTÉTICOS DE ERRORES RESIDUALES AL CUADRADO $\boldsymbol{ei2}$ FRENTE A $\boldsymbol{yi}$. FUENTE: GUJARATI (2004) 25

ILUSTRACIÓN 10. DIAGRAMA DE DISPERSIÓN DE $\boldsymbol{xi}$ FRENTE A $\boldsymbol{ei}$. FUENTE: ELABORACIÓN PROPIA CON EXCEL. 26

ILUSTRACIÓN 11. DIAGRAMAS DE DISPERSIÓN DE $\boldsymbol{xi2}$ FRENTE A $\boldsymbol{ei2}$. FUENTE: ELABORACIÓN PROPIA CON EXCEL 26

ILUSTRACIÓN 12. DIAGRAMAS DE DISPERSIÓN DE $\boldsymbol{Yi}$ FRENTE A $\boldsymbol{ei}$. FUENTE: RAWLINGS ET AL. (1998) 27

ILUSTRACIÓN 13. DATOS MENSUALES DE INGRESOS POR FAMILIA VS GASTOS EN ROPA. FUENTE: KMENTA (1986) 42

ILUSTRACIÓN 14. COMPARATIVA RECTAS DE REGRESIÓN MCO Y MCP. FUENTE: KMENTA (1986) 44

ILUSTRACIÓN 15. NUBE DE PUNTOS Y CURVA DE REGRESIÓN LOGÍSTICA. FUENTE: ELABORACIÓN PROPIA CON EXCEL 47

ILUSTRACIÓN 16. CARRERA DESPEGUE A BORDO. FUENTE: ELABORACIÓN PROPIA 50

ILUSTRACIÓN 17. TOMA VERTICAL A BORDO. FUENTE: ELABORACIÓN PROPIA 50

ILUSTRACIÓN 18. VUELO DE APOYO A TIERRA. FUENTE: REVISTA GENERAL DE MARINA (2023) 50

ILUSTRACIÓN 19. TOMA VERTICAL DE UN *F-35B* A BORDO DE UN PORTAERONAVES CLASE *AMERICA*. FUENTE: *US MARINE CORPS*. 51

ILUSTRACIÓN 20. DISPOSICIÓN GENERAL DEL *AV-8B PLUS (HARRIER II).* FUENTE: NAVAIR (2021) 52

ILUSTRACIÓN 21. RESULTADO BPG TEST CON RSTUDIO. FUENTE: PROPIA CON EXCEL 55

ILUSTRACIÓN 22. DIAGRAMA DE DISPERSIÓN DATOS DE PARTIDA. FUENTE: ELABORACIÓN PROPIA CON EXCEL. 57

ILUSTRACIÓN 23. DIAGRAMA DE DISPERSIÓN DE RESIDUOS $\boldsymbol{e}$ Y VARIABLE ESTIMADA $\hat{Y}$. FUENTE: ELABORACIÓN PROPIA CON EXCEL 58

ILUSTRACIÓN 24. DIAGRAMA DE DISPERSIÓN DE $\boldsymbol{e2}$ FRENTE A LA VARIABLE $\boldsymbol{x2}$. FUENTE: ELABORACIÓN PROPIA CON EXCEL 59

ILUSTRACIÓN 25. DIAGRAMA DE DISPERSIÓN DE $\boldsymbol{e2}$ FRENTE A $\boldsymbol{y}$, Y RECTA DE REGRESIÓN $\boldsymbol{e2 = 44867y - 6 \times 109}$. FUENTE: ELABORACIÓN PROPIA CON EXCEL 59

ILUSTRACIÓN 26. DIAGRAMA DE DISPERSIÓN DE $\boldsymbol{|e|/s}$ FRENTE A$\boldsymbol{y}$, Y RECTA DE REGRESIÓN $\boldsymbol{|e|/s = 0,0015y - 128,2}$. FUENTE: ELABORACIÓN PROPIA CON EXCEL 60

ILUSTRACIÓN 27. DIAGRAMA DE DISPERSIÓN CON RECTA DE REGRESIÓN. FUENTE: ELABORACIÓN PROPIA CON SPSS 62

ILUSTRACIÓN 28 RESULTADO COMPARATIVO RECTA REGRESIÓN MCO Y MCP. FUENTE: ELABORACIÓN PROPIA CON *R* 69

Tablas

TABLA 1. DATOS ANUALES INGRESOS VS GASTOS EN ROPA DE 20 FAMILIAS. FUENTE: KMENTA (1986) 42

TABLA 2. VALORES OBSERVADOS REGRESIÓN LOGÍSTICA. FUENTE: ELABORACIÓN PROPIA.....46

TABLA 3. RESULTADOS NUMÉRICOS EJEMPLO REGRESIÓN LOGÍSTICA. FUENTE: ELABORACIÓN PROPIA CON EXCEL. 47

TABLA 4. CÓDIGO APLICACIÓN MÉTODO MCP. FUENTE: ELABORACIÓN PROPIA CON R. 48

TABLA 5. RESULTADOS NUMÉRICOS EJEMPLO REGRESIÓN LOGÍSTICA. FUENTE: ELABORACIÓN PROPIA CON R. 48

TABLA.6. RESULTADOS MCO DE LAS SUBMUESTRAS N_1 Y N_2. FUENTE: PROPIA CON EXCEL.....53

TABLA 7. RESULTADOS MCO Y MCO DE BREUSCH-PAGAN. FUENTE: PROPIA CON EXCEL 54

TABLA 8. RESULTADOS MCO Y MCO DE KOENKER-BASSETT. FUENTE: PROPIA CON EXCEL.....55

TABLA 9. RESULTADOS MCO Y MCO DE WHITE. FUENTE: PROPIA CON EXCEL.....56

TABLA 10. RESULTADO APLICACIÓN MÉTODO MCO. FUENTE: ELABORACIÓN PROPIA CON EXCEL. 57

TABLA 11. CÓDIGO APLICACIÓN MÉTODO MCO CON *R.* FUENTE: ELABORACIÓN PROPIA.....60

TABLA 12. RESULTADO APLICACIÓN MÉTODO MCO CON *R.* FUENTE: ELABORACIÓN PROPIA.....61

TABLA 13. COEFICIENTE DETERMINACIÓN CON MCO. FUENTE: ELABORACIÓN PROPIA CON SPSS.....61

TABLA 14. COEFICIENTES DE REGRESIÓN Y VARIANZA CON MCO. FUENTE: ELABORACIÓN PROPIA CON SPSS 62

TABLA 15. RESULTADO ANOVA DE MCO. FUENTE: ELABORACIÓN PROPIA CON SPSS.....62

TABLA 16. RESULTADOS APLICACIÓN MÉTODO MCO. FUENTE: ELABORACIÓN PROPIA CON GRETL 63

TABLA 17. EXTRACTO PARÁMETROS UTILIZADOS EN APLICACIÓN DEL MÉTODO MCP. FUENTE: ELABORACIÓN PROPIA CON EXCEL 63

TABLA 18. RESULTADO APLICACIÓN MÉTODO MCP. FUENTE: ELABORACIÓN PROPIA CON EXCEL 64

TABLA 19. CÓDIGO APLICACIÓN MÉTODO MCP. FUENTE ELABORACIÓN PROPIA CON *R* 64

TABLA 20. RESULTADO APLICACIÓN MÉTODO MCP. FUENTE ELABORACIÓN PROPIA CON *R*.....65

TABLA 21. CÓDIGO APLICACIÓN MÉTODO MCP (MODELO PONDERADO). FUENTE ELABORACIÓN PROPIA CON *R* 65

TABLA 22. RESULTADO APLICACIÓN MÉTODO MCP (MODELO PONDERADO). FUENTE ELABORACIÓN PROPIA CON *R* 66

TABLA 23. CÓDIGO APLICACIÓN MÉTODO MCP (MODELO FUNCIÓN *WEIGHTS*). FUENTE ELABORACIÓN PROPIA CON *R*.....66

TABLA 24. RESULTADO APLICACIÓN MÉTODO MCP (*WEIGHTS* CON PESOS). FUENTE: ELABORACIÓN PROPIA CON R.....67

TABLA 25. CÓDIGO APLICACIÓN MÉTODO MCP (*WEIGHTS* CON PROBABILIDAD). FUENTE: ELABORACIÓN PROPIA CON R ..67

TABLA 26. RESULTADO APLICACIÓN MÉTODO MCP (*WEIGHTS* CON PROBABILIDAD) FUENTE ELABORACIÓN PROPIA CON *R*68

TABLA 27 CÓDIGO REPRESENTACIÓN COMPARATIVA MÉTODOS MCO Y MCP FUENTE ELABORACIÓN PROPIA CON *R* 68

TABLA 28. COEFICIENTE DETERMINACIÓN CON MCP. FUENTE: PROPIA CON SPSS 70

TABLA 29. RESULTADO ANOVA DE MCP. FUENTE: ELABORACIÓN PROPIA CON SPSS 70

TABLA 30. COEFICIENTES DE REGRESIÓN CON MCP. FUENTE: ELABORACIÓN PROPIA CON SPSS 70

TABLA 31. CUADRO RESUMEN REGRESIÓN MCP. FUENTE: ELABORACIÓN PROPIA CON GRETL 71

TABLA 32 DATOS DE HORAS DE VUELO Y CONSUMO DE COMBUSTIBLE NOVENA ESCUADRILLA. FUENTE: ARMA AÉREA ARMADA 80

TABLA.33. DATOS MUESTRA N1 PARA REALIZAR MCO GQ1. FUENTE: ELABORACIÓN PROPIA CON EXCEL 82

TABLA 34. DATOS MUESTRA N2 PARA REALIZAR MCO GQ2. FUENTE: ELABORACIÓN PROPIA CON EXCEL 83

TABLA 35. RESULTADOS OPERACIÓN GQ TEST EN LAS MUESTRAS GQ1 Y GQ2. FUENTE: ELABORACIÓN PROPIA CON EXCEL 84

TABLA 36. RESULTADOS DE OPERACIONES TEST DE BREUSCH-PAGAN-GODFREY. FUENTE: ELABORACIÓN PROPIA CON EXCEL 84

TABLA 37 RESULTADOS DE OPERACIONES TEST DE KOENKER-BASSET. FUENTE: ELABORACIÓN PROPIA CON EXCEL 85

TABLA 38. RESULTADOS DE OPERACIONES TEST DE WHITE. FUENTE: ELABORACIÓN PROPIA CON EXCEL 86

Lista de Acrónimos

AGA	Academia General del Aire
BLUE	Best Linear Unbiased Estimator
CUD	Centro Universitario de la Defensa
ELIO	Estimador Lineal Insesgado Óptimo
GRETL	Gnu Regression, Econometrics and Time-series Library
MCG	Mínimos Cuadrados Generalizados
MCO	Mínimos Cuadrados Ordinarios
MCP	Mínimos Cuadrados Ponderados
MUTAD	Máster Universitario de Técnicas de Ayuda a la Decisión
SSEX	Suma de Cuadrados Explicada
SSNEX	Suma de Cuadrados No Explicada
SST	Suma Total de Cuadrados
TFM	Trabajo Final de Máster
V/STOL	Vertical/Short Take-Off Landing

Capítulo 1. Introducción

La aplicación de las habilidades adquiridas en la asignatura "Modelos de Regresión" del Máster constituye la motivación fundamental del presente trabajo. Aunque el currículo del MUTAD solo contempla el **Método de Mínimos Cuadrados Ordinarios (MCO),** se considera de sumo interés como ampliación de la formación en este ámbito, el método destacado en el título de este trabajo, lo que constituye en sí la justificación fundamental de este TFM.

El modelo de inferencia sobre regresión estudiado mediante la aplicación de MCO exige la **hipótesis de homocedasticidad** de los errores, es decir, que la varianza de las perturbaciones permanezca constante con independencia del valor que tome el regresor. Si se suprime esta hipótesis, no se altera la centralidad de los coeficientes de regresión de MCO, pero sí cambian las varianzas y dejan de ser constantes para cada observación. En ese caso se presenta la **heterocedasticidad** ocasionando que los **estimadores** del modelo de regresión con MCO, aunque **insesgados**, dejen de ser **eficientes** y los **contrastes de validación** del modelo no aplicables bajo esta anomalía.

En general, la heterocedasticidad no es posible detectarla directamente por lo que se recurre como primera aproximación a la representación gráfica de los residuos, y de forma analítica, a la inferencia estadística. Partiendo de la estimación de las varianzas, o del valor conocido de éstas para cada observación, el método que permite transformar las variables para conseguir un modelo de regresión homocedástico es el de **Mínimos Cuadrados Ponderados (MCP).** Este método logra corregir el método de MCO bajo heterocedasticidad, mediante la minimización de la varianza residual con la introducción de pesos que introducen una ponderación de las observaciones permitiendo que los estimadores resulten insesgados y eficientes.

Finalmente, este trabajo se ha basado en el modelo de regresión lineal simple con la implementación de técnicas estadísticas para la detección, contraste y validación de la heterocedasticidad con forma lineal o no lineal. Es por ello que el enfoque del TFM se ha orientado particularmente al estudio de los fundamentos que a una recopilación exhaustiva de métodos estadísticos basados en diferentes variantes del método MCP.

1.1. Objetivos

El objetivo de este Trabajo Final de Master es, primeramente enunciar el método de mínimos cuadrados ordinarios para la obtención del modelo de regresión lineal simple en presencia de homocedasticidad y mostrar los contrastes de hipótesis para inferencia estadística básicos aplicables bajo esta hipótesis. A continuación, dar a conocer el problema de la heterocedasticidad, revisar el método de MCO bajo esta hipótesis y explicar métodos estadísticos de contraste que detecten la presencia de este fenómeno.

Seguidamente, comprobar que el método de mínimos cuadrados ordinarios del modelo de regresión lineal deja de ser eficiente en presencia de heterocedasticidad, para lo cual se demostrará que el método de mínimos cuadrados ponderados corrige el método de MCO comprobando además que la varianza de los estimadores bajo el método de MCP es menor que en el primer caso, para lo cual se acompañarán ejemplos concretos.

Finalmente, a partir de datos reales de una aeronave de la Armada, realizar análisis de los mismos como aplicación directa de la metodología empleada en el presente TFM

1.2. Estructura

El contenido del trabajo se estructura conforme se detalla a continuación:

El primer capítulo, es el elemento iniciador donde se introduce la motivación de la temática abordada, seguida de unos objetivos directos, junto a una sucinta estructuración del contenido finalizando con la metodología empleada en la confección del trabajo.

A lo largo del **segundo capítulo**, se introduce la regresión lineal simple, mediante la explicación del método de mínimos cuadrados ordinarios, aplicado al modelo teórico y las inferencias aplicables bajo la hipótesis de homocedasticidad. A partir de aquí, se

aborda el problema de la heterocedasticidad, las causas, consecuencias y detección gráfica del fenómeno además de las herramientas estadísticas de inferencia aplicables.

En el **tercer capítulo**, se analizan las propiedades de los estimadores de MCO bajo la hipótesis de heterocedasticidad, y se introduce el método de MCP bajo este supuesto. Además, se aportan ejemplos comparativos en presencia heterocedástica de ambos métodos, así como de varianza conocida y estimada, culminando con la aplicación directa del método MCP en el modelo de regresión logística.

El **cuarto capítulo** representa un ejercicio práctico de aplicación directa del método de MCP, en base a datos reales de consumo de combustible y horas de vuelo recopilados durante 11 años de operación de la Novena Escuadrilla de aeronaves de la Armada, que ilustra mediante tablas y resultados computados con diverso software, *Excel, SPSS, R y GRETL*, lo expuesto en los capítulos precedentes.

El **quinto capítulo** sintetiza las conclusiones obtenidas tras el estudio realizado en el capítulo anterior, donde se reseñan posibles líneas futuras de desarrollo y apertura de nuevos contenidos asociados a la temática del TFM.

En último lugar, en los **apéndices A y B,** se disponen los datos de partida, los procesos utilizados para la obtención de los resultados de MCO, MCP y contrastes de hipótesis en condiciones de heterocedasticidad.

1.3. Metodología

La metodología utilizada para la elaboración de este trabajo consta de tres partes diferenciadas:

La primera, centrada en el capítulo 2, donde a partir de la teoría introductoria del método de mínimos cuadrados ordinarios y la inferencia asociada con la hipótesis de homocedasticidad, se da paso a la situación heterocedástica, sus causas y detección preliminar en el modelo de regresión junto con los contrastes aplicables en dicha situación. Seguidamente, en el capítulo 3 se incorpora el método de mínimos cuadrados ponderados a partir de la heterocedasticidad como corrección al método de MCO, donde entre otros, se compara la eficiencia e insesgadez de los estimadores obtenidos mediante ambos métodos.

La segunda, localizada en el capítulo 4 en un ejercicio práctico de aplicación, diseñada a partir de datos reales de consumo de combustible y tiempo de vuelo de un tipo de aeronave de la Armada, donde se han implementado las técnicas de modelos de regresión antedichas, la resolución numérica comparativa con los paquetes software $Excel, SPSS, R\ y\ GRETL$, para concluir con una validación del modelo de regresión en situación heterocedástica.

La última parte, capítulo 5, se focaliza en la discusión de los resultados obtenidos tras el ejercicio práctico y las conclusiones derivadas del mismo junto con la apertura de posibles líneas futuras.

Capítulo 2. La regresión lineal simple

La regresión lineal simple estriba en hallar la forma de relacionar estadísticamente dos variables, una independiente X (explicativa, endógena o regresora) y otra Y (explicada, exógena o respuesta) dependiente de la anterior. El modelo teórico que fundamenta dicha relación será capaz de establecer una función que más se aproxime, bajo cierto criterio métrico, a los datos de partida, que describa su tendencia y estime valores intermedios con cierta precisión (Baenas 2022). Esta relación[1] queda supeditada a la aleatoriedad de la variable X, que controlará el comportamiento del modelo dado, Uriel (2019). A priori, la relación funcional más sencilla es la *lineal*.

Reseñar que la regresión múltiple no se aborda en este trabajo por falta de tiempo, aunque se introduzca, en capítulos posteriores, por razones didácticas.

2.1. El método de Mínimos Cuadrados Ordinarios (MCO)

Partiendo de una muestra de n pares de observaciones finitas $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, o nube de puntos, procedentes de una variable aleatoria bidimensional (x, y), el método de mínimos cuadrados ordinarios (MCO) establece que los pares de datos pueden ser ajustados a una función de tal manera que los errores (suma de las desviaciones entre los datos y el valor teórico) pueden ser minimizados. Es decir, considerando el modelo teórico a partir de la función $f(x; \alpha_j)$ se ha de determinar el conjunto de parámetros α_j $(j = 1, 2, \dots, p)$ que minimiza la suma

[1] Las relaciones pueden ser clasificadas como *deterministas* o *estocásticas*. La relación entre X e Y caracterizada por $y=f(x)$ es *determinista* si a cada valor de X le corresponde solo un valor de Y. Asimismo, dicha relación es *estocástica* si dado para cada valor de X existe una distribución de probabilidad de valores de Y, Kmenta (1986).

cuadrática de las desviaciones (o desviación cuadrática) entre los valores observados y_i con el correspondiente valor teórico $f(x;\ \alpha_j)$. En resumen, se trata de minimizar el estadístico *varianza residual (V_r)* o *error cuadrático medio (e^2)*,

$$V_r\ (\alpha_j) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - f(x_i;\alpha_j)\right]^2. \tag{2.1}$$

El ajuste óptimo de los datos, precisará de la definición de estadísticos diferentes a la de V_r para poder decidir sobre el grado de ajuste, además de que se pueda realizar inferencia estadística, con la que posteriormente comprobar la bondad del ajuste del modelo teórico a las observaciones propuestas.

Para hacer mínima la función $V_r\ (\alpha_j)$ hay que establecer la siguiente condición necesaria[2] para funciones diferenciables,

$$\frac{\partial V_r}{\partial \alpha_0} = 0;\ \frac{\partial V_r}{\partial \alpha_1} = 0;\ \dots,\frac{\partial V_r}{\partial \alpha_p} = 0. \tag{2.2}$$

Resolviendo el sistema de *p* ecuaciones (2.2) se obtienen los parámetros estimados, $\hat{\alpha}_j$ $(j = 1, 2, \dots, p)$ o estimador muestral, para los que la varianza residual es extrema. Finalmente, la función definida por $y = f(x;\ \hat{\alpha}_j)$ se denomina curva de regresión.

2.1.1. El modelo de regresión lineal simple

El modelo teórico es una recta, de *pendiente* α_1 e *intercepto en y u ordenada en el origen* α_0, tal que

$$f(x;\ \alpha_0, \alpha_1\ ) = \alpha_0 + \alpha_1 x\ . \tag{2.3}$$

Sustituyendo el resultado de (2.3) en (2.1) queda esta última expresión como sigue,

$$V_r\ (\alpha_0, \alpha_1) = \frac{1}{n}\sum_{i=1}^{n}[y_i - (\alpha_0 + \alpha_1 x_i\ )]^2 \tag{2.4}$$

Los estimadores $\hat{\alpha}_0, \hat{\alpha}_1$ de α_0, α_1 respectivamente, se obtienen aplicando a la ec. (2.4) las ecuaciones de (2.2) resultando,

[2] La condición suficiente de mínimo se asegura con el estudio del carácter de la matriz hessiana asociada, que se omite por brevedad.

$$\frac{\partial V_r}{\partial \alpha_0} = 0; \quad \sum_{i=1}^{n} [y_i - (\alpha_0 + \alpha_1 x_i)] = 0,$$
$$\frac{\partial V_r}{\partial \alpha_1} = 0; \quad \sum_{i=1}^{n} [y_i - (\alpha_0 + \alpha_1 x_i)]\, x_i = 0. \tag{2.5}$$

Las ecuaciones resultante forman un sistema compatible determinado, cuando la matriz del sistema es regular ($Var(x) > 0$), de dos ecuaciones con dos incógnitas denominadas *ecuaciones normales* o *condiciones de primer orden de mínimos cuadrados* conforme se indica en Uriel (2019),

$$n\alpha_0 + \alpha_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$
$$\alpha_0 \sum_{i=1}^{n} x_i + \alpha_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{2.6}$$

Resolviendo analíticamente el sistema lineal anterior se obtiene que,

$$\hat{\alpha}_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2},$$
$$\hat{\alpha}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \tag{2.7}$$

Finalmente, las estimaciones de y en función de la variable regresora proporcionan la *recta de regresión*,

$$\hat{y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i. \tag{2.8}$$

La interpretación de cada valor calculado desde la recta de regresión $\hat{y}_i$ tiene dos vertientes, una de *estimación* de la media de Y para cada valor de X, y otra de *predicción* del valor de *Y* que se debiera obtener en una observación futura para ese nivel de X. El signo de $\hat{\alpha}_1$ determina si X e Y están relacionados directa o inversamente (pendiente positiva o negativa).

En este contexto, el error entre el modelo teórico y los datos ofrece dos partes diferenciadas (Ilustración 1),

1) *Error aleatorio*, dado por la diferencia entre el modelo teórico y los datos (parte aleatoria o no "explicada" de la variable),

$$e_i = y_i - f(x_i; \alpha_j) \tag{2.9}$$

2) *Error residual*, definido por la diferencia entre los datos y la predicción teórica, dad por los estimadores muestrales,

$$\hat{e}_i = y_i - \hat{y}_i \tag{2.10}$$

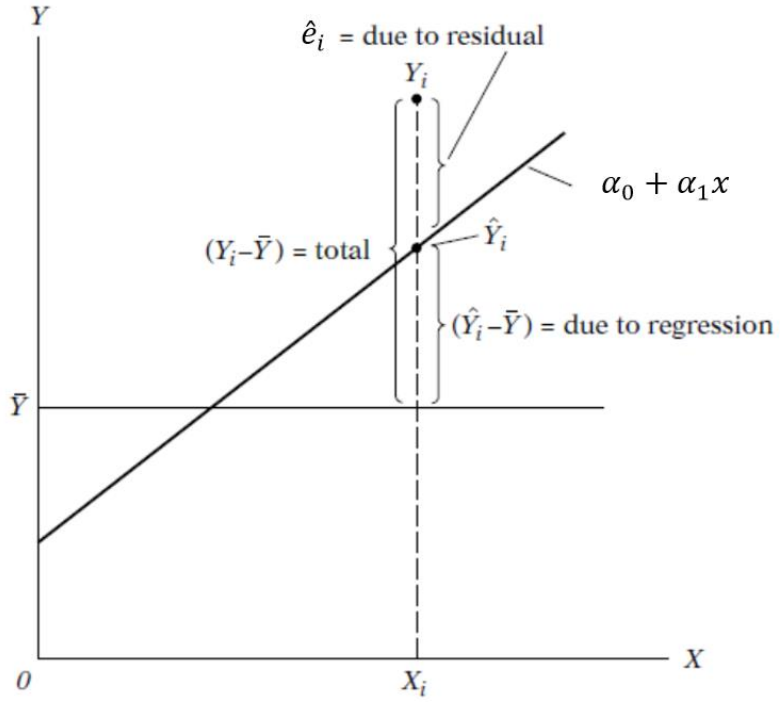


Ilustración 1. Error aleatorio y error residual de un valor x_i en una ecuación de regresión. Fuente: Gujarati (2004)

Partiendo de dichos errores, una medida cuantitativa preliminar de la bondad del ajuste se deriva a partir de la descomposición de la varianza de los datos. La varianza total de Y se descompone en dos componentes, una suma cuadrática debida al modelo ajustado y otra debida a la suma cuadrática de errores aleatorios en los datos:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2.11}$$

esto es, podemos descomponer la varianza muestral en dos componentes, la explicada por la regresión lineal, y la no explicada o residual (sumandos primero y segundo de

2.11, respectivamente). Esto permite el uso de una técnica de inferencia convencional de análisis de varianza (*ANOVA*).

Asimismo, considerando de las ecuaciones de regresión (2.5) que,

$$\sum_{i=1}^{n} e_i = 0, \sum_{i=1}^{n} e_i\, x_i = 0, \tag{2.12}$$

y a partir de la siguiente expresión,

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} (\alpha_0 + \alpha_1 x_i + e_i\,) = \alpha_0 + \alpha_1 \bar{x} + \frac{1}{n}\sum_{i=1}^{n} e_i, \tag{2.13}$$

se llega finalmente a la ecuación de regresión que viene dada por los promedios de y, y x respectivamente

$$\bar{y} = \hat{\alpha}_0 + \hat{\alpha}_1 \bar{x}. \tag{2.14}$$

Dado que la *varianza muestral* ($\hat{\sigma}_x^2$) de la variable x se obtiene a partir de la expresión

$$\hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2, \tag{2.15}$$

y la *covarianza muestral* (σ_{xy}) de $x\ e\ y$ viene dada por,

$$\hat{\sigma}_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\bar{y}, \tag{2.16}$$

despejando $\hat{\alpha}_1$ de (2.14) se reformulan los parámetros de regresión en función de las expresiones de las ec. (2.15) y (2.16) para llegar, finalmente, a la recta de regresión en función de los promedios muestrales:

$$\begin{aligned} \hat{\alpha}_1 &= \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}, \\ \hat{\alpha}_0 &= \bar{y} - \hat{\alpha}_1 \bar{x}, \\ y - \bar{y} &= \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}(x - \bar{x}), \end{aligned} \tag{2.17}$$

concluyendo que esta ecuación simplifica la expresión formal de la recta de regresión.

2.1.2. El coeficiente de correlación lineal

A partir de la recta de regresión lineal obtenida mediante MCO en el apartado anterior, se establece el *coeficiente de determinación* $\hat{r}^2$, a través de la proporción entre las varianzas residuales de la recta de regresión y la de la varianza de la variable explicada Y de tal manera que

$$\hat{r}^2 = 1 - \frac{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}; \hat{r}^2 \in [0,1] \quad (2.18)$$

Este parámetro constituye una medida de la bondad de ajuste debido a que:

1. Representa el porcentaje de varianza residual que puede ser explicado por el modelo
2. Si $\hat{r}^2 = 0$, no puede establecerse una relación lineal entre las variables
3. Si $\hat{r}^2 = 1$, se establece un ajuste completo entre los datos y la recta de regresión

Por conveniencia se puede expresar la varianza residual de la recta de regresión como

$$V_r(\hat{\alpha}_0, \hat{\alpha}_1) = \hat{\sigma}_y^2(1 - \hat{r}^2). \quad (2.19)$$

El coeficiente de determinación lineal, $\hat{r}^2$, se obtiene elevando al cuadrado el *coeficiente de correlación lineal de Pearson*, $\hat{r}$, definido este como el grado de asociación o independencia entre las dos variables X *e* Y como sigue

$$\hat{r} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\sqrt{\overline{y^2} - \bar{y}^2}}; \ \hat{r} \in [-1, 1] \quad (2.20)$$

Algunas de las propiedades de $\hat{r}$, se reseñan a continuación (ver ilustración 2):

1. Dado que $\hat{r}^2 \leq 1$, entonces se tiene que $-1 \leq \hat{r} \leq 1$
2. Si $\hat{r} = \pm 1$ la varianza residual Vr es nula, con lo que la recta de regresión se ajusta completamente a los datos.
3. Si $\hat{r} = 1 \quad (\hat{\sigma}_{xy} > 0)$ la pendiente de la recta es creciente $(+)$
4. Si $\hat{r} = -1 \ (\hat{\sigma}_{xy} < 0)$ la pendiente de la recta es decreciente $(-)$
5. Si $\hat{r} = 0 \quad (\hat{\sigma}_{xy} = 0)$ la recta de regresión es $y = \bar{y}$ (independiente de x)

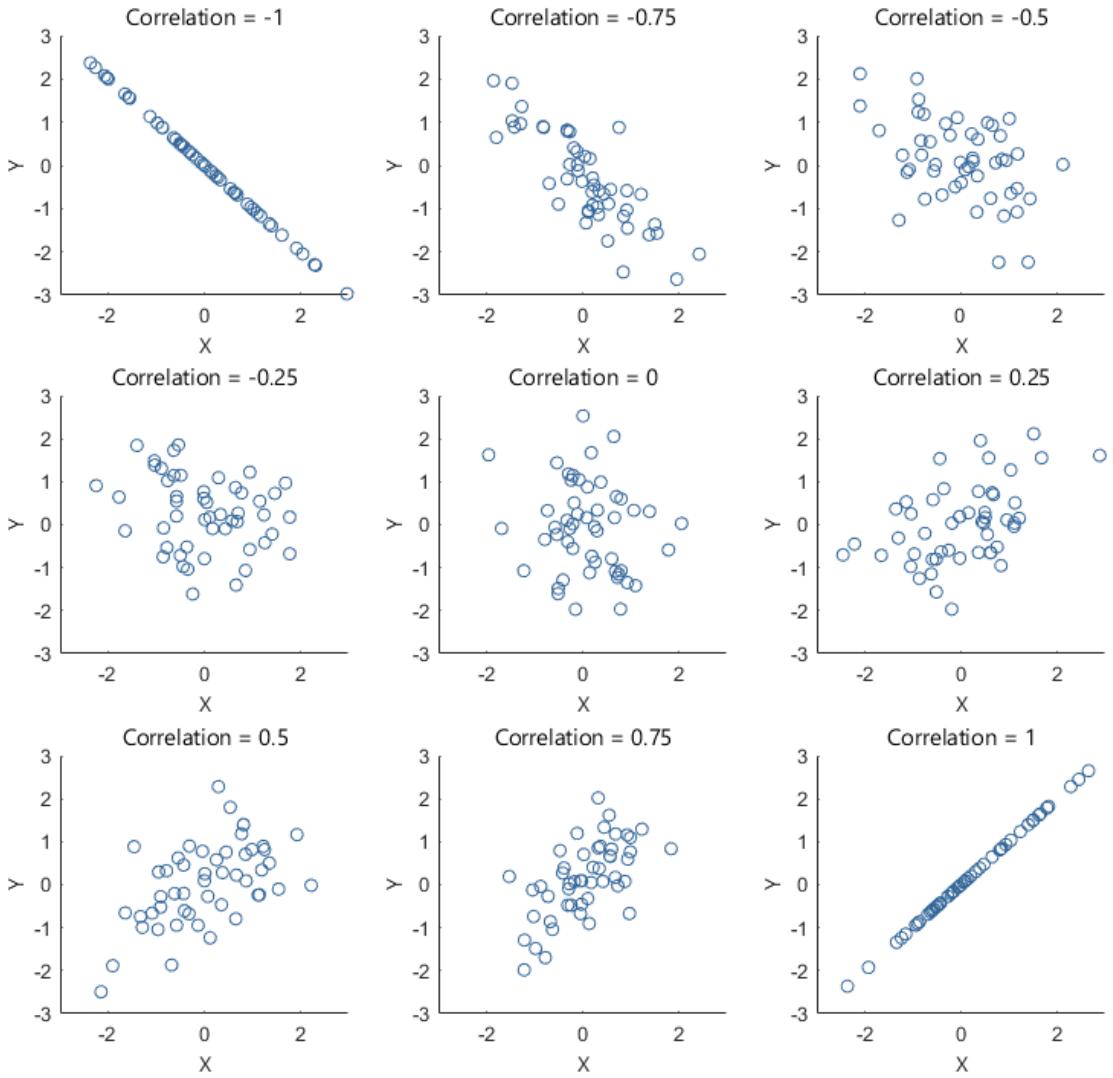


Ilustración 2. Patrones de correlación en función del valor de $\hat{r}$.Fuente: (www.statlect.com, s. f.)

Finalizando este apartado se introducen definiciones que serán utilizadas en apartados posteriores. Formalizando definiciones previas, a partir de la ec. (2.19) se descompone la varianza de la variable y en dos términos diferenciados, atendiendo a la realizada en (2.11)

$$\hat{\sigma}_y^2 = \hat{\sigma}_y^2(1 - \hat{r}^2) + \hat{r}^2\hat{\sigma}_y^2, \tag{2.21}$$

donde al término $\hat{r}^2\hat{\sigma}_y^2$ se le denomina *varianza explicada* (debida a la regresión lineal), y al término de la varianza residual $\hat{\sigma}_y^2(1 - \hat{r}^2)$ se le denomina *varianza no explicada (o residual)*. Obviamente, la varianza total de la variable y es la suma de dichas varianzas. A modo de resumen, lo anterior se simplifica mediante la suma de cuadrados con la expresión $SST = SSEX + SSNEX$, cuyo desglose queda así:

Suma total de cuadrados $SST = \sum_{i=1}^{n}(y_i - \bar{y}\,)^2 = n\hat{\sigma}_y^2$

Suma de cuadrados explicada $SSEX = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}\,)^2 = n\hat{\alpha}_1^2\hat{\sigma}_x^2$, (2.22)

Suma de cuadrados no explicada $SSNEX = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{e}_i^2$.

2.2. Inferencias estadísticas sobre un parámetro poblacional

La estimación muestral realizada de los parámetros poblacionales para la obtención de la recta de regresión y del coeficiente de correlación lineal, precisa de un contraste mediante técnicas de inferencia estadística.

Para ello, se recurre al modelo de regresión lineal simple dado por,

$$y_i = \alpha_0 + \alpha_1 x_i + e_i = f(x_i) + e_i, i = 1, 2, \dots, n, (n > 2), \quad (2.23)$$

siendo y_i variables aleatorias incorrelacionadas[3]; x_i variables no aleatorias o controladas; $\alpha_0\ y\ \alpha_1$, parámetros poblacionales estimados mediante MCO; y e_i variables aleatorias incorrelacionadas.

Conocer el valor esperado y la varianza de los estimadores de MCO permite describir la precisión de los estimadores de MCO. No obstante, para realizar una inferencia estadística se requiere conocer la distribución muestral conforme a los supuestos de Gauss-Markov conforme indica Wooldridge (2010)[4], suponiendo que cada variable y_i sigue una distribución normal de media $f(x_i) = \alpha_0 + \alpha_1 x_i$, y desviación típica σ_i. Es decir, la función de densidad de cada variable observable es,

$$f(y/x_i) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{[f(x_i)-y]^2}{2\sigma_i^2}} \quad (2.24)$$

[3] Estadísticamente independientes.

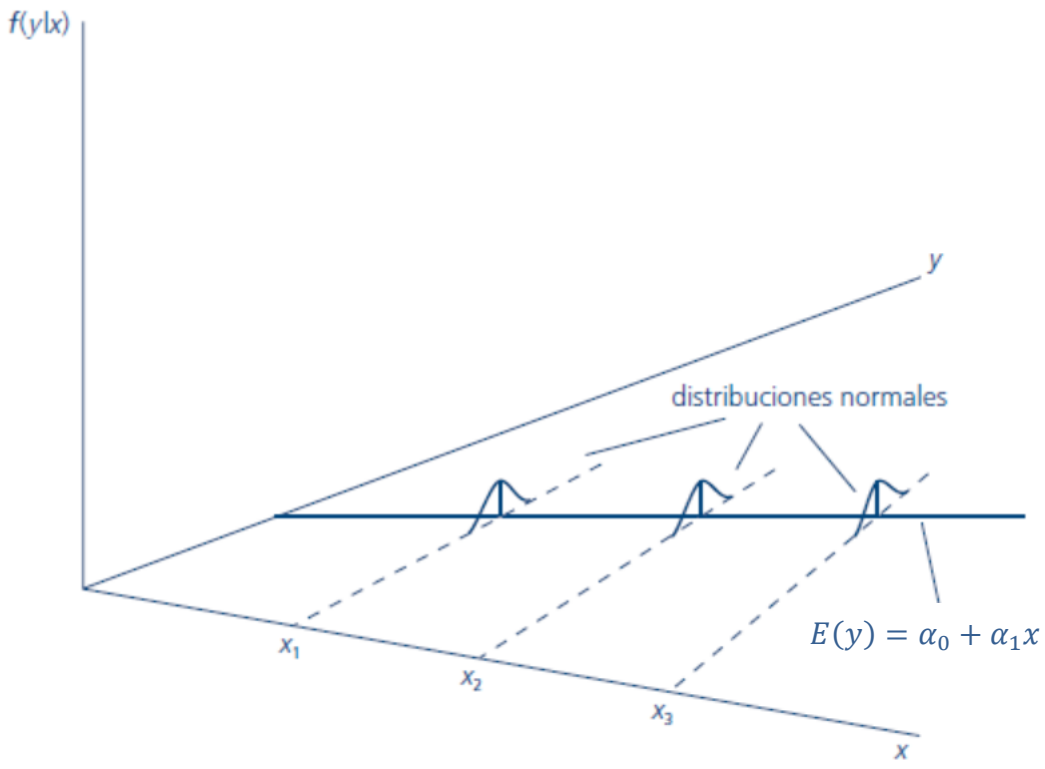
[4] El *Teorema de Gauss Markov* justifica el uso del método de MCO mostrando que sus estimadores son insesgados y suponiendo la distribución normal de su variable aleatoria observable y_i.

lo que se denota como y_i se distribuye con una $N(f(x_i), \sigma_i)$

$$\frac{y_i - f(x_i)}{\sigma_i} = \frac{e_i}{\sigma_i} \sim N(0,1), i = 1, 2, \dots, n. \tag{2.25}$$

En la ilustración 3, se muestra gráficamente que la variable explicada se distribuye normalmente con media $f(x_i)$ y varianza σ_i^2. La función de probabilidad de y condicionada a x se representa mediante $f(y/x)$ y la esperanza matemática de y, $E(y) = \alpha_0 + \alpha_1 x$, es el valor medio poblacional de la variable, cuya estimación se realiza mediante la recta de regresión.

En base a la hipótesis de *homocedasticidad* (ya que con carácter general se desconoce, a priori, si es una propiedad de la variable aleatoria) se simplifica el modelo de regresión lineal, suponiendo que la varianza del error, $Var(e_i)$, es la misma para cada observación de la variable, con lo que para una población dada, $\sigma_i = \sigma = constante$.



.

Ilustración 3. Distribución normal homocedástica con la variable explicativa x. Fuente: Wooldridge (2010)

En resumen, asumiendo que las variables x e y se observan sin abordar error de medida[5] alguno, se muestran las condiciones en las que el modelo de regresión lineal (2.23) es aplicable:

a) $y_i = \alpha_0 + \alpha_1 x_i + e_i, i = 1, 2, \dots, n, (n > 2)$
b) $Media\ del\ error\ e_i, E(e_i) = 0$
c) $Homocedasticidad, Var(e_i) = E(e_i^2) = \sigma^2$
d) $Error\ aleatorio\ incorrelado, Cov(e_i, x_i) = 0, y\ Cov(e_i, e_j) = 0$
e) $Variabilidad\ de\ los\ valores\ de\ la\ variable\ independiente, x, Var(x) > 0$
f) $Distribución\ del\ modelo\ para\ inferencia, \frac{e_i}{\sigma} \sim N(0,1).$

2.2.1. Contraste de la regresión lineal simple

Estableciendo de forma breve las bases del contraste de hipótesis asociado al análisis de varianzas en la regresión simple, la suma del cuadrado de las normales tipificadas (2.25) sigue una distribución χ^2_{n-2},[6]

$$\sum_{i=1}^{n} \frac{\hat{e}_i^2}{\sigma^2} = \frac{SSNEX}{\sigma^2} \sim \chi^2_{n-2}. \qquad (2.26)$$

Por otro lado, asumiendo la hipótesis de que $y_i \sim N(\mu, \sigma)$, se distribuye respecto a $\mu = cte$, $x\ e\ y$ no están relacionadas, se establece para la varianza muestral, la siguiente relación en base al *Teorema de Fisher*[7] reseñado en Nortes (1993),

[5] En el modelo de regresión se asume que tanto la variable dependiente como las regresoras se miden sin errores inherentes al proceso de obtención de los valores muestrales de la variable (p.ej. cálculo de incertidumbres). Aunque en la práctica esto no ocurre, en general, se puede ampliar esta aspecto en Gujarati (2004) o John O. Rawlings et al. (1998).

[6] Los $n - 2$ grados de libertad, son debidos a las dos ecuaciones (2.12) que se precisan para obtener la recta de regresión lineal.

[7] El Teorema de Fisher establece que para una variable aleatoria $X \sim N(\mu, \sigma)$, se pueden obtener muestras aleatorias simples de tamaño n tal que se verifica que $\bar{x}$ y s^2 (cuasivarianza muestral), son independientes de tal manera que, $\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ y $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1.}$.

$$\frac{n\hat{\sigma}_y^2}{\sigma^2} = \frac{SST}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{2.27}$$

Por tanto, considerando la hipótesis nula H_0 que $x\ e\ y$ no se encuentran relacionadas linealmente, se puede expresar la suma de cuadrados explicada $SSEX$ a partir de $SSNEX$ (depende del modelo de regresión) y SST (no depende del modelo de regresión) ya definidos:

$$\frac{SSEX}{\sigma^2} = \frac{SST}{\sigma^2} - \frac{SSNEX}{\sigma^2} \sim \chi^2_{1.} \tag{2.28}$$

A partir de lo establecido en (2.26) y (2.28) y recordando que $SSNEX$ y $SSEX$ son independientes se deduce el estadístico *F de Snedecor-Fisher* Nortes (1993)[8], resultando que

$$F = \frac{\chi^2_1}{\chi^2_{n-2}} = \frac{\frac{1}{1}\frac{SSEX}{\sigma^2}}{\frac{1}{n-2}\frac{SSNEX}{\sigma^2}} = (n-2)\frac{SSEX}{SSNEX} \sim F_{1,n-2}. \tag{2.29}$$

En consecuencia, considerando lo antedicho y dado un nivel de significación α se puede establecer el contraste de hipótesis (unilateral) siguiente:

$$H_0: x, y\ no\ están\ relacionadas\ linealmente$$

$$H_1: x, y\ están\ relacionadas\ linealmente \tag{2.30}$$

- Se acepta H_0 si $F < F_{1,n-2;\alpha}$
- Se rechaza H_0 si $F \geq F_{1,n-2;\alpha}$,

siendo $F_{1,n-2;\alpha}$ el punto crítico de la distribución para los grados de libertad y significación dados.

2.2.2. Contraste para el coeficiente de correlación lineal

Expresando el estadístico de F en términos del coeficiente de determinación $\hat{r}^2$ a partir de (2.19) y (2.29) se obtiene despejando $\hat{r}^2$ que,

[8] La distribución de Snedecor-Fisher $(F_{m,n})$ de (m, n) grados de libertad se expresa mediante el cociente $F_{m,n} = (\frac{1}{m}\sum_{i=1}^{m} x_i^2)/(\frac{1}{n}\sum_{i=1}^{n} y_i^2)$

$$F = (n-2)\frac{SSEX}{SSNEX} = (n-2)\frac{\hat{r}^2}{1-\hat{r}^2}. \tag{2.31}$$

Partiendo en esta ocasión de un contraste de hipótesis bilateral se pude relacionar F con el estadístico t, considerando[9] $t^2 = F$, que se distribuye conforme a una $t\ de\ Student$ a partir de (2.31) como sigue,

$$t = \sqrt{F} = \sqrt{(n-2)\frac{SSEX}{SSNEX}} = \hat{r}\sqrt{\frac{(n-2)}{1-\hat{r}^2}} \sim t_{n-2}. \tag{2.32}$$

Por tanto, el contraste de hipótesis, con un coeficiente de significación α, para el coeficiente de correlación lineal se puede plantear como

$$H_0: r = 0$$

$$H_1: r \neq 0 \tag{2.33}$$

- Se acepta H_0 si $|t| < t_{\ n-2;\alpha/2}$
- Se rechaza H_0 si $|t| \geq t_{\ n-2;\alpha/2}$,

siendo $t_{n-2;\alpha/2}$, el punto crítico de la distribución para los grados de libertad y significación dados.

2.2.3. Contraste para el coeficiente de regresión lineal

Para poder realizar la inferencia del coeficiente de regresión lineal, se precisa obtener la distribución asociada al muestreo de los estimadores de pendiente e intercepto.

Comenzando con el estimador de la pendiente, $\hat{\alpha}_1$, a partir de (2.17) se obtiene

$$\hat{\alpha}_1 = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\hat{\sigma}_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})y_i - \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\bar{y}}{\hat{\sigma}_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\hat{\sigma}_x^2}. \tag{2.34}$$

Considerando que el estimador $\hat{\alpha}_1$ sigue una distribución normal, conforme al teorema de adición de distribuciones normales, por ser combinación lineal de las variables

[9] Se demuestra que la relación es idéntica entre las distribuciones $F_{1,n} = t_n^2$.

independientes normales Kmenta (1986) $y_i = \alpha_0 + \alpha_1 x_i + e_i$, se determina la esperanza y la varianza del coeficiente de regresión **suponiendo la homocedasticidad de la variable aleatoria**, esto es, que la varianza del error permanece constante, $Var(e_i) = \sigma_i^2 = \sigma^2$. Además se demuestra que $\hat{\alpha}_1$ es insesgado,

$$E(\hat{\alpha}_1) = E\left(\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}\right) = \frac{1}{n\hat{\sigma}_x^2}\sum_{i=1}^{n}(x_i - \bar{x})E(y_i) = \frac{1}{n\hat{\sigma}_x^2}\sum_{i=1}^{n}(x_i - \bar{x})(\alpha_0 + \alpha_1 x_i)$$

$$= \frac{\alpha_0}{n\hat{\sigma}_x^2}\sum_{i=1}^{n}(x_i - \bar{x}) + \frac{\alpha_1}{n\hat{\sigma}_x^2}\sum_{i=1}^{n}(x_i - \bar{x})x_i = \alpha_1\frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2} = \boldsymbol{\alpha_1},$$

$$Var(\hat{\alpha}_1) = \sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})}{n\hat{\sigma}_x^2}\right]^2 Var(y_i) = \tag{2.35}$$

$$\frac{1}{\hat{\sigma}_x^2}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{n}\frac{Var(e_i)}{n\hat{\sigma}_x^2} = \frac{\sigma^2}{n\hat{\sigma}_x^2} = \frac{\boldsymbol{\sigma^2}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{x'^2_i}},$$

siendo $x'_i = (x_i - \bar{x})$.

De forma similar se demuestra que el estimador del intercepto, $\hat{\alpha}_0$, también es insesgado a partir de las ecuaciones (2.17) y (2.35)

$$E(\hat{\alpha}_0) = \boldsymbol{\alpha_0}$$

$$Var(\hat{\alpha}_0) = \frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}\left(\frac{\overline{\boldsymbol{x}}^{\boldsymbol{2}}}{\hat{\boldsymbol{\sigma}}_{\boldsymbol{x}}^{\boldsymbol{2}}} + \mathbf{1}\right) \tag{2.36}$$

Como consecuencia de que los estimadores $\hat{\alpha}_0$ y $\hat{\alpha}_1$, son combinaciones lineales de variables independientes normales, se puede escribir que

$$\hat{\alpha}_0 \sim N\left[\alpha_0, \frac{\sigma^2}{n}\left(\frac{\bar{x}^2}{\hat{\sigma}_x^2} + 1\right)\right] \text{ y } \hat{\alpha}_1 \sim N\left(\alpha_1, \frac{\sigma^2}{n\hat{\sigma}_x^2}\right). \tag{2.37}$$

Dado que la varianza poblacional σ^2, en general, desconocida, mediante (2.26) y (2.27), bajo la hipótesis de homocedasticidad, la cuasivarianza muestral, s^2, es un estimador insesgado, obtenido a partir de la varianza residual $V_r(\hat{\alpha}_0, \hat{\alpha}_1)$

$$\sum_{i=1}^{n}\frac{\hat{e}_i^2}{\sigma^2} = \frac{SSNEX}{\sigma^2} = \frac{(n-2)s^2}{\sigma^2} \rightarrow s^2 = \frac{SSNEX}{n-2}. \tag{2.38}$$

En consecuencia, en esta ocasión se plantea el estadístico $t\ de\ Student$ con $n-2$ grados de libertad, cuyo cálculo se determina a partir de (2.22, 2.35 y 2.38) aplicando la hipótesis de $H_0: \alpha_1 = 0$ quedando como,

$$t = \frac{\hat{\alpha}_1 - \alpha_1}{\sqrt{Var(\hat{\alpha}_1)}} = \frac{\hat{\alpha}_1}{\sqrt{\frac{s^2}{n\hat{\sigma}_x^2}}} = \frac{\hat{\alpha}_1}{\sqrt{\frac{\hat{\alpha}_1^2}{(n-2)}\frac{SSNEX}{SSEX}}} \sim t_{n-2}, \tag{2.39}$$

Finalmente, el contraste de hipótesis, con un coeficiente de significación α para el coeficiente de regresión lineal se expresa como,

$$H_0: \alpha_1 = 0$$

$$H_1: \alpha_1 \neq 0 \tag{2.40}$$

- Se acepta H_0 si $|t| < t_{\ n-2;\alpha/2}$
- Se rechaza H_0 si $|t| \geq t_{\ n-2;\alpha/2}$,

siendo $t_{n-2;\alpha/2}$, el punto crítico de la distribución para los grados de libertad y significación dados.

2.3. Heterocedasticidad

En el apartado anterior se asumía el modelo de regresión homocedástico en donde la varianza de e_i permanecía constante, es decir

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

$$E(e_i) = 0\ , \tag{2.41}$$

$$E(e_i^2) = \sigma^2.$$

Esta situación, a modo de ejemplo, se utiliza en modelos relacionados con observaciones de conjuntos temporales, donde los valores de la variable explicativa son de un orden de magnitud similar en todos los puntos de observación, al igual que para los valores de la variable dependiente Kmenta (1986). Así por ejemplo, en una función de consumo el nivel de consumo en los últimos años es de un orden de magnitud similar que el de hace 20 años, y lo mismo se verifica para los ingresos. A no ser que existan algunas circunstancias especiales o el período de tiempo considerado sea excesivo, el supuesto

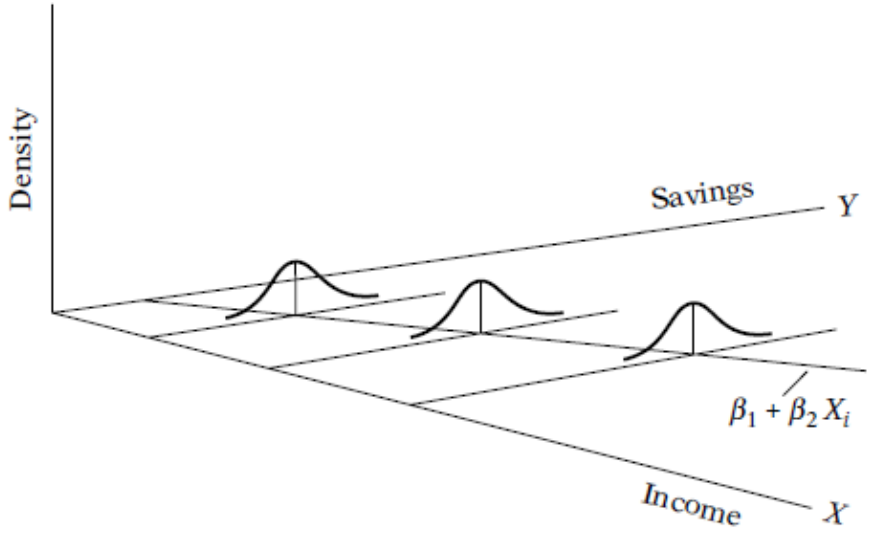
de homocedasticidad en modelos de conjuntos, a priori, se podría considerar razonable siempre y cuando se disponga de información adecuada que lo avale.

En el caso en que la varianza de los errores deja de ser constante, $Var(e_i) \neq \sigma^2$, en las distintas observaciones que integran la muestra, se dice que el término de error presenta *heterocedasticidad*. La detección de este fenómeno está supeditada al tamaño muestral, por lo que la estimación de σ_i^2 pudiera no considerarse aceptable si el número de valores de los que se dispone es reducido.

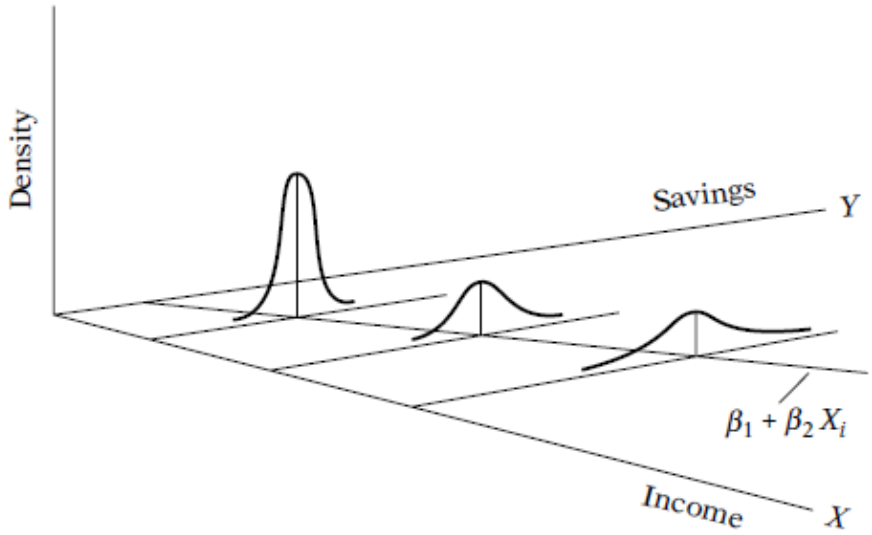
En general, la condición heterocedástica de la perturbación de la regresión se caracteriza por

$$\begin{aligned} Var(e_i) &= \sigma_i^2, \\ E(e_i) &= 0, \\ E(e_i^2) &= \sigma_i^2. \end{aligned} \tag{2.42}$$

Como se verá, en el modelo de regresión lineal esta situación provocará que aunque los estimadores de MCO permanezcan insesgados dejarán de ser eficientes, en el sentido de que no son de varianza mínima, Baenas (2022).



Homoscedastic disturbances.



Heteroscedastic disturbances.

Ilustración 4. Distribución normal homocedástica VS heterocedástica. Fuente: Gujarati (2004)

Como ejemplo introductorio, en la ilustración 4 se muestran dos casos, homocedástico y heterocedástico, donde se establece la diferencia a partir de un modelo de dos variables, en la que Y representa ahorros y X ingresos. En ambas situaciones, se observa que cuando los ingresos aumentan, los ahorros en promedio también lo hacen. Sin embargo, en la figura superior la varianza de los ahorros es uniforme en todos los niveles de ingreso (homocedasticidad), a diferencia de la figura inferior, donde las familias con mayores ingresos, en promedio, ahorran más que las familias con menos, por lo que aparece variabilidad en los ahorros (heterocedasticidad).

2.3.1. Causas de la heterocedasticidad

Existen diversas razones por las que las $Var(e_i)$ dejan de ser constantes, algunas de las cuales se indican a continuación:

a) **Seguir modelos de *error-learning***, por los que, mientras se aprende, los errores de comportamiento disminuyen con el tiempo. Por ese motivo, se espera que σ_i^2 disminuya (ilustración 5).

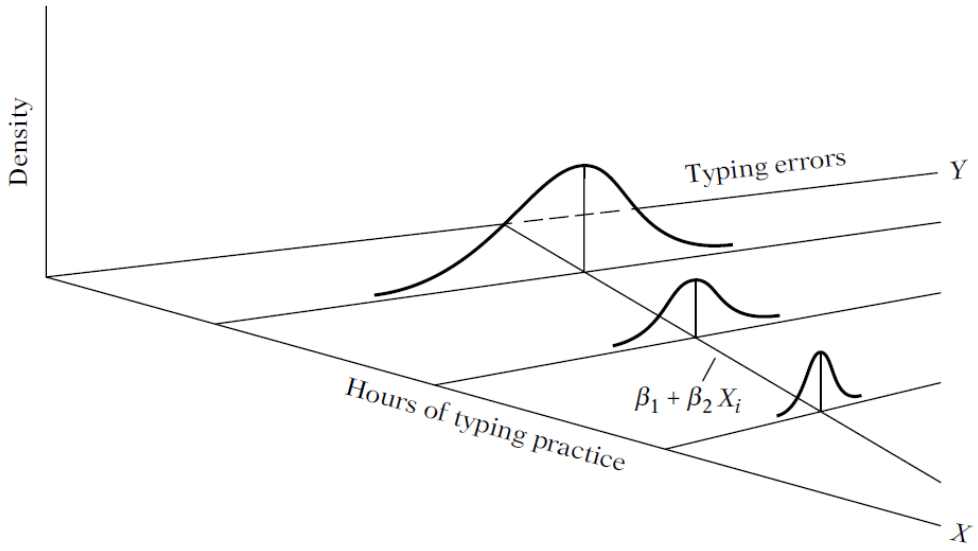


Ilustración 5. Disminución errores conforme se incrementa el tiempo de prácticas. Fuente: Gujarati (2004)

b) La **presencia de valores atípicos (*outliers***) puede ser una evidencia de que existe heterocedasticidad en los datos. Particularmente, si la muestra es pequeña puede alterar significativamente los resultados del análisis de regresión (ilustración 6).

c) La **asimetría en la distribución de uno o más regresores del modelo** puede ser otra fuente de heterocedasticidad.

d) La **incorrecta transformación de los datos**, como por ejemplo en las transformaciones de razón o de primeras diferencias.

e) La **aplicación de una incorrecta forma funcional**, en el caso de modelos lineales frente a modelos logarítmico-lineales, como es el caso de la regresión logística (próximo capítulo).

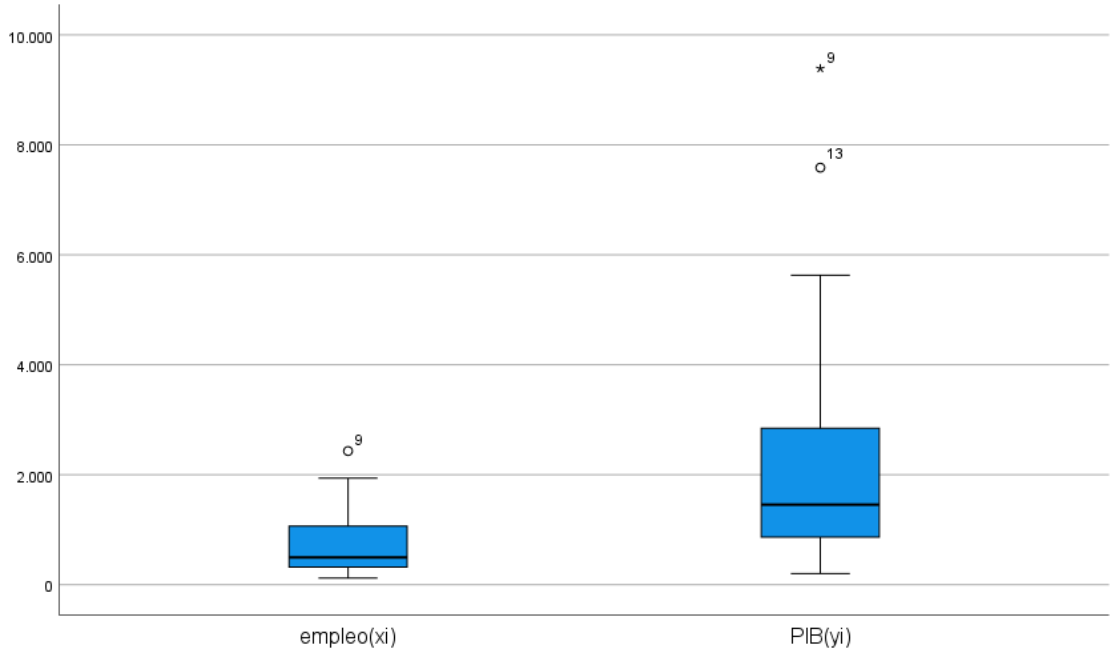


Ilustración 6. Box-plot con presencia de valores atípicos. Fuente: elaboración propia con SPSS

Por último, cabe destacar que se observa la heterocedasticidad más presente en **datos transversales** (donde los ítems suelen ser empresas, individuos, ciudades, etc.) cuyo comportamiento no es homogéneo, que en el caso de **datos de series temporales** (donde las variables tienden a ser de similar orden de magnitud).

Un ejemplo de heterocedasticidad en el análisis de datos transversales se ofrece en la ilustración 7 donde se aprecia una considerable variabilidad cuando la tasa de empleo supera los 1000 empleados.

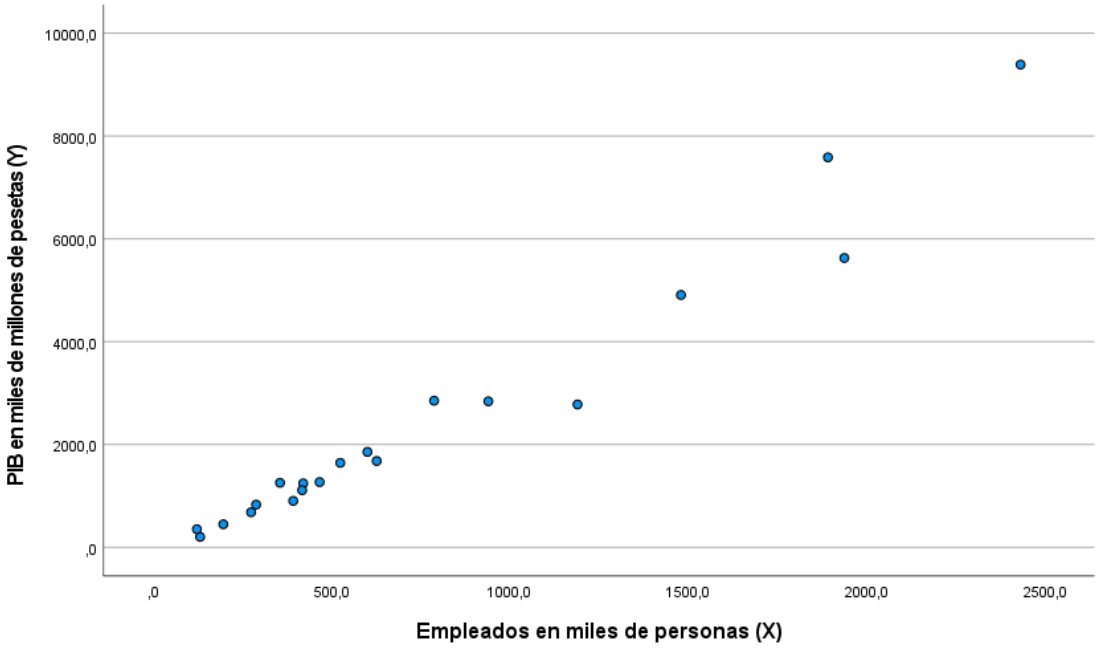


Ilustración 7. Tasa de empleo frente al Producto Interior Bruto (PIB) Fuente: elaboración propia con SPSS

2.3.2. Detección gráfica de la heterocedasticidad

En la detección de la heterocedasticidad es conveniente analizar la relación entre e_i^2 y la variable explicativa del modelo. Retomando el modelo de regresión lineal simple (2.23) y la igualdad (2.42) se tiene que:

$$e_i^2 = \alpha_0 + \alpha_1 x_i$$
$$E(\hat{e}_i^2) = \sigma_i^2. \tag{2.43}$$

El análisis preliminar de la relación entre e_i^2 y la variable explicativa se puede realizar mediante los métodos gráficos para a partir de ellos entender la intuición subyacente a los contrastes estadísticos más formales, Gallego (2008).

2.3.2.1 Gráficos de residuos

El gráfico de los residuos representa la dispersión de $\hat{e}_i$ o $\hat{e}_i^2$ constituyendo una herramienta visual para detectar la heterocedasticidad. Cuando se examina un gráfico de residuos, se han de formar grupos de observaciones y comprobar si la varianza individual resulta constante en cada grupo. En la ilustración 8 se muestra que la varianza local de los grupos 30-40 y 55-65 son superiores al resto. Tal apreciación, formulada

como hipótesis, podrá someterse entonces a un contraste estadístico de igualdad de varainzas.

Por otra parte, como primera aproximación, se puede realizar el análisis de la regresión representando $\hat{e}_i^2$ frente a $\hat{y}_i$, donde se obtienen patrones con los que se puede detectar sistemáticamente si los errores son crecientes o decrecientes (ilustración 9).

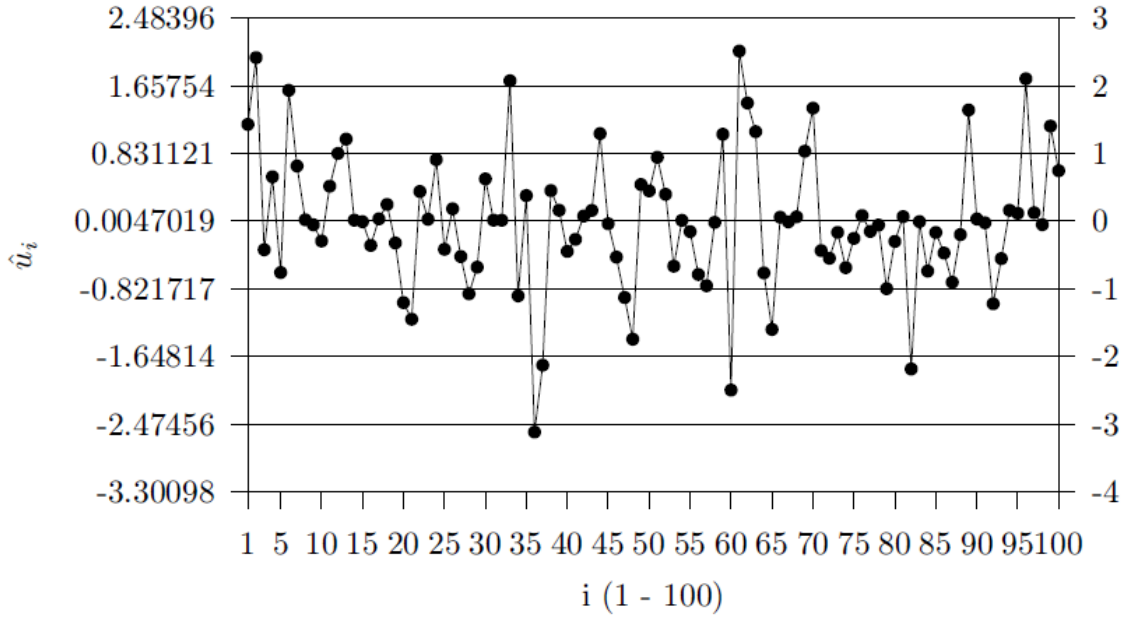


Ilustración 8. Gráfico de residuos por grupos. Fuente: Gallego (2008)

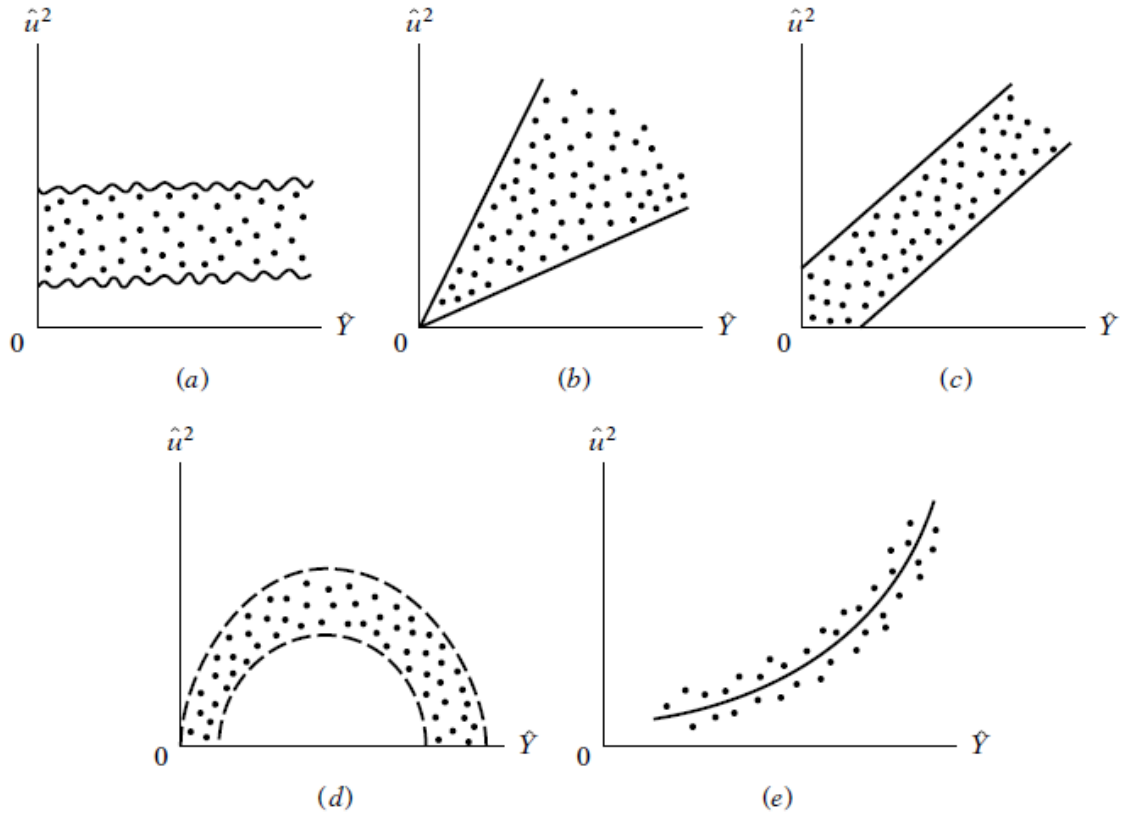


Ilustración 9. Patrones hipotéticos de errores residuales al cuadrado $(\hat{e}_i^2)$ frente a $\hat{y}_i$. Fuente: Gujarati (2004)

2.3.2.2 Gráficos de dispersión

Este tipo de gráficos se considera más útil que el anterior en el caso de que se sospeche que una variable explicativa pueda causar heterocedasticidad. Partiendo de un mismo conjunto de datos, en las ilustraciones 10 y 11 se observa que el valor de los residuos aumenta con el valor de la variable explicativa.

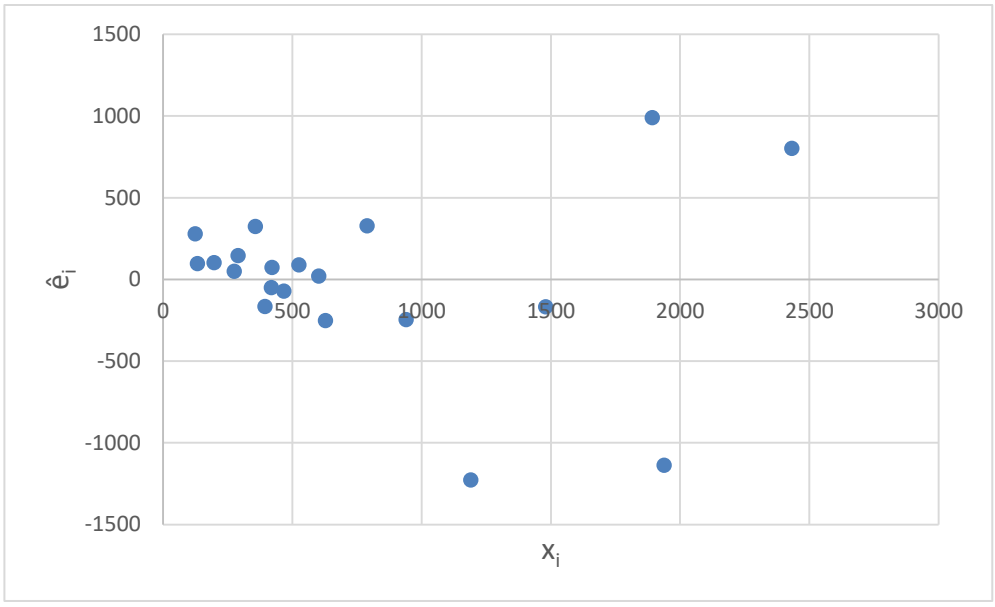


Ilustración 10. Diagrama de dispersión de x_i frente a $\hat{e}_i$. Fuente: elaboración propia con Excel.

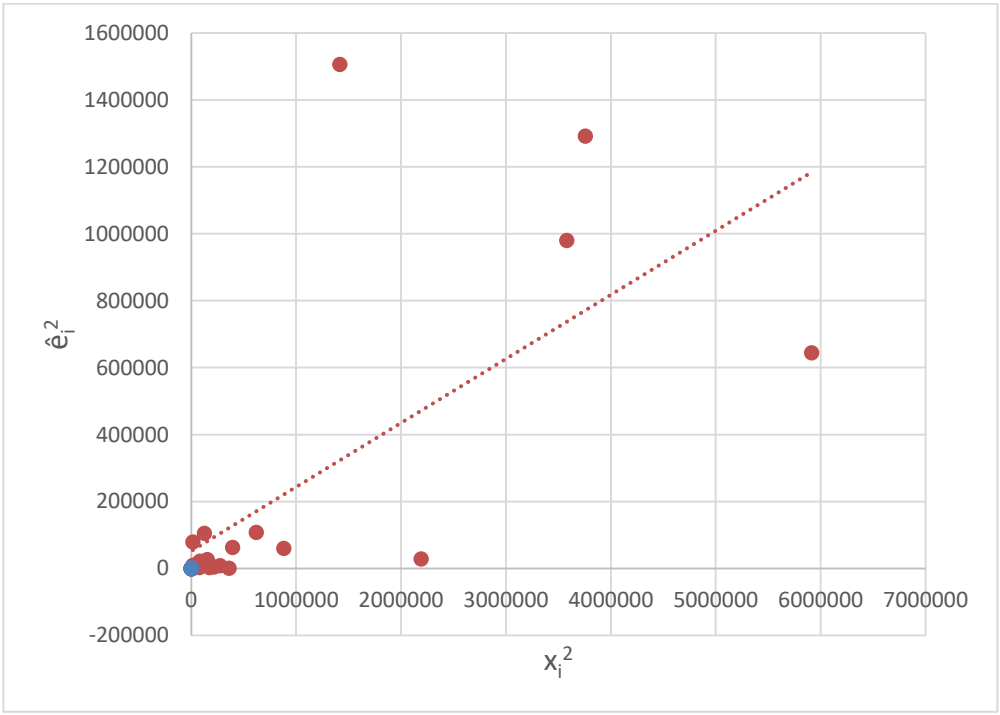


Ilustración 11. Diagramas de dispersión de x_i^2 frente a $\hat{e}_i^2$. Fuente: elaboración propia con Excel

Por otra parte, extraído de Rawlings et al. (1998) se representa en la ilustración 12 la estimación de la variable explicada frente a los residuos, ilustrando este tipo de gráficos un comportamiento característico de heterocedasticidad.

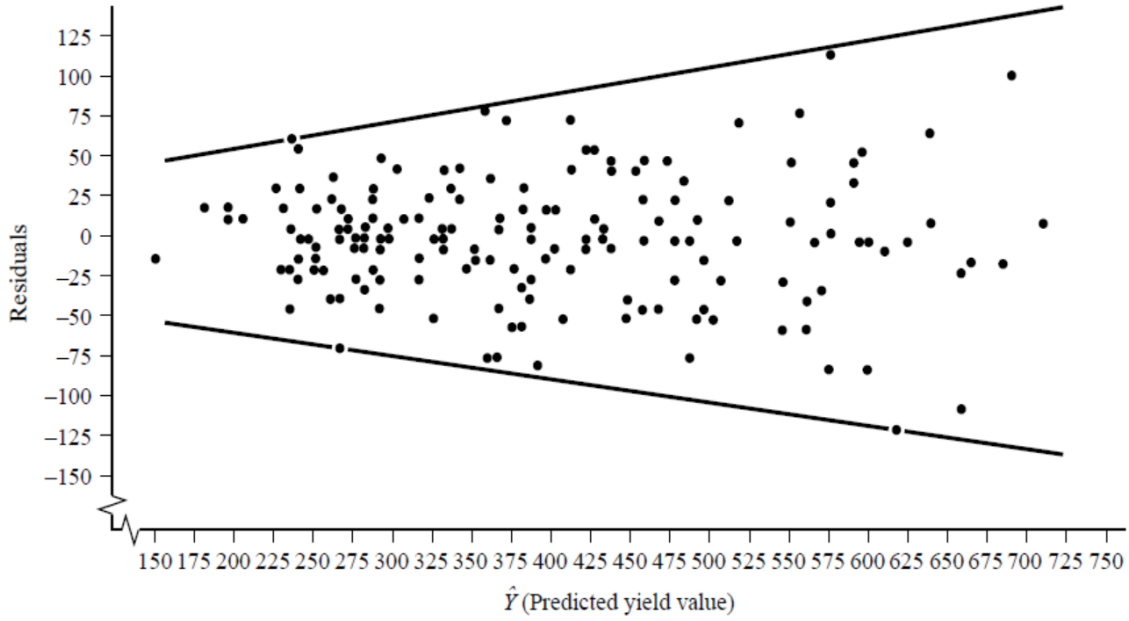


Ilustración 12. Diagramas de dispersión de $\widehat{Y}_i$ frente a $\widehat{e}_i$. Fuente: Rawlings et al. (1998).

2.4. Contrastes de heterocedasticidad

Para detectar la presencia de heterocedasticidad se proponen una serie de contrastes cuyo nexo de unión es que parten de la hipótesis nula de *ausencia de heterocedasticidad*. Sin embargo, no todos sugieren la misma forma funcional de la heterocedasticidad cuando se rechaza la hipótesis nula, Novales (1994)[10]. En cualquier caso, en este apartado se pretende únicamente reseñar los tests de contraste más relevantes sin entrar en detalle, para en un capítulo posterior, utilizarlos en un caso práctico.

2.4.1. El contraste de *Goldfeld* y *Quandt* (1965)

Este contraste se basa en el supuesto de que la varianza de la perturbación, σ_i^2, depende de una variable z_i , que normalmente es la variable explicativa, aunque no es preciso que lo sea. En cualquier caso debe ser una variable observable, por lo que se necesita disponer de información muestral de dicha variable.

[10] Conviene aclarar que no se incluye la demostración matemática formal de los contrastes de hipótesis, pudiéndose consultar la descripción de los fundamentos en Novales (1994), Uriel (2019) y Gujarati (2004).

Para contrastar la hipótesis nula de ausencia de heterocedasticidad

$$H_o: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2,$$
$$H_1: \sigma_i^2 = \sigma^2 g(z_i), \tag{2.44}$$

donde g se supone que es una función monótona creciente con z_i, por lo que se procede como sigue:

a) Se ordenan las observaciones de todas las variables del modelo en orden creciente de los valores de z_i.
b) Se divide la muestra en dos bloques de tamaño muestral n_1 y n_2 respectivamente, pudiendo dejar fuera p observaciones centrales para acentuar la independencia de los dos grupos. El número de observaciones de cada grupo tiene que ser de orden similar y mayor que el número de parámetros a estimar.
c) Se estima por MCO el modelo de regresión, por separado, para cada grupo de observaciones y se extrae la suma de cuadrados residual ($SSNEX$) de cada regresión
d) Dado que cada varianza residual (recordando 2.26) se distribuye según una chi-cuadrado de $n-2$ grados de libertad, el cociente de ambas distribuciones se aproxima a una distribución F de $Snédecor$, teniendo en cuenta que las varianzas son independientes.

   Finalmente, el estadístico para un nivel de significación α queda como

$$GQ = \frac{SSNEXn_2}{SSNEXn_1} \sim \frac{\chi^2_{n_2-2}}{\chi^2_{n_1-2}} \sim F_{n_2-2,\, n_1-2}. \tag{2.45}$$

   La interpretación del contraste sugiere que si existe homocedasticidad las varianzas tienen que ser iguales, pero si hay presencia de heterocedasticidad, con la ordenación propuesta, la varianza del residuo será mayor al final de la muestra por lo que debería aparecer que $SSNEXn_2 > SSNEXn_1$.

   Por tanto, cuanto más diverjan los $SSNEX$, mayor será el valor del estadístico y por ello conforme a Novales (1994) se rechazará H_0, a un nivel de significación α si

$$GQ > F_{n_2-2,\, n_1-2;\alpha}. \tag{2.46}$$

Observaciones a este contraste:

1. Si se sospecha que la varianza del residuo depende inversamente de los valores de z_i, entonces se debería ordenar la muestra en orden decreciente para cada variable, conforme al procedimiento reseñado.
2. El test, si bien es preciso, no muestra mucha potencia[11],[12] cuando las perturbaciones son heterocedásticas, aunque su varianza "promedio" en n_1 no sea demasiado diferente de n_2
3. La elección del valor de p es importante, si el valor de éste es demasiado pequeño no habrá independencia entre n_1 y n_2 y la homocedasticidad prevalece frente a la heterocedasticidad. En este sentido, Harvey y Phillips (1974) sugieren fijar p a 1/3 del total de la muestra Novales (1994).
4. Si se concluye que el residuo del modelo no presenta heterocedasticidad, pudiera deberse a una errónea especificación del parámetro σ_i^2, originada por una variable diferente a la que se ha supuesto. En consecuencia, el contraste debería llevarse a cabo repetidamente con variables, de las que se pueda sospechar a priori que puede depender la varianza del término error.

2.4.2. El contraste de *Breusch-Pagan-Godfrey* (1979)

El éxito del test *Goldfeld-Quandt* depende no solo del número de valores centrales que se suprimen, sino también en identificar la correcta variable explicativa con la que ordenar las observaciones. Esta limitación desaparece si se considera el test *Breusch-Pagan-Godfrey (BPG)* Gujarati (2004).

Considerando el modelo de regresión lineal de variable k

[11] Potencia de un contraste es la probabilidad de rechazar la hipótesis nula cuando es falsa, y también que la hipótesis nula en este caso es la de homocedasticidad, Novales (1994)

[12] Por ejemplo, muestra una alta probabilidad de aceptar H_o cuando es falso, afirma Kmenta (1986)

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki} + e_i{}^{13}, \tag{2.47}$$

suponiendo que la varianza del error en cada período se describe como

$$\sigma_i^2 = f(\delta_0 + \delta_1 Z_{1i} + \cdots + \delta_m Z_{mi}), \tag{2.48}$$

donde σ_i^2, es una función de variables $Z´s$ no estocásticas, por lo que concretamente se considera que σ_i^2 es una función lineal de las $Z´s$

$$\sigma_i^2 = \delta_0 + \delta_1 Z_{1i} + \cdots + \delta_m Z_{mi}. \tag{2.49}$$

Si $\delta_1 = \delta_2 = \cdots = \delta_m = 0$, entonces $\sigma_i^2 = \delta_0$ es una constante.

Para comprobar si σ_i^2 es homocedástico se plantea la hipótesis $\delta_1 = \delta_2 = \cdots = \delta_m = 0$.

Teniendo en cuenta lo anterior, el procedimiento a seguir es

a) Estimar el modelo por MCO y obtener los residuos $\hat{e}_0, \hat{e}_1, \ldots, \hat{e}_n$.
b) Obtener la serie de residuos normalizados al cuadrado:
   $\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{e}_i^2/n = SSNEX/n$.[14]
c) Construir variables p_i como sigue: $p_i = \hat{e}_i^2/\hat{\sigma}^2$
d) Estimar una regresión de $\hat{\sigma}^2$ sobre una constante y las variables $Z´s$

$$p_i = \delta_0 + \delta_1 Z_{1i} + \cdots + \delta_m Z_{mi} + \varepsilon_i, \tag{2.50}$$

   donde ε_i es el residuo de esta regresión.

e) Obtener la suma explicada $SSEX$ de la expresión (2.50), suponiendo una distribución normal para el término error, bajo la hipótesis de homocedasticidad y si el tamaño de la muestra se incrementa indefinidamente, entonces, la variable se distribuye como una chi-cuadrado de $m-1$ grados de libertad, y el estadístico para un nivel de significación α, se distribuye asintóticamente como

$$\mathrm{n}\hat{R}^2 \sim_{as} \chi_{m-1}^2 \,. \tag{2.51}$$

La interpretación del contraste BPG radica en que si los residuos no fueran heterocedásticos, entonces las variables $Z´s$ no mostrarían poder explicativo alguno sobre los residuos transformados p_i y por tanto, $SSEX$ debería ser pequeño. Si $SSEX/2$

[13] El modelo de regresión lineal múltiple se considera únicamente por conveniencia para la formalización teórica de este contraste.

[14] $\hat{\sigma}^2$ es el estimador de máxima verosimilitud de σ^2 bajo la hipótesis nula (homocedasticidad).

fuese mayor que el valor de chi-cuadrado tabulado al nivel de significación α, entonces se consideraría lo suficientemente alto y se rechazaría la hipótesis nula de ausencia de heterocedasticidad.
Observaciones a este contraste:

1. Si la heterocedasticidad implica que la varianza del término error crece con el tiempo, lo hará en valor absoluto.
2. Permite cierta flexibilidad, ya que no se precisa especificar la función que explica la dependencia de la magnitud de los residuos respecto a las $Z´s$. No importa la forma funcional de dicha dependencia.
3. La lista de variables Z debería ser corta e incluir pocas variables que no estén ya incluidas como variables explicativas del modelo original.

2.4.3. El contraste de *Koenker-Basset* (1979)

Al igual que el test BPG, el test de *Koenker-Basset (KB)* se basa en los residuos cuadráticos, $\hat{e}_i^2$, pero a diferencia de ser regresados en uno o más regresores, dichos residuos cuadráticos son ajustados en los valores estimados cuadráticos del regresando. Es decir, si el modelo original es de la forma

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki} + e_i, \tag{2.52}$$

al estimarse se obtiene $\hat{e}_i$, y a continuación se considera el modelo

$$\hat{e}_i^2 = \delta_0 + \delta_1 \hat{y}_i{}^2 + \varepsilon_i, \tag{2.53}$$

donde $\hat{y}_i$ son los valores estimados mediante (2.52). Si $\delta_1 = 0$, entonces se acepta H_0, por lo que se podría concluir que no existe heterocedasticidad. Esta hipótesis puede ser comprobada por el test F,

$$H_0: \delta_1 = 0$$
$$H_1: \delta_1 \neq 0 \tag{2.54}$$

- Se acepta H_0 si $F_{1,k} > F$
- Se rechaza H_0 si $F_{1,k} < F$

siendo $F_{1,k}$ el punto crítico de la distribución para los grados de libertad y significación dados.

2.4.4. El contraste de *White* (1980)

A diferencia del test *Goldfeld-Quandt,* el test de *White* no precisa especificar la forma que puede adoptar la heterocedasticidad, por lo que no depende del supuesto de la normalidad de acuerdo con Gujarati (2004).

Partiendo de un ejemplo ilustrativo, se considera un modelo de regresión de 3 variables

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + e_i, \tag{2.55}$$

El procedimiento a seguir es:

a) Emplear el modelo (2.55) por MCO y obtener los residuos, $\hat{e}_i$.
b) Aplicar una regresión auxiliar, $\hat{e}_i^2$, sobre una constante, los regresores del modelo original, sus cuadrados y productos cruzados de los regresores

$$\hat{e}_i^2 = \delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \delta_3 x_{1i}^2 + \delta_4 x_{2i}^2 + \delta_5 x_{1i} x_{2i} + \varepsilon_i, \tag{2.56}$$

y obtener $\hat{r}^2$ de esta regresión.
c) Bajo la hipótesis cero (no heterocedasticidad) en (2.56) el estadístico nR^2 se distribuye asintóticamente conforme a una distribución chi-cuadrado cuyos grados de libertad (gl) son igual al número de regresores (menos el término constante) de la regresión auxiliar

$$\mathrm{n}R^2 \sim_{as} \chi_{gl}^2 \,. \tag{2.57}$$

En este caso, como hay 5 regresores en la regresión auxiliar, la distribución tiene 5 gl .
d) En el caso de que el valor de chi-cuadrado resultante de (2.57) excediera el valor chi-cuadrado crítico, al nivel de significación α elegido, existe heterocedasticidad. En caso contrario, no hay heterocedasticidad, lo que equivale a afirmar que en la regresión auxiliar (2.56),

$$\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = 0. \tag{2.58}$$

Observaciones a este contraste:

1. Dado que el tamaño muestral crece con el número de observaciones, R^2 tenderá a cero bajo la hipótesis nula de ausencia de heterocedasticidad. En el caso en que la varianza del término error depende de las variables explicativas del modelo, R^2 no tenderá cero.
2. En d) se asume que el error de la varianza de e_i, σ_i^2, está relacionada funcionalmente con los regresores, sus cuadrados, y sus productos cruzados. Si todas los coeficientes de pendiente parciales en esta

regresión son simultáneamente igual a cero, entonces el error de la varianza es la constante homocedástica igual a δ_0.

3. Si un modelo dispone de muchos regresores, al introducir todos los regresores, sus términos cuadráticos, y sus productos cruzados, se incrementan muy rápido los grados de libertad con el número de variables.

Capítulo 3. El método de Mínimos Cuadrados Ponderados

3.1. Introducción

Cuando en presencia de heterocedasticidad en el modelo de regresión se asume la homocedasticidad como aproximación, los estimadores, permanecen insesgados, pero se vuelven ineficientes debido a que no ofrecen una varianza mínima.

En el supuesto de que se disponga de una estimación de las varianzas del caso heterocedástico, el método de mínimos cuadrados ponderados, es capaz de asegurar la homocedasticidad de los errores gracias a una transformación del modelo de regresión, que permite la aplicación del método MCO.

3.1.1. Propiedades de los estimadores de MCO bajo heterocedasticidad

En la ecuación (2.42) se tiene la condición heterocedástica de la perturbación de la regresión, lo que implica que la varianza de la perturbación puede variar de una observación a otra. El efecto de dicha varianza afecta a las propiedades de los estimadores de mínimos cuadrados de los coeficientes de regresión.

Siguiendo lo establecido en (2.34) el estimador de mínimos cuadrados de α_1 es

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n} x'_i y'_i}{\sum_{i=1}^{n} x'^2_i} = \alpha_1 + \frac{\sum_{i=1}^{n} x'_i e_i}{\sum_{i=1}^{n} x'^2_i}, \tag{3.1}$$

siendo $x'_i = (x_i - \bar{x})$ e $y'_i = (y_i - \bar{y})$ respectivamente.

A continuación, aplicando la ecuación (2.41) donde en situación heterocedástica $E(e_i) = 0$, se tiene

$$E(\hat{\alpha}_1) = \alpha_1 + E\left(\frac{\sum_{i=1}^{n} x'_i e_i}{\sum_{i=1}^{n} x'^2_i}\right) = \alpha_1. \tag{3.2}$$

De forma similar, a partir de

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \bar{x} = (\alpha_0 + \alpha_1 \bar{x} + \bar{e}\,) - \hat{\alpha}_1 \bar{x}, \tag{3.3}$$

se obtiene que

$$E(\hat{\alpha}_0) = \alpha_0 + \alpha_1 \bar{x} + E(\bar{e}) - E(\hat{\alpha}_1)\bar{x} = \alpha_0. \tag{3.4}$$

Por lo que se demuestra que la centralidad de los estimadores de mínimos cuadrados o MCO no se altera en condiciones de heterocedasticidad.

Por otra parte, aplicando ahora

$$Var(\hat{\alpha}_1) = E(\hat{\alpha}_1 - \alpha_1)^2 = E\left(\frac{\sum_{i=1}^{n} x'_i e_i}{\sum_{i=1}^{n} x'^2_i}\right)^2, \tag{3.5}$$

y teniendo en cuenta que $E({e_i}^2) = \sigma_i^2$ y $E(e_i, e_j) = 0$, se obtiene que

$$Var(\hat{\alpha}_1) = \frac{\sum_{i=1}^{n} x'^2_i \, \sigma_i^2}{(\sum_{i=1}^{n} x'^2_i)^2}. \tag{3.6}$$

De igual forma, partiendo de (3.4) y procediendo como en (3.5) se aplica

$$Var(\hat{\alpha}_0) = E(\hat{\alpha}_0 - \alpha_0)^2 = E\,(\bar{x}(\alpha_1 - \hat{\alpha}_1) + \bar{e})^2 = \bar{x}^2 E\left(\frac{\sum_{i=1}^{n} x'_i e_i}{\sum_{i=1}^{n} x'^2_i}\right)^2, \tag{3.7}$$

y asumiendo la misma condición que en (3.6) para el estimador del intercepto se tiene

$$Var(\hat{\alpha}_0) = E(\hat{\alpha}_0 - \alpha_0)^2 = \bar{x}^2 \frac{\sum_{i=1}^{n} x'^2_i \, \sigma_i^2}{(\sum_{i=1}^{n} x'^2_i)^2}, \tag{3.8}$$

por lo que **se obtiene que la varianza de los estimadores de MCO deja de ser insesgada, respecto a la varianza poblacional, bajo la hipótesis de heterocedasticidad**. Por ello, se debe recurrir al método MCP para transformar el modelo mediante la asignación de pesos, con el fin de poder aplicar el método de MCO y obtener los estimadores de la ecuación de regresión.

Conviene añadir también que en el caso homocedástico se recupera (2.38), aunque ahora este estimador no es insesgado. Por tanto para demostrar $E(s^2) \neq \sigma^2$ en presencia de heterocedasticidad se parte de un desarrollo similar al anterior que se puede consultar en Gujarati (2004), donde se demuestra que bajo homocedasticidad, $\sigma_i^2 = \sigma^2$, para cada i. Por tanto, el valor esperado del valor convencional de $s^2 = \sum_{i=1}^{n} \hat{e}_i^2 /(n-2)$ no será igual a σ^2 en presencia de heterocedasticidad.

En resumen:

- No se altera la centralidad de los coeficientes de regresión de MCO bajo heterocedasticidad.
- Bajo heterocedasticidad, la varianza de $\hat{\alpha}_1$ y de $\hat{\alpha}_0$ se vuelve sesgada respecto a la varianza poblacional, ya no son los mismos estimadores de MCO que cumplen esas condiciones, con lo que los estimadores de mínimos cuadrados ($\hat{\alpha}_1$, $\hat{\alpha}_0$) dejan de ser eficientes en ese contexto.

3.2. El método de mínimos cuadrados ponderados

Como se advirtió en el apartado anterior, bajo heterocedasticidad además de la ineficiencia de los estimadores (en el sentido de varianza mínima), la asignación de distribuciones asociadas al muestreo del capítulo de inferencia no es válida. Por lo tanto, lo más adecuado es tratar de trasladar el problema al caso homocedástico para disponer de dichas herramientas, restringiendo el estudio a la regresión lineal simple.

Mediante el método MCO, se minimizaba la expresión

$$V_r(\alpha_0, \alpha_1) = \sum_{i=1}^{n} [y_i - (\alpha_0 + \alpha_1 x_i)]^2 = \sum_{i=1}^{n} e_i^2 \tag{3.9}$$

por contra, con MCP, se minimiza la ecuación (3.9) suponiendo que σ_i^2 es conocida, se realizando la siguiente transformación

$$V_r(\alpha_0, \alpha_1) = \sum_{i=1}^{n} w_i [y_i - (\alpha_0 + \alpha_1 x_i)]^2 = \sum_{i=1}^{n} w_i e_i^2, \tag{3.10}$$

$$\text{donde } w_i = 1/\sigma_i^2.$$

Es decir, se va a minimizar una suma ponderada de residuos cuadráticos donde $w_i = 1/\sigma_i^2$, representa los pesos para cada observación, frente a MCO que se minimizaba una suma de residuos cuadráticos de pesos uniformes. Por MCP, el peso asignado a cada observación es inversamente proporcional a su σ_i, por lo que si las observaciones son

más precisas, proporcionalmente aportarán más peso al minimizar $V_r(\alpha_0, \alpha_1)$ Peña (1989). Por tanto, dado que se minimiza una suma ponderada de residuos cuadráticos este método se denomina de mínimos cuadrados ponderados siendo un caso especial de la técnica de estimación general (MCG)[15]. Por ello, la técnica de MCP es generalizable solo a otras situaciones con menos restricciones, como correlación y heterogeneidad de errores.

Partiendo del modelo de regresión lineal simple heterocedástica

$$y_i = \alpha_0 + \alpha_1 x_i + e_i, \tag{3.11}$$

a continuación, se busca una transformación del modelo que consiga nuevos errores aleatorios distribuidos con varianza constante. Para ello, se divide por σ_i la ecuación de (3.11), con lo que se permite la estimación con MCO

$$\frac{y_i}{\sigma_i} = \alpha_0 \left(\frac{1}{\sigma_i}\right) + \alpha_1 \left(\frac{x_i}{\sigma_i}\right) + \frac{e_i}{\sigma_i},$$

$$y_i^* = \alpha_0 w_i^* + \alpha_1 x_i^* + e_i^*, \tag{3.12}$$

donde $y_i^* = \frac{y_i}{\sigma_i}$, $w_i^* = \frac{1}{\sigma_i}$, $x_i^* = \frac{x_i}{\sigma_i}$ y $e_i^* = \frac{e_i}{\sigma_i}$.

Desde el principio se comprueba que el proceso equivalente es redefinir la varianza residual con los pesos $w_i = 1/\sigma_i^2$, por lo que se plantea ahora la construcción de los estimadores ELIO [16]($\tilde{\alpha}_1$, $\tilde{\alpha}_0$). Dado que la varianza ahora es constante, se reúnen las condiciones de igualdad entre los estimadores ELIO y los de MCO ($\hat{\alpha}_1$, $\hat{\alpha}_0$). Minimizando conforme a (2.5) la expresión

$$V_r\ (\alpha_0, \alpha_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i^* - \alpha_0 w_i^* - \alpha_1 x_i^*)^2, \tag{3.13}$$

[15] En rigor, la técnica de mínimos cuadrados generalizados (MCG) es un método más general que el MCP, ya que permite errores correlacionados. En el contexto de heterocedasticidad, se pueden adoptar los términos MCP y MCG indistintamente Gujarati (2004).

[16] El método de estimación ELIO (Estimador Lineal Insesgado Óptimo) o *BLUE* (*Best Linear Unbiased Estimator*) requiere que el estimador sea una combinación lineal insesgada de observaciones muestrales y que su varianza sea la menor de cualquier otro estimador lineal insesgado. Se empleará este método para la obtención del estimador ELIO de α, denotado, $\tilde{\alpha}$, conforme a Kmenta (1986).

se obtienen las ecuaciones normales de mínimos cuadrados.

$$\begin{aligned}\sum_{i=1}^{n} w_i^* \, y_i^* &= \tilde{\alpha}_0 \sum_{i=1}^{n} w_i^{*2} + \tilde{\alpha}_1 \sum_{i=1}^{n} w_i^* \, x_i^*, \\ \sum_{i=1}^{n} x_i^* \, y_i^* &= \tilde{\alpha}_0 \sum_{i=1}^{n} w_i^* x_i^* + \tilde{\alpha}_1 \sum_{i=1}^{n} x_i^{*2}.\end{aligned} \tag{3.14}$$

Por aligerar la notación de estas ecuaciones se eliminan los asteriscos

$$\begin{aligned}\sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} &= \tilde{\alpha}_0 \sum_{i=1}^{n} \frac{1}{\sigma_i^2} + \tilde{\alpha}_1 \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}, \\ \sum_{i=1}^{n} \frac{x_i y_i}{\sigma_i^2} &= \tilde{\alpha}_0 \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} + \tilde{\alpha}_1 \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2}.\end{aligned} \tag{3.15}$$

Introduciendo el cambio de variable $w_i = 1/\sigma_i^2$ se transforman las ecuaciones de (3.16) en las siguientes expresiones

$$\begin{aligned}\sum_{i=1}^{n} w_i y_i &= \tilde{\alpha}_0 \sum_{i=1}^{n} w_i + \tilde{\alpha}_1 \sum_{i=1}^{n} w_i \, x_i, \\ \sum_{i=1}^{n} w_i x_i y_i &= \tilde{\alpha}_0 \sum_{i=1}^{n} w_i x_i + \tilde{\alpha}_1 \sum_{i=1}^{n} w_i x_i^2.\end{aligned} \tag{3.16}$$

Resolviendo estas ecuaciones se obtienen los estimadores MCP de los coeficientes de regresión, $\alpha_0 \; y \; \alpha_1$,

$$\begin{aligned}\tilde{\alpha}_1 &= \frac{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i y_i - \sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i^2 - (\sum_{i=1}^{n} w_i x_i)^2} = \\ \tilde{\alpha}_1 &= \frac{\sum_{i=1}^{n} w_i (x_i - \tilde{x})(y_i - \tilde{y})}{\sum_{i=1}^{n} w_i (x_i - \tilde{x})^2}, \\ \tilde{\alpha}_0 &= \tilde{y} - \frac{\sum_{i=1}^{n} w_i (x_i - \tilde{x})(y_i - \tilde{y})}{\sum_{i=1}^{n} w_i (x_i - \tilde{x})^2} \tilde{x} = \\ \tilde{\alpha}_0 &= \frac{\tilde{y} \sum_{i=1}^{n} w_i (x_i - \tilde{x})^2 - \tilde{x} \sum_{i=1}^{n} w_i (x_i - \tilde{x})(y_i - \tilde{y})}{\sum_{i=1}^{n} w_i (x_i - \tilde{x})^2},\end{aligned} \tag{3.17}$$

donde $\tilde{\alpha}_0 = \tilde{y} - \tilde{\alpha}_1 \tilde{x}$, siendo $\tilde{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$ e $\tilde{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$.

Estas expresiones de los mejores estimadores insesgados lineales (ELIO) de $\alpha_0\ y\ \alpha_1$, y son diferentes a las propias de mínimos cuadrados. Por tanto, se concluye que los estimadores de mínimos cuadrados de los coeficientes de regresión no son ELIO bajo el supuesto de heterocedasticidad. Además, se deduce que los estimadores de mínimos cuadrados no obtienen la mínima varianza entre todos los estimadores insesgados y, por ende, no son eficientes.

Las varianzas de los estimadores bajo heterocedasticidad quedan entonces como

$$\begin{aligned} Var(\tilde{\alpha}_1) &= \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i^2 - (\sum_{i=1}^{n} w_i x_i)^2} = \frac{1}{\sum_{i=1}^{n} w_i (x_i - \tilde{x})^2} \\ Var(\tilde{\alpha}_0) &= \frac{\sum_{i=1}^{n} w_i x_i^2}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i^2 - (\sum_{i=1}^{n} w_i x_i)^2} = \frac{1}{\sum_{i=1}^{n} w_i} + \frac{\tilde{x}^2}{\sum_{i=1}^{n} w_i (x_i - \tilde{x})^2}. \end{aligned} \tag{3.18}$$

Reseñar que si $w_i = w = 1/\sigma^2$ para todo i , entonces las expresiones de (3.12) y (3.18) son las mismas que las dadas en (2.35) y (2.36) para el modelo homocedástico.

Como se ha visto, la heterocedasticidad no deshace las propiedades de consistencia e insesgadez de los estimadores de MCO, pero provoca que éstos ya no sean eficientes, (por ejemplo en muestras grandes). Para corregir esta falta de eficiencia existen medidas correctoras de la heterocedasticidad, para lo cual se podría plantear, además del supuesto ya visto en (3.10) de que σ_i^2 es conocida, la hipótesis de que σ_i^2 es desconocida.

En el caso de que σ_i^2 no sea conocida y no se disponga certeza del origen de la heterocedasticidad, no cabe más remedio que confiar en los datos de la muestra y estimar las varianzas de la perturbación a partir de los datos.

Un primer método, indica Kmenta (1986), consistiría en obtener los estimadores a partir de la función de máxima verosimilitud[17], pero la solución precisaría de un procedimiento iterativo laborioso hasta que los valores de las estimaciones converjan, es decir, hasta

[17] La teoría de los estimadores de máxima verosimilitud puede consultarse en la bibliografía general de Estadística. En las referencias de este TFM se puede acudir a Peña (1989) o Rawlings et al. (1998)

que las diferencias entre las sucesivas series de estimaciones se consideren despreciables[18].

Un segundo método alternativo consiste en estimar σ_i^2 mediante la varianza muestral

$$s_i^2 = \frac{1}{n}\sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_i)^2}{(n_i - 1)}, \tag{3.19}$$

donde $\overline{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$.

Se puede apreciar que s_i^2 es un estimador consistente de σ_i^2. Si se sustituye w_i por $1/s_i^2$ en (3.17), se obtienen las estimaciones de los coeficientes de regresión; y efectuando la misma sustitución en (3.18), se consigue la estimación de sus respectivos errores estándar. Los estimadores resultantes tienen las mismas propiedades asintóticas que los estimadores de máxima verosimilitud.

En resumen:

1. Si no se mantiene la homocedasticidad, los estimadores MCO de los coeficientes de regresión no son eficientes.
2. La varianza de MCO ya no es mínima (óptima), no se puede aplicar el Teorema de Gauss-Markov y los estimadores no son eficientes.
3. La varianza de los estimadores es sesgada, y por ello, inconsistente[19]
4. La varianza muestral se convierte en un estimador consistente de la varianza poblacional cuando ésta se desconoce.

3.2.1. Ejemplo comparativo σ_i^2 conocida y desconocida con MCP

De la mano de Kmenta (1986) se pretende demostrar que si se desconoce la varianza de cada una de las observaciones, es posible estimarla mediante la varianza muestral conforme a la expresión (3.19). Para ello se dispone de los datos de la tabla 1 donde se representan los ingresos y gastos en ropa mensuales de 20 familias,

[18] La prueba de este procedimiento se puede consultar en el paper de W. Oberhofer and J.Kmenta, "A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models", *Econometrica*, 42 (May 1974), pp.579-590.

[19] Un estimador es consistente, si mantiene un sesgo nulo cuando aumenta el tamaño muestral indefinidamente. Por tanto, si $n \to \infty$ la función de densidad del estimador converge al valor del parámetro, es decir, $p \lim_{n\to\infty} \hat{\alpha}_1 = \alpha_1$ ($\hat{\alpha}_1$converge en probabilidad a α_1)

X=Ingresos (€)	Número de familias (i)	Y=Gasto en ropa (€)
2000	8	160, 160, 180, 200, 210, 220, 230, 250
4000	7	200, 220, 230, 300, 310, 340, 350
6000	5	300, 300, 400, 450, 540

Tabla 1. Datos anuales ingresos VS gastos en ropa de 20 familias. Fuente: Kmenta (1986)

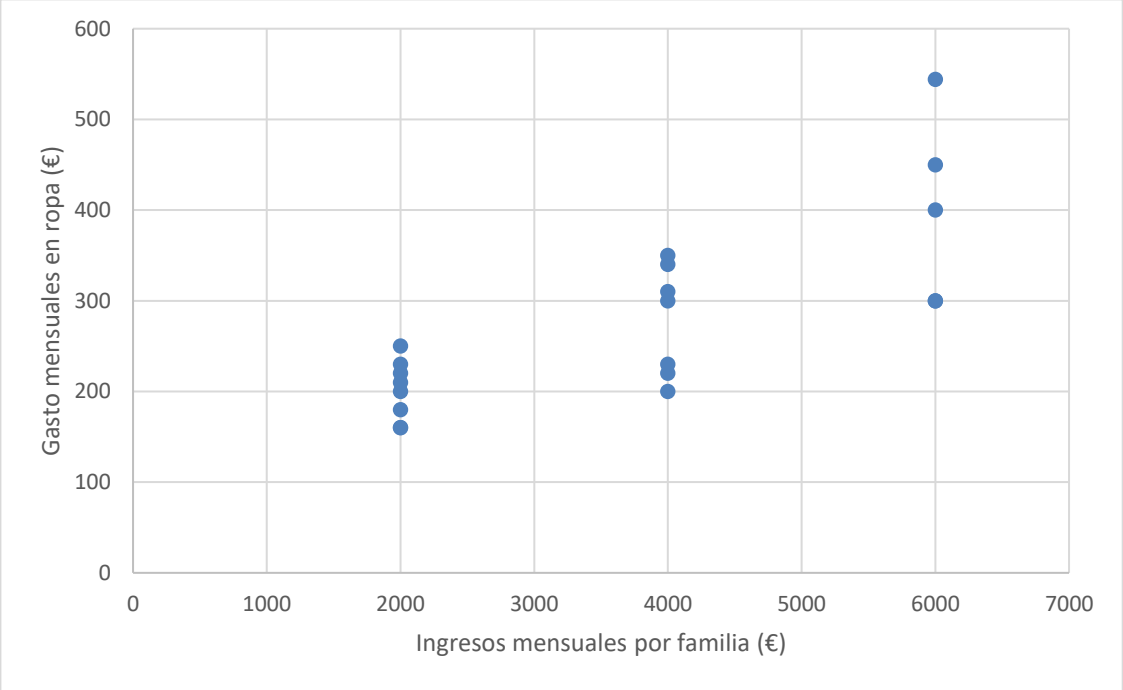


Ilustración 13. Datos mensuales de Ingresos por familia vs Gastos en ropa. Fuente: Kmenta (1986)

A priori, se va a considerar que el modelo de partida para las variables X e Y, representadas en la ilustración 13, se corresponde con el de la ecuación (3.11).

Por un lado, se hallan los estimadores de los coeficientes de regresión por **MCO**

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.04849$$

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \bar{x} = 98.28685$$

$$\boldsymbol{y_i = 98.28685 + 0.04849 x_i,}$$

siendo $\bar{x} = 3700$ *e* $\bar{y} = 277.7$.

Asumiendo que el modelo es heterocedástico, los estimadores de mínimos cuadrados son centrados pero no asintóticamente eficientes.

Si ahora se considera, desde un punto de vista comparativo, que $\sigma_i^2 = x_i^2$, la ecuación de regresión (3.12) pasa a ser homocedástica dividiendo ambos miembros por $\sigma_i = x_i$, que aplicando **MCP** a los resultados son

$$\frac{y_i}{x_i} = 0.04512 + \frac{109.73537}{x_i},$$

que revertiendo a la ecuación de regresión original, se obtiene

$$\boldsymbol{y_i = 109.73537 + 0.04512 x_i},$$

donde $\tilde{\alpha}_1 = 0.04512$ y $\tilde{\alpha}_0 = 109.73537$

Por último, hallando la varianza de los estimadores bajo heterocedasticidad se comprueba que

$$Var(\hat{\alpha}_1) = \frac{\sum_{i=1}^{n} x'^2_i \, \sigma_i^2}{(\sum_{i=1}^{n} x'^2_i)^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \, x_i^2}{(\sum_{i=1}^{n} (x_i - \bar{x})^2)^2} = \frac{2.03 \times 10^{14}}{(8.27 \times 10^6)^2} = \mathbf{2.97}.$$

$$Var(\tilde{\alpha}_1) = \frac{1}{\sum_{i=1}^{n} w_i (x_i - \tilde{x})^2} = \frac{1}{\sum_{i=1}^{n} \frac{(x_i - \tilde{x})^2}{x_i^2}} = \frac{1}{0.875} = \mathbf{1.14}.$$

que comparando los estimadores y las varianzas de ambos se tiene que

$$\frac{\hat{\alpha}_1}{\tilde{\alpha}_1} = \frac{0.048}{0.045} = \mathbf{1.07 \sim 1},$$

$$\frac{Var(\hat{\alpha}_1)}{Var(\tilde{\alpha}_1)} = \frac{2.97}{1.14} = \mathbf{2.62 \sim 3},$$

con lo que la eficiencia de ambos estimadores es aproximadamente del mismo orden y la varianza del estimador de α_1 es casi tres veces la del estimador ELIO de α_1.

Representando en la ilustración 14 las regresiones de MCO (rojo) y MCP (azul)

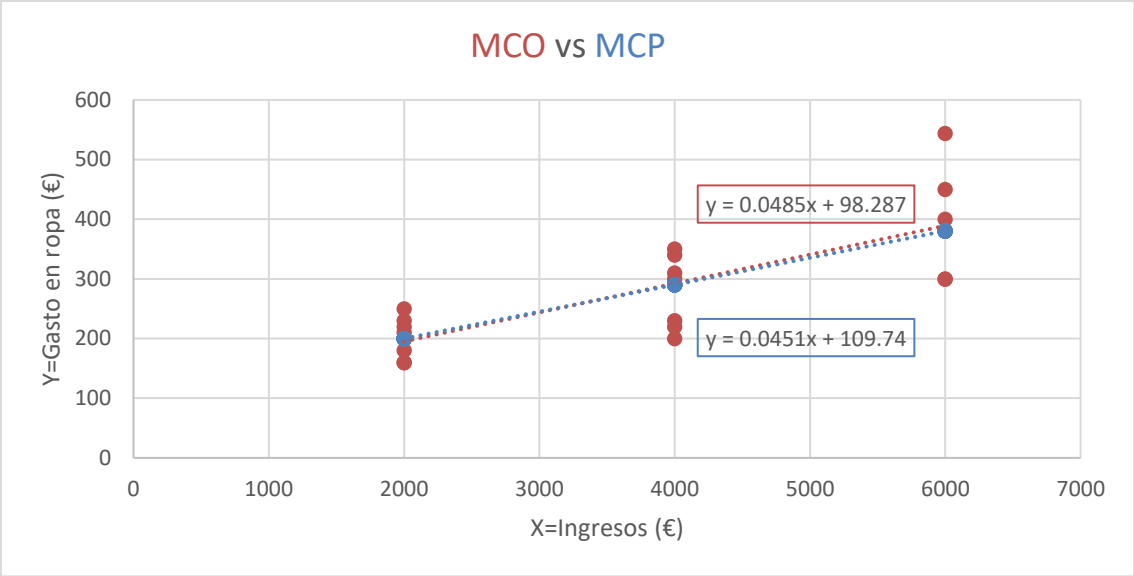


Ilustración 14. Comparativa rectas de regresión MCO y MCP. Fuente: Kmenta (1986)

Finalmente, si desconocemos σ_i^2 , a partir de los datos de la tabla 1 se estiman las varianzas con la varianza muestral de (3.19), ya que se demuestra que s_i^2es un estimador consistente de σ_i^2, con lo que considerando dicha expresión se estiman las varianzas de las perturbaciones obteniéndose para cada familia.

$$\boldsymbol{s_1^2 = 1069.64286},$$

$$\boldsymbol{s_2^2 = 3714.28571},$$

$$\boldsymbol{s_3^2 = 10520}.$$

Aplicando las fórmulas de **MCP** de (3.13), con w_i sustituido con $(1/s_i^2)$ se obtiene $\hat{\alpha}_1 = 0.0446$ y $\hat{\alpha}_0 = 111.02$, con lo que resulta que la ecuación de regresión queda como

$$\boldsymbol{y_{ij} = 111.02370 + 0.04457x_i}.$$

En este caso, se observa que la estimación de $\sigma_i^2 = 1/s_i^2$ arroja resultados similares. En este caso, a aquellos obtenidos bajo la suposición de que $\sigma_i^2 = x_i^2$. Al ser un ejemplo meramente ilustrativo, las varianzas de los estimadores no se han obtenido en este caso.

3.2.2. El método de MCP con el modelo de regresión logística simple

La regresión logística da respuesta a variables de carácter categórico o cualitativo, que mediante la regresión lineal simple no es viable, Baenas (2022). El modelo de regresión logística pertenece al conjunto de modelos generalizados, ofreciendo una respuesta

cualitativa que según el caso puede ser de tipo binario[20]. Sin atender a las características de modelado, la función asociada al modelo de regresión logística viene dada por la transformación logística de la probabilidad o *logit.*

$$L(p_i) = log\frac{p_i}{1-p_i} = \alpha_0 + \alpha_1 x_i + e_i, \tag{3.20}$$

donde e_i es la variable aleatoria de error y p_i la probabilidad de ocurrencia o éxito de un suceso.

Para la estimación de parámetros en el modelo logístico se recurre a la estimación de la probabilidad de éxito del suceso para cada valor de x_i, dada por la proporción

$$\hat{p}_i = \frac{n_i}{N_i}. \tag{3.21}$$

Puesto que esta estimación de la probabilidad precisa de un valor N_i considerable, la variable $\hat{p}_i$ se distribuye asintóticamente conforme a una distribución normal y se demuestra que $L(\hat{p}_i)$ se distribuirá según una normal con media $L(p_i)$ y desviación típica obtenida a partir de

$$\sigma[L(\hat{p}_i)] \cong \sigma(\hat{p}_i)\left(\frac{dL(\hat{p}_i)}{\hat{p}_i}\right)_{p_i} = \sqrt{\frac{1}{N_i p_i(1-p_i)}}, \tag{3.22}$$

por lo que teniendo presente el modelo de regresión logística se tiene que

$$\sigma_i^2 = Var(e_i) = Var[L(\hat{p}_i)] = \frac{1}{N_i p_i(1-p_i)}, \tag{3.23}$$

y dado que el valor de σ_i se desconoce, ésta se puede estimar mediante s_i

$$s_i^2 = \frac{1}{N_i \hat{p}_i(1-\hat{p}_i)}. \tag{3.24}$$

El modelo de regresión (3.23) es heterocedástico por construcción del propio modelo de regresión, por lo que si se asume la homocedasticidad como aproximación, se desprende que los estimadores resultarán insesgados e ineficientes. Por ello, al estimarse la varianza por (3.24), mediante el método de MCP se garantiza la

[20] Las variables dicotómicas o binarias sólo admiten dos valores, denotados por 0 y 1, para categorías mutuamente excluyentes, tales como verdadero-falso, sí-no, éxito-fracaso, etc. Según el texto a este tipo de variables se les denomina *dummy*, de *diseño* o *indicadoras* conforme a Baenas (2022).

homocedasticidad de los errores y en consecuencia la aplicación del método MCO. Para el modelo logístico simple se sintetiza como sigue:

a) Para cada observación x_i de la muestra, se obtiene la estimación de la probabilidad $\hat{p}_i$, y el modelo de regresión

$$L(\hat{p}_i) = log\frac{\hat{p}_i}{1-\hat{p}_i} = log\frac{n_i}{N_i - n_i}, con\ N_i \geq 5 \tag{3.25}$$

b) Introduciendo los pesos $\sqrt{w_i} = 1/s_i$ como una variable en el modelo binario de regresión logística (3.20), éste se convierte en

$$\sqrt{w_i}L(p_i) = \sqrt{w_i}\alpha_0 + \sqrt{w_i}\alpha_1 x_i + \sqrt{w_i}e_i, \tag{3.26}$$

por lo que sustituyendo convenientemente $\sqrt{w_i}$ en la ecuación (3.26) el modelo se transforma en

$$L^*(p_i) = \alpha_0\sqrt{w_i} + \alpha_1 x_i^* + e_i^*. \tag{3.27}$$

c) Una vez transformada la heterocedasticidad de los errores en un modelo homocedástico, se aplica MCO en (3.27) teniendo en cuenta que el proceso conlleva una regresión múltiple de variables regresoras, x_i^* y $\sqrt{w_i}$, y la aparición del intercepto nulo.

EJEMPLO

Partiendo de la tabla 2 de valores observados de cierta variable x_i, distribuida por clases de N_i observaciones y n_i, ocurrencias de un suceso y:

x_i	N_i	n_i
7	50	9
9	60	13
11	70	19
14	90	29
16	110	46
21	80	37
26	75	40
31	60	34
36	50	31
41	35	21

Tabla 2. Valores observados regresión logística. Fuente: elaboración propia

Seguidamente, se muestra el cálculo de las variables $\sqrt{w_i}$, x_i^* y $L^*(p_i)$ efectuado con $Excel$ cuyos valores de regresión obtenidos aparecen en la tabla 3, junto a la

representación gráfica de este ejemplo (ilustración 15). El plano de regresión obtenido es:

$$\boldsymbol{L^*(p_i) = -1.496\sqrt{w_i} + 0.057x_i^*,\ \widehat{R}^2 = 0.9}$$

Ejemplo de Regresión logística simple con MCP

xi	Ni	ni	pi	logit	$w^{0.5}$	x*	logit*	p(x)
7	50	9	0.18	-1.51634749	2.71661554	19.0163088	-4.11933316	0.23305125
9	60	13	0.21666667	-1.28519824	3.19113355	28.7202019	-4.10123924	0.25431425
11	70	19	0.27142857	-0.98738665	3.72059903	40.9265893	-3.67366983	0.27681703
14	90	29	0.32222222	-0.74357803	4.43345864	62.068421	-3.29662246	0.31277805
16	110	46	0.41818182	-0.33024169	5.17335833	82.7737332	-1.70845858	0.33810835
21	80	37	0.4625	-0.1502822	4.45954034	93.650347	-0.67018955	0.40536368
26	75	40	0.53333333	0.13353139	4.3204938	112.332839	0.57692155	0.4763688
31	60	34	0.56666667	0.26826399	3.83840245	118.990476	1.02970514	0.54834222
36	50	31	0.62	0.48954823	3.43220046	123.559217	1.68022764	0.61834864
41	35	21	0.6	0.40546511	2.89827535	118.829289	1.17514953	0.68376204

$\hat{\alpha}_1$	**0.05679698**	**-1.49633389**	$\hat{\alpha}_0$
$\sqrt{Var(\hat{\alpha}_1)}$	0.00775736	0.17670184	$\sqrt{Var(\hat{\alpha}_0)}$
$\hat{r}^2$	0.90056291	0.91337345	$s=\sqrt{s^2}$
F	36.2264372	8	$n-2$
$SSEX$	60.4438873	6.67400849	$SSNEX$

Tabla 3. Resultados numéricos ejemplo regresión logística. Fuente: elaboración propia con Excel.

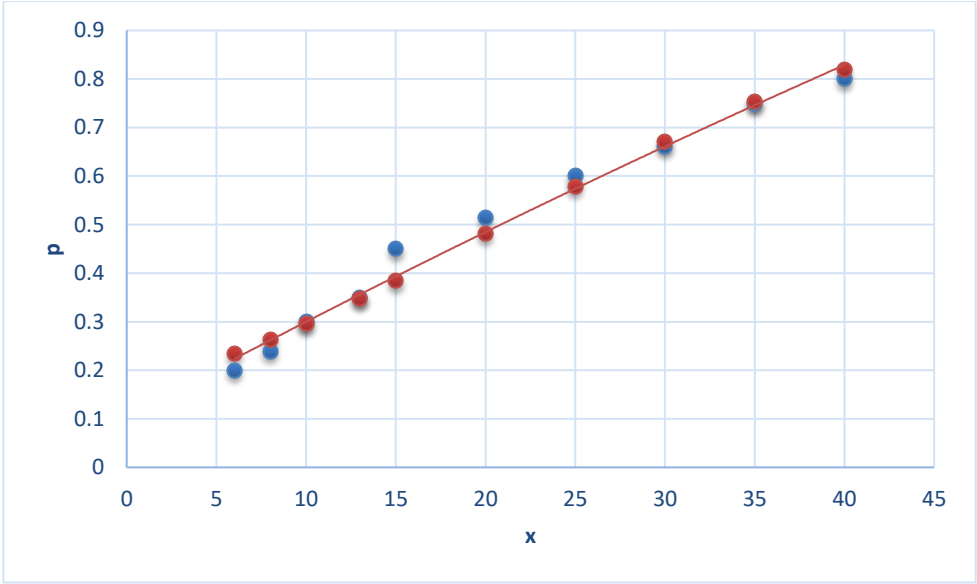


Ilustración 15. Nube de puntos y curva de regresión logística. Fuente: elaboración propia con Excel

Adicionalmente introduciendo el código correspondiente en la tabla 4, se ha rehecho con R los cálculos del ejemplo anterior, obteniéndose los resultados de la tabla 5

```
#LOGIT MCP
ej <-read.table('data_logit2.txt', header=TRUE)
p <- ej$n/ej$N
rw <- sqrt(ej$N*p*(1-p)) # Pesos para MCP
logit <- log(p/(1-p))
x2 <- ej$x*rw # x*
logit2 <- logit*rw # logit*
ej2 <- cbind(ej,p,rw,x2,logit2) # Tabla de datos ampliada
modelo_log1 <- lm(logit2 ~rw+x2+0,data=ej2)
summary(modelo_log1)
```

Tabla 4. Código aplicación método MCP. Fuente: elaboración propia con R.

```
Call:
lm(formula = logit2 ~ rw + x2 + 0, data = ej2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2372 -0.8258 -0.1949  0.5000  1.3313

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
rw -1.496334   0.176702  -8.468 2.89e-05 ***
x2  0.056797   0.007757   7.322 8.21e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9134 on 8 degrees of freedom
Multiple R-squared:  0.9006,    Adjusted R-squared:  0.8757
F-statistic: 36.23 on 2 and 8 DF,  p-value: 9.777e-05
```

Tabla 5. Resultados numéricos ejemplo regresión logística. Fuente: elaboración propia con R.

Capítulo 4. Ejercicio práctico de aplicación

El presente capítulo se centra en el análisis de datos reales cuyo tratamiento supone la aplicación directa de la metodología empleada en el presente TFM. Los datos reseñados proceden de fuentes oficiales del Arma Aérea de la Armada, por lo que su difusión se enmarca dentro del Uso Oficial.

El ejercicio consta de un análisis regresivo con el método de MCO de hasta $n = 132$ observaciones mensuales de la variable independiente, *horas de vuelo* (x_i), frente a la variable dependiente *consumo de combustible* (y_i), comenzando con la identificación de la presencia de heterocedasticidad desde el punto de vista gráfico, por medio de diagramas de dispersión, y analítico mediante la aplicación de tests de inferencia estudiados cuya hipótesis nula será la presencia de homocedasticidad. Si se detecta la presencia de heterocedasticidad se aplicará el método de MCP como corrección del MCO, donde el estudio regresivo se centrará en la aplicabilidad del método, conocida σ_i^2 (en función de la variable explicativa) y desconocida (estimada por la cuasivarianza muestral, s_i^2). Además, se realizará la comparativa de los estimadores bajo heterocedasticidad, para comprobar la pérdida de eficiencia de los mismos entre MCO y MCP. En su mayoría, todos los cálculos asociados a las operaciones antedichas se realizarán mediante $Excel$ R, $SPSS$ y $GRETL$.

Los valores se han obtenido a partir de los datos de un tipo de avión, con capacidad de aterrizaje y despegue vertical, cuyo consumo depende fundamentalmente de como haya sido la maniobra durante la hora de vuelo. En cualquier caso, se ha supuesto que el vuelo de crucero (diurno/nocturno) se ha combinado con las siguientes maniobras:

- carrera de despegue y toma en tierra o a bordo
- despegue y toma vertical en tierra o a bordo

- carrera de despegue y toma vertical a bordo (ilustraciones 16 y 17).

Ilustración 16. Carrera despegue a bordo. Fuente: elaboración propia

Ilustración 17. Toma vertical a bordo. Fuente: elaboración propia

4.1. Motivación del estudio

El ala fija embarcada de la Armada la componen los aviones de ataque *V/STOL*[21] de la Novena Escuadrilla de Aeronaves, *AV-8B plus (Harrier II),* que constituyen un importante vector de proyección de la Fuerza Naval española. Entre sus misiones destacan, la defensa aérea de la Fuerza, el ataque a unidades de superficie, el ataque y apoyo a tierra, así como el apoyo aéreo próximo (Ilustración 18).

Ilustración 18. Vuelo de apoyo a tierra. Fuente: Revista General de Marina (2023)

[21] Vertical / Short Take-Off and Landing

Actualmente, las 13 aeronaves de que consta la Escuadrilla, disponen de numerosas horas de vuelo y años de servicio con lo que se encuentran en su último tercio de vida operativa. Dado que al parecer no está clara la adquisición a corto plazo del *F-35B (Lightning II),* ilustración 19*,* se han iniciado contactos con *NAVAIR (Navy Air and System Command)* de la Armada de los Estados Unidos, con el fin de prolongar la vida útil del *Harrier* hasta el 2030 y mantener la capacidad de proyección que proporciona la aviación de caza y ataque embarcada.

Ilustración 19. Toma vertical de un *F-35B* a bordo de un portaeronaves clase *America*. Fuente: *US Marine Corps*.

Se pretende por tanto, comprobar que los residuos de los valores aportados al modelo de regresión MCO siguen un patrón heterocedástico y aplicar el modelo de regresión por MCP para corregir la ineficiencia de los estimadores de MCO. Para ello, se emplearán las cuatro herramientas informáticas reseñadas al comienzo de este capítulo junto con la aplicación de cinco contrastes de heterocedasticidad relacionados en el apartado 2.4 para la validación del modelo de regresión. La información resultante del presente estudio se considera que pudiera ser conveniente y de utilidad, para la toma de decisiones que el Mando estime oportuno, en relación con la prolongación de la vida operativa de esta aeronave.

4.2. Características técnicas de la aeronave

Las dimensiones de la aeronave se muestran en la ilustración 20 de disposición general.

En cuanto al resto de características técnicas:

- Especificaciones: Peso máximo despegue vertical 8596 kg/carga útil 4899 kg
- Velocidad máxima: 1170 km/h o 1065 km/h al nivel del mar

- Techo máximo: 15420 m
- Autonomía: máxima de 3600 km (3500 kg combustible interno + depósitos auxiliares) y de combate 870 km.
- Motor: 1 Rolls-Royce F402-RR-408 de 104.5 kN de empuje
- Combustible: JP-8 (en tierra) y JP-5 (embarcado).

Por razones de conveniencia el consumo mixto del avión procede de la suma de consumos en despegue rodado, en vuelo de ferry y toma vertical a bordo del buque de proyección logística (BPE) Juan Carlos I.

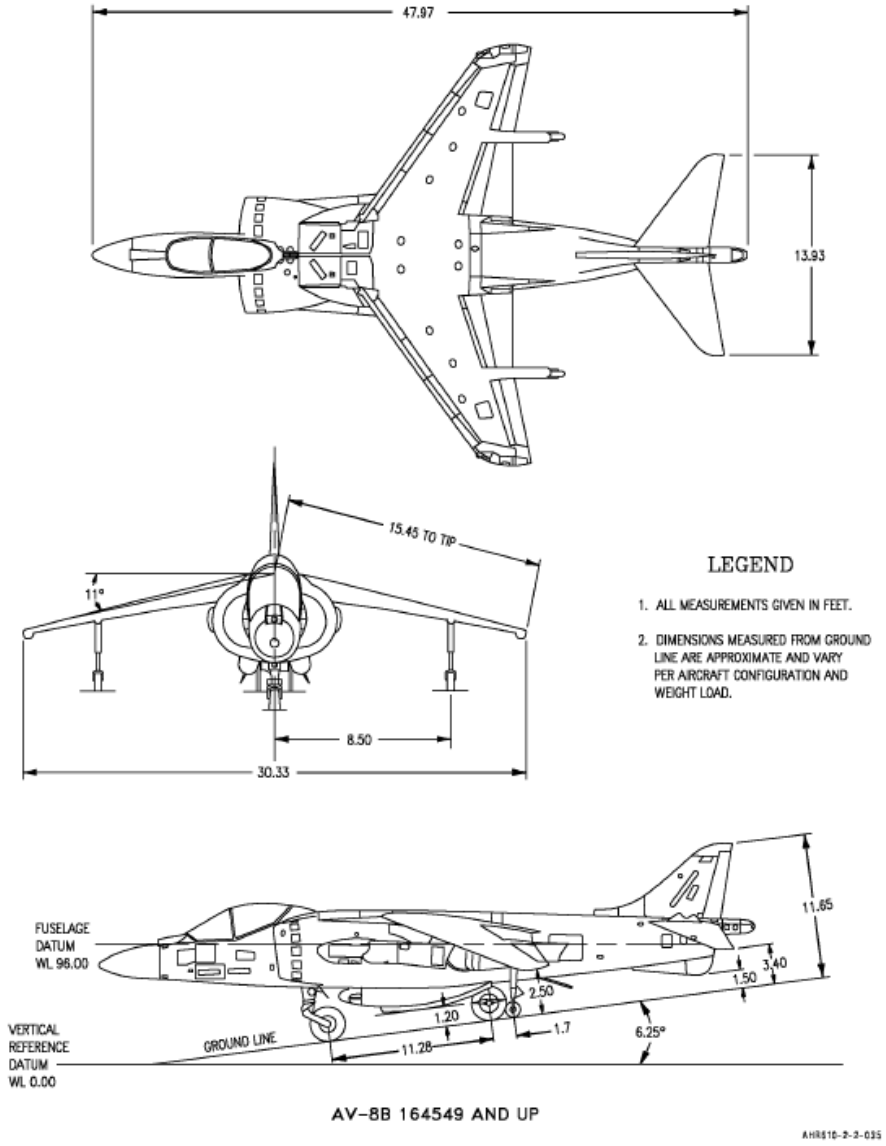


Figure 1. Aircraft Dimensions (Sheet 2)

Ilustración 20. Disposición general del *AV-8B plus (Harrier II).* Fuente: NAVAIR (2021)

4.3. Validación del modelo de regresión. Aplicación de contrastes de heterocedasticidad

Para que las conclusiones del modelo de regresión sean consistentes es necesario verificar las condiciones de aplicación en lo que respecta a la adecuación de sus estimaciones e inferencias. El modelo de inferencia visto en el apartado 4.3.1 precisa la condición de homocedasticidad, esto es, que la varianza de los errores sea constante, $Var(e_i) = \sigma^2$. Dado que en este apartado se ha detectado gráficamente la presencia de heterocedasticidad, $Var(e_i) = \sigma_i^2$, mediante la aplicación de las herramientas de inferencia estadística reseñadas en el apartado 3.3 se pretende detectar analíticamente la presencia de heterocedasticidad.

En general, los resultados de que cada uno de los tests de contraste se realizan, a continuación, bajo la hipótesis nula de ausencia de heterocedasticidad, es decir, en aplicación de la condición homocedástica con un nivel de significación de $\alpha = 0.05$, se plantea la siguiente hipótesis:

$$H_0: presencia\ de\ homocedasticidad$$

$$H_1: ausencia\ de\ homocedasticidad$$

4.3.1. Goldfeld y Quandt

Bajo las indicaciones concretas de la implementación de este test se ha confeccionado una tabla de operaciones, obrante en el apéndice B1, cuyos resultados de regresión son los que se reflejan en la tabla 6

MCO GQ1	
614.3902956	**8802.455544**
115.3753074	35419.14866
0.313834231	49917.99718
28.35717426	62
70660589528	1.54492E+11

MCO GQ2	
477.2855306	**51981.90879**
327.6493059	140067.307
0.033092625	78625.61815
2.121964102	62
13117956251	3.83283E+11

Tabla.6. Resultados MCO de las submuestras n_1 y n_2. Fuente: propia con Excel

A partir de la ecuación (3.18), resolviendo el cociente de la SSNEX de cada submuestra (resaltado en amarillo), se obtiene el estadístico $GQ = 2.48092618$. A su vez el

estadístico crítico de contraste, $F_{64,64;0.05}$, obtenido mediante Excel con $INV.F(1-0.05;64;64)=1.51328717$.

En aplicación del estadístico (3.20), con el nivel de significación $\alpha=0.05$, se rechaza H_0 porque $GQ > F_{64,64;0.05}$. En conclusión, se detecta analíticamente la presencia de heterocedasticidad.

4.3.2. Breusch-Pagan-Godfrey

Para comprobar si se admite la hipótesis de homocedasticidad con los datos aportados, es preciso matizar que este test es sensible a formas lineales de heterocedasticidad, Dado que sólo se dispone de una variable explicativa en el modelo de estudio, la validez del resultado se considera preliminar. El proceso de obtención de los resultados de regresión de MCO y del modelo MCO BPG, o auxiliar, se encuentra en el apéndice B3, de cuyas tablas se extraen los siguientes resultados (tabla 7):

MCO	
526.6284321	**33348.69285**
74.36770128	27828.87182
0.278364768	65264.65459
50.14641492	130
2.13597E+11	**5.53732E+11**

MCO BPG	
24695706.37	**4851778116**
4485957.014	1678673948
0.189052377	3936849331
30.3062841	130
4.69711E+20	2.01484E+21

Tabla 7. Resultados MCO y MCO de Breusch-Pagan. Fuente: propia con Excel

A partir del estadístico de la expresión (2.51), se obtiene que el estadístico es $BP = n\,\hat{R}^2 = 24.95491$, mientras que el estadístico crítico de contraste, ${\chi^2}_{2-1}$, se calcula con $Excel$ con $DIST.CHICUAD.CD(24.95491;2) = 5.86868 \times 10^{-7}$. Confirmándose que el $p-valor < n\,\hat{R}^2$ y que $BP > {\chi^2}_{m-1}$, con $\alpha=0.05$, se rechaza H_0 por lo que se aprecia la presencia de heterocedasticidad en los residuos, habida cuenta de que el coeficiente regresor muestra mayor poder explicativo sobre la variable dependiente y un $SSEX$ mucho mayor, que en el caso MCO.

Mediante el lenguaje R, de *RStudio* [22]se puede hallar el estadístico BP así como el $p - valor$ mediante el siguiente código y resultado asociado (ilustración 21)

```
> bptest(mod_MCO)

        studentized Breusch-Pagan test

data:  mod_MCO
BP = 24.955, df = 1, p-value = 5.87e-07
```

Ilustración 21. Resultado BPG test con RStudio. Fuente: propia con Excel

4.3.3. Koenker-Bassett

El proceso de obtención de los resultados de regresión del MCO y del modelo MCO KB, obtenido por regresión de los residuos al cuadrado sobre los valores estimados de la variable dependiente al cuadrado, se encuentra en el apéndice B4, extrayendo los siguientes resultados (tabla 8):

MCO		MCO KB	
526.6284321	**33348.69285**	**0.109609382**	**1594080578**
74.36770128	27828.87182	0.019646609	1092476604
0.278364768	65264.65459	**0.193176909**	3926825035
50.14641492	130	31.12578014	130
2.13597E+11	**5.53732E+11**	4.79958E+20	**2.00459E+21**

Tabla 8. Resultados MCO y MCO de Koenker-Bassett. Fuente: propia con Excel

Como en este test la hipótesis nula se establece en virtud al valor nulo del estimador del regresor, discutido a partir del estadístico F obtenido de la regresión $F = 31.1257801$, mientras que el estadístico crítico de contraste, $F_{1,130;0.05}$, se calcula con $Excel$ con $INV.F(1 - 0.05; 1; 130) = 3.913989$. Visto que $F > F_{1,130;0.05}$, con $\alpha = 0.05$, se rechaza H_0 por lo que se aprecia la presencia de heterocedasticidad en los residuos.

[22] RStudio es un entorno de desarrollo para el lenguaje de programación R, especialmente diseñado para la computación estadística y gráficos. Incluye una consola y editor de sintaxis como soporte del código además de diversas utilidades asociadas. Fuente: Wikipedia.

4.3.4. White

Para comprobar si se admite la hipótesis de homocedasticidad con los datos aportados, es preciso matizar que este test es sensible a formas lineales de heterocedasticidad, Dado que sólo se dispone de una variable explicativa en el modelo de estudio, la validez del resultado se considera preliminar. El proceso de obtención de los resultados de regresión de MCO y del modelo MCO auxiliar se encuentra en el apéndice B5, extrayendo los siguientes resultados (tabla 9):

MCO	
526.6284321	**33348.69285**
74.36770128	27828.87182
0.278364768	65264.65459
50.14641492	130
2.13597E+11	**5.53732E+11**

MCO WHITE	
43062.76676	**-5009467.32**
14253.36782	5704584.815
0.583222042	3925878370
90.95834373	130
2.80379E+21	**2.00363E+21**

Tabla 9. Resultados MCO y MCO de White. Fuente: propia con Excel

A partir de la expresión (2.57), se obtiene el estadístico de White $n\,\hat{R}^2 = 76.9853$, mientras que el estadístico crítico de contraste, ${\chi^2}_{76.985,2}$, definido a partir del valor anterior y $(gl = 2)$ grados de libertad, se ha obtenido con el comando $DIST.CHICUAD.CD(76.985; 2) = 1.918 \times 10^{-17} \approx 0$.

Dado que $n\,\hat{R}^2 > {\chi^2}_{76.985,2}$, con $\alpha = 0.05$, se rechaza H_0 por lo que se descarta la presencia de homocedasticidad en detrimento de la heterocedasticidad en los residuos, habida cuenta de que el coeficiente regresor muestra mayor poder explicativo sobre la variable dependiente y un $SSEX$ mucho mayor, que en el caso MCO.

4.4. Análisis de los datos

Los datos de partida se encuentran relacionados en el Apéndice A, correspondientes a la variable exógena, horas de vuelo $(\boldsymbol{x_i})$ y a la endógena $(\boldsymbol{y_i})$, consumo de combustible (en litros) y se asume que siguen el modelo lineal (2.3). Estos datos se han registrado mensualmente y abarcan un período de 11 años, desde enero de 2012 a diciembre de 2022.

4.4.1. Análisis de los datos mediante el método MCO

A priori, representando los datos se observa que se disponen en una forma de embudo, característica conforme se indica en la ilustración 22.

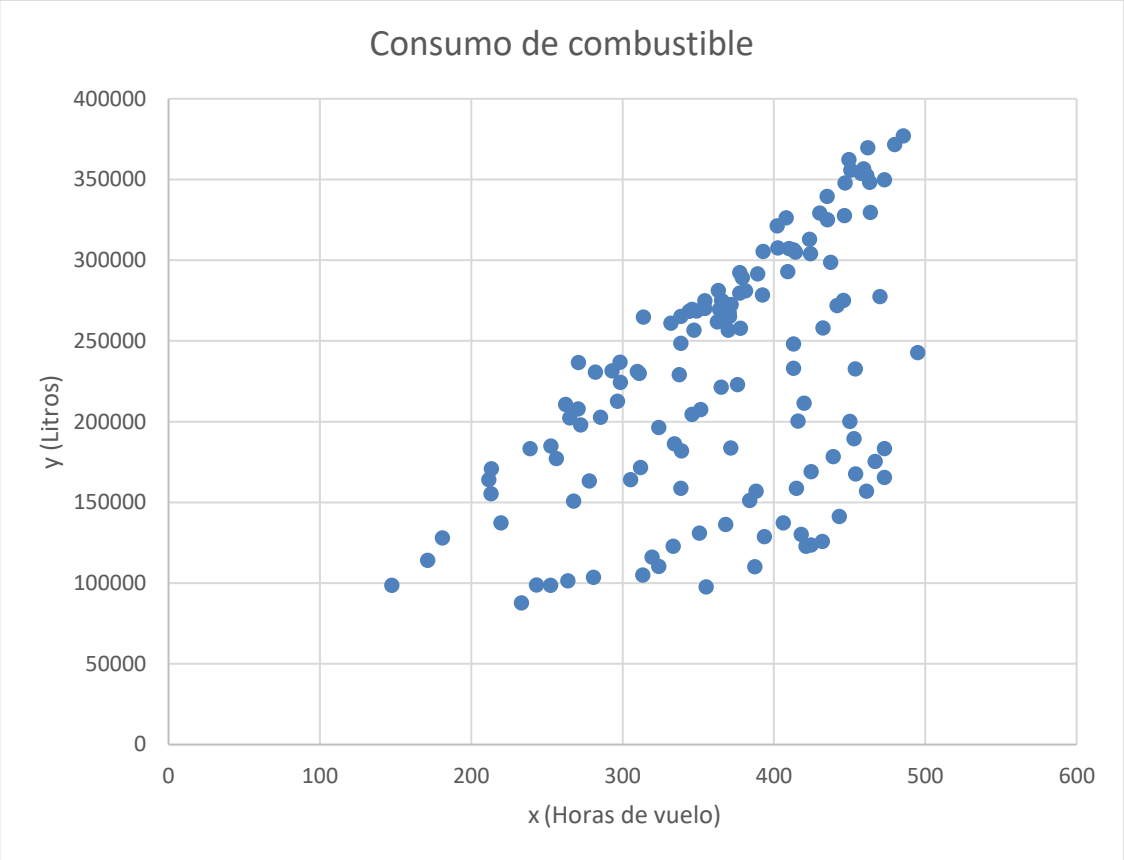


Ilustración 22. Diagrama de dispersión datos de partida. Fuente: elaboración propia con Excel.

Esta disposición gráfica subyace la presencia de heterocedasticidad creciente en los datos (véase el apéndice A), cuestión que se confirmará mediante la oportuna aplicación de los métodos de inferencia.

Los resultados analíticos de la regresión de MCO se han calculado mediante el comando *ESTIMACIÓN.LINEAL* de Excel, con los siguientes resultados de la tabla 10.

	MCO		
$\hat{\alpha}_1$	**526.6284321**	**33348.69285**	$\hat{\alpha}_0$
$\sqrt{Var(\hat{\alpha}_1)}$	**74.36770128**	27828.87182	$\sqrt{Var(\hat{\alpha}_0)}$
$\hat{r}^2$	**0.278364768**	65264.65459	$s = \sqrt{s^2}$
F	50.14641492	130	$n-2$
$SSEX$	2.13597E+11	5.53732E+11	$SSNEX$

Tabla 10. Resultado aplicación método MCO. Fuente: elaboración propia con Excel.

La recta de regresión lineal obtenida es

$$\hat{y} = 33348.693 + 526.628x,\ \hat{R}^2 = 0.278,$$

siendo el estimador de la pendiente, $\hat{\alpha}_1 = 526.628$, y el estimador del intercepto, $\hat{\alpha}_0 = 33348.693$, de cuyo valor de $\hat{R}^2$ se desprende que existe poca correlación entre las variables, por lo que esta medida de la bondad del ajuste indica, a priori, que la relación lineal entre las variables es débil.

Como primera aproximación en las ilustraciones 23 y 24 se observa que tanto los residuos como el cuadrado de los mismos aumenta con el valor de la variable estimada y explicada respectivamente, adoptando la nube de puntos, la forma de "embudo" típica que caracteriza a la heterocedasticidad.

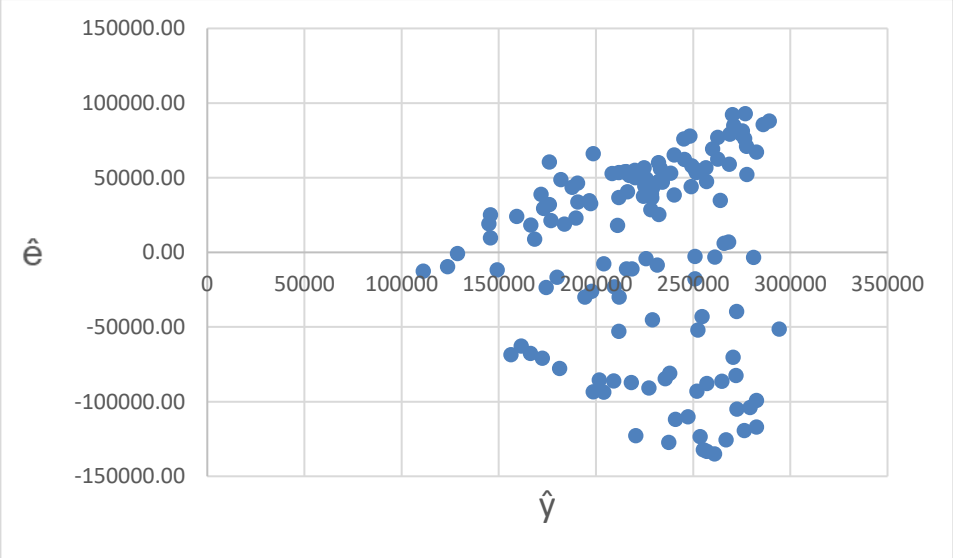


Ilustración 23. Diagrama de dispersión de residuos $\hat{e}$ y variable estimada $\hat{y}$. Fuente: elaboración propia con Excel

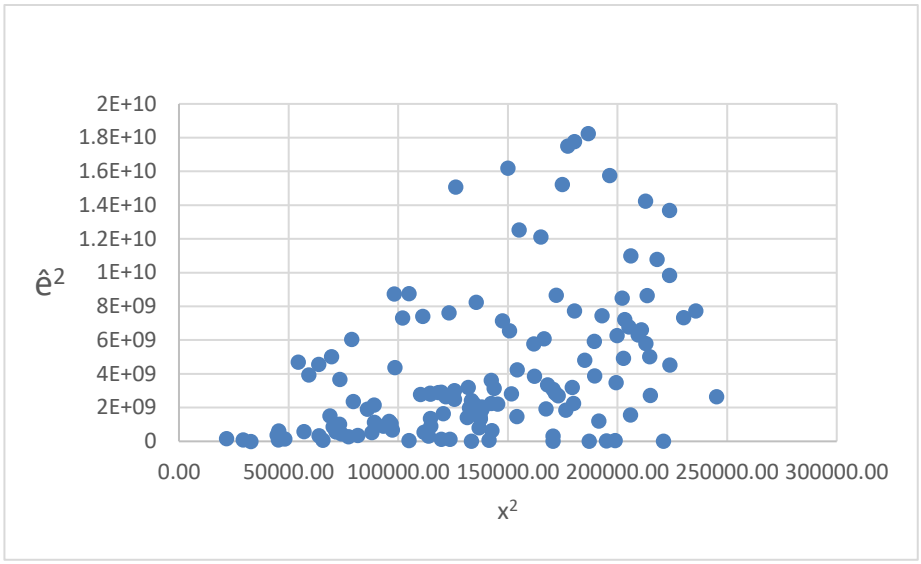


Ilustración 24. Diagrama de dispersión de $\hat{e}^2$ frente a la variable x^2. Fuente: elaboración propia con Excel

Asimismo, se considera útil representar por un lado, el valor de la estimación de la variable explicada, $\hat{y}$, frente al cuadrado de los residuos, $\hat{e}^2$, (ilustración 25), y por otro, frente a los errores residuales tipificados en valor absoluto, $|\hat{e}|/s$, (ilustración 26). En ambos casos se mantiene el patrón de heterocedasticidad, y el coeficiente de correlación es muy bajo.

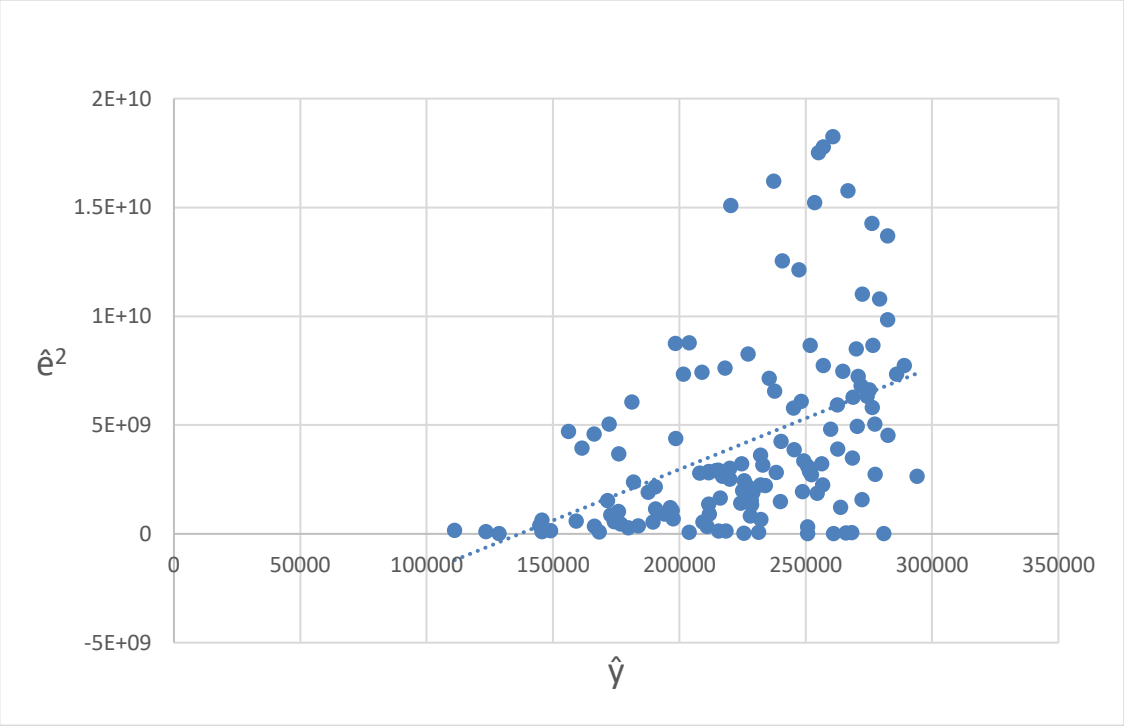


Ilustración 25. Diagrama de dispersión de $\hat{e}^2$ frente a $\hat{y}$, y recta de regresión $\hat{e}^2 = 44867\hat{y} - 6 \times 10^9$. Fuente: elaboración propia con Excel

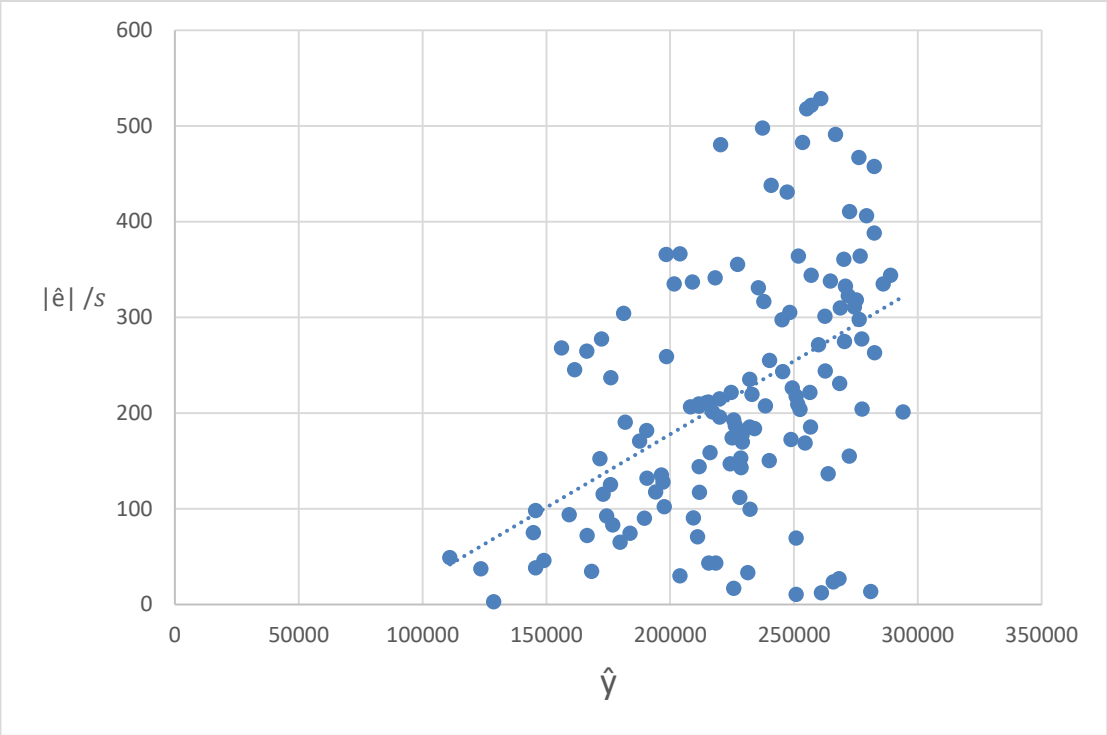


Ilustración 26. Diagrama de dispersión de $|\hat{e}|/s$ frente a $\hat{y}$, y recta de regresión $|\hat{e}|/s = 0,0015\hat{y} - 128,2$.

Fuente: elaboración propia con Excel

En cuanto a la inferencia del modelo de regresión, se contrasta al nivel de $\alpha = 0.05$ la hipótesis nula H_0 de que las variables x e y no están relacionadas linealmente frente a la hipótesis alternativa H_1 de que sí lo están. Se obtiene el valor del estadístico F de *Snedecor* asociado a la estimación del resultado del MCO como $F = 50.146$ y el punto crítico, $F_{1,n-2,\alpha}$, con el comando de Excel $INV.F(1-\alpha; m; n)$ de manera que $F_{1,130;0.05} = INV.F(1-0.05; 1; 130) = 3.913$.

En virtud del contraste (2.30), como $F > F_{1,130;\,0.05}$ se rechaza H_0, es decir, con un 95% de confianza se admite que existe una relación lineal entre x e y. Procediendo igualmente al 99% de confianza, se obtiene que el punto crítico $F_{1,130;\,0.01} = 6.834$, con lo que se mantiene dicha relación lineal.

Mediante el lenguaje de la aplicación $RStudio$, o R, introduciendo el código siguiente (tabla 11):

```
#MCO
ej<-read.table('data_harrier.txt', header=TRUE)
x<-ej$X
y<-ej$Y
mod_MCO<-lm(y~x, data=ej)
coef(mod_MCO)
summary(mod_MCO)
```

Tabla 11. Código aplicación método MCO con *R*. Fuente: elaboración propia

donde se obtiene también la misma recta de regresión y datos de estimación asociados que en Excel (tabla 12),

```
> summary(mod_MCO)

Call:
lm(formula = y ~ x, data = ej)

Residuals:
    Min      1Q  Median      3Q     Max
-135059  -52288   22146   52876   92995

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33348.04   27828.87   1.198    0.233
x             526.63      74.37   7.081 8.02e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65260 on 130 degrees of freedom
Multiple R-squared:  0.2784,    Adjusted R-squared:  0.2728
F-statistic: 50.15 on 1 and 130 DF,  p-value: 8.022e-11
```

Tabla 12. Resultado aplicación método MCO con *R*. Fuente: elaboración propia

De igual forma, mediante la aplicación SPSS, se obtienen resultados similares, como es el caso de $\widehat{R}^2$ (tabla 13)

Resumen del modelo MCO

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,528[a]	,278	,273	65264,655

a. Predictores: (Constante), X

Tabla 13. Coeficiente determinación con MCO. Fuente: elaboración propia con SPSS

y los coeficientes de estimación del modelo (tabla 14), ANOVA (tabla 15) además del diagrama de dispersión (ilustración 27)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Desv. Error	Beta		
1	(Constante)	33348,693	27828,872		1,198	,233
	X	526,628	74,368	,528	7,081	<,001

a. Variable dependiente: Y

Tabla 14. Coeficientes de regresión y varianza con MCO. Fuente: elaboración propia con SPSS

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	2,136E+11	1	2,136E+11	50,146	<,001[b]
	Residuo	5,537E+11	130	4259475138,5		
	Total	7,673E+11	131			

a. Variable dependiente: Y

b. Predictores: (Constante), X

Tabla 15. Resultado ANOVA de MCO. Fuente: elaboración propia con SPSS

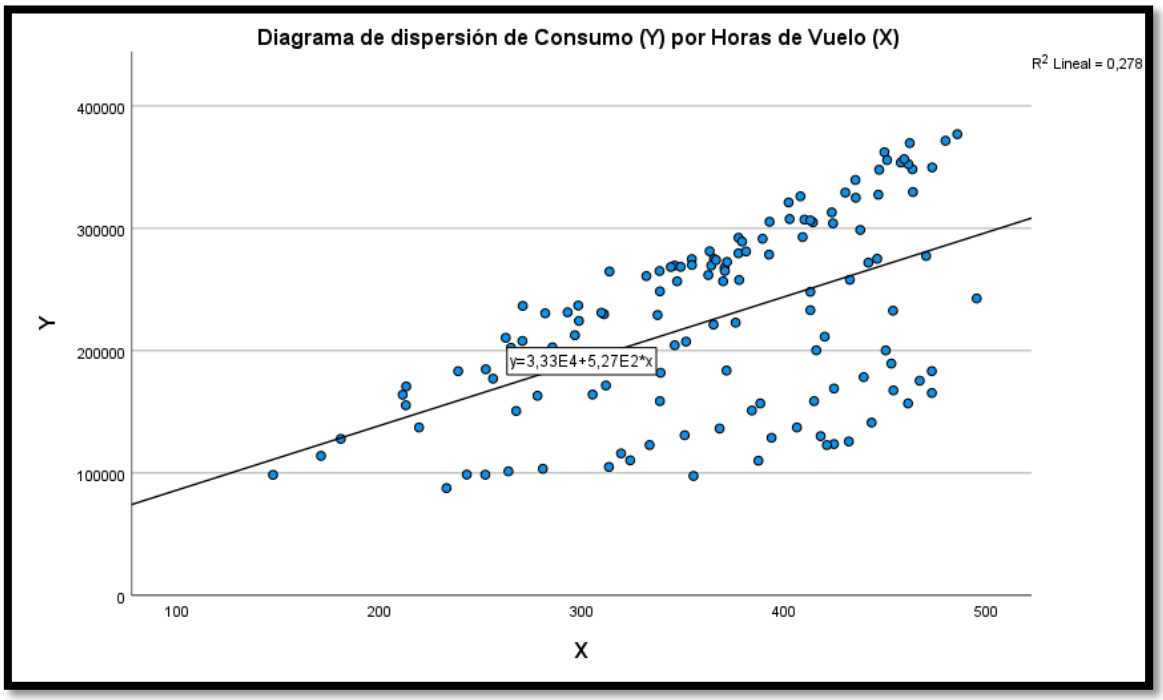


Ilustración 27. Diagrama de dispersión con recta de regresión. Fuente: elaboración propia con SPSS

Finalmente, con la aplicación GRETL[23] mediante las operaciones oportunas se obtiene el resultado de la regresión (Tabla 16).

Modelo MCO, usando las observaciones 1-132
Variable dependiente: y_i

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	33348.0	27828.9	1.198	0.2330	
X	526.630	74.3677	7.081	<0.0001	***

Media de la variable. dep.	226267.2	D.T. de la variable dep.	76534.14
Suma de cuad. residuos	5.54e+11	D.T. de la regresión	65264.59
R-cuadrado	0.278366	R-cuadrado corregido	0.272815
F(1, 130)	50.14676	Valor p (de F)	8.02e-11

Tabla 16. Resultados aplicación método MCO. Fuente: elaboración propia con GRETL

4.4.2. Análisis de los datos mediante el método MCP con σ_i^2 desconocida

Asumiendo la situación heterocedástica con los elementos gráficos anteriores se procede a analizar los datos mediante el método MCP desconociendo el valor de σ_i^2 que será estimado, en este caso, mediante s_i^2.

Procederemos, en primer lugar, con Excel mediante la transformación del modelo según se ha visto. Para ello, se Introducen los pesos dados por $w_i = 1/s_i^2$, o por $\sqrt{w}_i = 1/s_i$, en x_i e y_i obteniéndose x_i^* e y_i^*respectivamente, (tabla 17), a los que se le aplicará el método MCO, ya que los errores e_i^*, se distribuyen de forma homocedástica, es decir, $Var(e_i^*) = 1$.

s_i^2	s_i	$sw_i = 1/s_i$	$x_i^* = x_i sw_i$	$y_i^* = y_i * sw_i$
52273774.2	723.006.046	0.00013831	0.06265508	262.066.965
82952457.4	910.782.397	0.0001098	0.0512746	192.624.496
105310694	102.620.999	9.74E+00	0.04609193	161.212.619
57352822.4	757.316.462	0.00013205	0.05801142	235.522.412
121259561	110.117.919	9.08E+00	0.0402523	128.247.066
44576072.9	66.765.315	0.00014978	0.06911024	527.920.822
50891377.2	713.381.926	0.00014018	0.06438814	4.998.511
12644332.4	355.588.701	0.00028122	0.09763152	721.957.361

Tabla 17. Extracto parámetros utilizados en aplicación del método MCP. Fuente: elaboración propia con Excel

[23] *GRETL (Gnu Regression, Econometrics and Time-series Library*) es un software libre desarrollado en lenguaje C que se utiliza para realizar análisis econométricos cuyo uso está extendido en diversos departamentos de economía de universidades de todo el mundo.

Actuando como en el apartado anterior se obtiene con Excel la regresión lineal por MCP, donde el intercepto nulo desaparece, (tabla 18),

	MCP1			
$\tilde{\alpha}_1$	524.8515157	32973.46113	0	$\tilde{\alpha}_0$
$\sqrt{Var(\tilde{\alpha}_1)}$	6.664055457	1695.779794	#N/A	$\sqrt{Var(\tilde{\alpha}_0)}$
$\hat{r}^2$	0.997912516	11.42288749	#N/A	$s = \sqrt{s^2}$
F	31072.96518	130	#N/A	$n-2$
$SSEX$	8108947.576	16962.70663	#N/A	$SSNEX$

Tabla 18. Resultado aplicación método MCP. Fuente: elaboración propia con Excel

cuya recta de regresión lineal es de la forma

$$\boldsymbol{y = 32973.461 + 524.852x},\ \boldsymbol{\widehat{R}^2 = 0.998}, \text{ n}$$

siendo $\tilde{\alpha}_1 = \mathbf{524.852}$ y $\tilde{\alpha}_0 = \mathbf{32973.461}$, de cuyo valor de $\boldsymbol{\widehat{R}^2}$ se desprende que existe alta correlación entre las variables x e y.

Haciendo uso del lenguaje R, se llega al mismo resultado introduciendo, en primer lugar, el siguiente código, donde se introducen los pesos w_i mediante la inversa de la cuasivarianza $(1/s_i^2)$, las variables explicativa y explicada como el producto de la variable $\sqrt{w_i}$ por cada una de ellas, regresando x_i y $\sqrt{w_i}$ con intercepto nulo sobre y_i mediante el comando lm de R (tabla 19):

```
#MCP
ej<-read.table('data_harrier.txt', header=TRUE)
wi<-1/ej$si2
x2<-ej$X*sqrt(wi)
y2<-ej$Y*sqrt(wi)
ej2<-cbind(ej,wi,x2,y2)
mod_MCP<-lm(y2~0+sqrt(wi)+x2, data=ej2)
coef(mod_MCP)
summary(mod_MCP)
```

Tabla 19. Código aplicación método MCP. Fuente elaboración propia con *R*

el resultado de MCP bajo R muestra los mismos resultados que Excel, recordando del apartado 3.2.4 c) que el intercepto en estos casos es nulo (tabla 20),

```
> summary(mod_MCP)

Call:
lm(formula = y2 ~ 0 + sqrt(wi) + x2, data = ej2)

Residuals:
   Min     1Q Median     3Q    Max
-11.31 -11.17  11.57  11.65  13.60

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
sqrt(wi) 32973.527   1695.780   19.44   <2e-16 ***
x2         524.851      6.664   78.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.42 on 130 degrees of freedom
Multiple R-squared:  0.9979,    Adjusted R-squared:  0.9979
F-statistic: 3.107e+04 on 2 and 130 DF,  p-value: < 2.2e-16
```

Tabla 20. Resultado aplicación método MCP. Fuente elaboración propia con *R*

Otra forma de ajustar el modelo ponderado se expresa con el siguiente código (tabla 21), consistente en introducir los pesos w_i mediante la inversa de la cuasivarianza $(1/s_i^2)$, y regresando la variable $\sqrt{w_i}$, el producto$\sqrt{w_i} \times x_i$ e intercepto nulo sobre el el producto$\sqrt{w_i} \times y_i$ mediante el comando lm

```
#MCP1_OK (modelo PONDERADO)
ej<-read.table('data_harrier.txt', header=TRUE)
wi<-1/ej$si2
mod_MCP1 <- lm(I(sqrt(wi) * Y) ~ 0 + sqrt(wi) + I(sqrt(wi) * X))
coef(mod_MCP1)
summary(mod_MCP1)
```

Tabla 21. Código aplicación método MCP (modelo ponderado). Fuente elaboración propia con *R*

cuyo resultado es necesariamente idéntico al anterior (tabla 22),

```
> summary(mod_MCP1)

Call:
lm(formula = I(sqrt(wi) * Y) ~ 0 + sqrt(wi) + I(sqrt(wi) * X))

Residuals:
   Min     1Q Median     3Q    Max
-11.31 -11.17  11.57  11.65  13.60

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
sqrt(wi)        32973.584   1695.783   19.44   <2e-16 ***
I(sqrt(wi) * X)   524.851      6.664   78.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.42 on 130 degrees of freedom
Multiple R-squared:  0.9979,    Adjusted R-squared:  0.9979
F-statistic: 3.107e+04 on 2 and 130 DF,  p-value: < 2.2e-16
```

Tabla 22. Resultado aplicación método MCP (modelo ponderado). Fuente elaboración propia con *R*

Sin embargo, R dispone de un modo específico de incluir los pesos en el modelo mediante el comando $weights$, esto es, por regresión de x sobre y mediante el comando lm pudiéndose efectuarse de dos formas diferentes. La primera, haciendo uso de la función $weights$ con pesos, w_i, conforme al siguiente código (tabla 23),

```
#MCP2_OK (modelo con función WEIGHT con pesos)
ej<-read.table('data_harrier.txt', header=TRUE)
wi<-1/ej$si2#con pesos
mod_MCP2<-lm(Y~X, weights = wi)
coef(mod_MCP2)
summary(mod_MCP2)
```

Tabla 23. Código aplicación método MCP (modelo función *WEIGHTS*). Fuente elaboración propia con *R*

que proporciona el resultado que se muestra a continuación en la tabla 24

```
> summary(mod_MCP2)

Call:
lm(formula = Y ~ X, weights = wi)

Weighted Residuals:
   Min     1Q Median     3Q    Max
-11.31 -11.17  11.57  11.65  13.60

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32973.584   1695.783   19.44   <2e-16 ***
X             524.851      6.664   78.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.42 on 130 degrees of freedom
Multiple R-squared:  0.9795,    Adjusted R-squared:  0.9793
F-statistic:  6203 on 1 and 130 DF,  p-value: < 2.2e-16
```

Tabla 24. Resultado aplicación método MCP (*WEIGHTS* con pesos). Fuente: elaboración propia con R

La segunda manera, se introduce la probabilidad observada de los pesos con el comando $weights$ utilizando el siguiente código (tabla 25),

```
#MCP3_OK (modelo con función WEIGHT con probabilidades)
ej<-read.table('data_harrier.txt', header=TRUE)
w <- ej$wi/sum(ej$wi)#con probabilidades
mod_MCP3 <-lm(Y ~ X, weights = w)
coef(mod_MCP3)
summary(mod_MCP3)
```

Tabla 25. Código aplicación método MCP (*WEIGHTS* con probabilidad). Fuente: elaboración propia con R

cuyo resultado se adjunta a continuación (tabla 26),

```
> summary(mod_MCP3)

Call:
lm(formula = Y ~ X, weights = w)

Weighted Residuals:
   Min     1Q Median     3Q    Max
-653.0 -644.8  667.9  672.7  785.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32973.583   1695.782   19.44   <2e-16 ***
X             524.851      6.664   78.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 659.6 on 130 degrees of freedom
Multiple R-squared:  0.9795,    Adjusted R-squared:  0.9793
F-statistic:  6203 on 1 and 130 DF,  p-value: < 2.2e-16
```

Tabla 26. Resultado aplicación método MCP (*WEIGHTS* con probabilidad) Fuente elaboración propia con *R*

Representando gráficamente el resultado de la aplicación de los métodos MCO y MCP, se puede observar la diferencia entre los modelos. Para ello se introduce el siguiente código de R (tabla 27),

```
#COMPARATIVA GRÁFICA MCO Y MCP
x_seq <- data.frame(X = seq(from = 0, to = 600, by = 100))

plot(data.frame(X = x_seq, Y = predict(mod_MCO, x_seq)), type = "l",
     main = expression(bar(y) == alpha[0] + alpha[1]*X), col = 2,
     xlab = "Horas de vuelo")


abline(h = 0, lty = 2)
lines(data.frame(X = x_seq, Y = predict(mod_MCP, x_seq)), col = 4))

legend("topleft", bty = "n", legend = c("MCO","MCP"), col = c(2, 4), lty = 1)
```

Tabla 27 Código representación comparativa métodos MCO y MCP Fuente elaboración propia con *R*

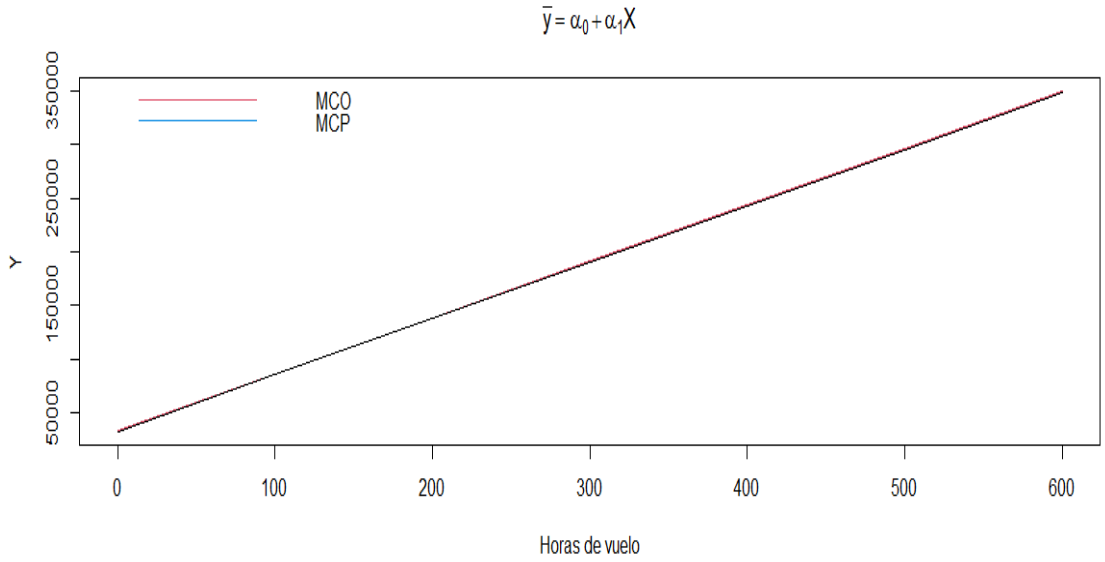


Ilustración 28 Resultado comparativo recta regresión MCO y MCP. Fuente: elaboración propia con *R*

Aunque aparentemente las rectas de regresión sean iguales el intercepto es inferior en el modelo propuesto para MCP (ilustración 28).

En resumen, se observa en los resultados con R, que los cuatro modelos utilizados muestran coeficientes regresores iguales pero, en realidad, el resultado de algunos parámetros difieren. Entre los modelos ponderados (MCP y MCP1) y los modelos con la función $weights$ (MCP2 y MCP3) el valor del estadístico F de *Snédecor* es superior en el primer caso, aunque el modelo resulta claramente significativo en las dos situaciones al ser $F > F_{1,130,0.1}$ al 99%. El asunto principal está enla precisión de los estimadores y que eso se tratará en la próxima sección En cuanto a los modelos de MCP2 y MCP3 los valores de los residuos se han resaltado porque no resultan iguales; el vector de pesos utilizado en MCP2 ofrece los mismos residuos que los modelos ponderados en MCP y MCP1. Consultada la ayuda de (RStudio, s. f.)R[24] sobre la utilización de $weights$ aclara que "*si se desconoce la varianza dentro de cada grupo se ha utilizar el vector de pesos del modelo ponderado*", por tal motivo, el modelo MCP3 se desecha. Por lo que se

[24] $weights$: an optional vector of weights to be used in the fitting process. Should be NULL or a numeric vector. If non-NULL, weighted least squares is used with $weights$ (that is, minimizing $sum(w * e^2)$); otherwise ordinary least squares is used. See also 'Details'... Non-NULL weights can be used to indicate that different observations have different variances (with the values in weights being inversely proportional to the variances); or equivalently, when the elements of weights are positive integers w_i, that each response y_i is the mean of w_i unit-weight observations (including the case that there are w_i observations equal to y_i and the data have been summarized).

corrobora que el tratamiento que hace R con los pesos es el mismo que se ha desarrollado en este TFM.

En cuanto al cálculo de MCP mediante el software *SPSS*, se comprueba que los resultados son iguales que en los cálculos precedentes con *Excel y R,* para el cálculo del coeficiente de determinación, el análisis de la varianza (ANOVA) y los coeficientes de regresión (tablas 28, 29 y 30)

Resumen del modelo MCP con si^2[b,c]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,990[a]	,979	,979	11,423

a. Predictores: (Constante), X

b. Variable dependiente: Y

c. Regresión de mínimos cuadrados ponderada - Ponderada por si^2

Tabla 28. Coeficiente determinación con MCP. Fuente: propia con SPSS

ANOVA[a,b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	809370,750	1	809370,750	6202,913	<,001[c]
	Residuo	16962,707	130	130,482		
	Total	826333,456	131			

a. Variable dependiente: Y

b. Regresión de mínimos cuadrados ponderada - Ponderada por si^2

c. Predictores: (Constante), X

Tabla 29. Resultado ANOVA de MCP. Fuente: elaboración propia con SPSS

Coeficientes[a,b]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Desv. Error	Beta	t	Sig.
1	(Constante)	32973,461	1695,780		19,444	<,001
	X	524,852	6,664	,990	78,759	<,001

a. Variable dependiente: Y

b. Regresión de mínimos cuadrados ponderada - Ponderada por si^2

Tabla 30. Coeficientes de regresión con MCP. Fuente: elaboración propia con SPSS

A su vez, el resultado obtenido con GRETL viene a corroborar que el procedimiento realizado con anterioridad es correcto, tal y como se indica en la tabla 31 estimando la varianza de las perturbaciones con s_i^2,

MCP, usando las observaciones 1-132
Variable dependiente: Y
Variable utilizada como ponderación: s_i^2

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	32973.6	1695.78	19.44	<0.0001	***
X	524.851	6.66406	78.76	<0.0001	***

Estadísticos basados en los datos ponderados:

Suma de cuad. residuos	16962.80	D.T. de la regresión	11.42292
R-cuadrado	0.979472	R-cuadrado corregido	0.979314
F(1, 130)	6202.881	Valor p (de F)	1.4e-111

Estadísticos basados en los datos originales:

Media de la vble. dep.	226267.2	D.T. de la vble. dep.	76534.14
Suma de cuad. residuos	5.54e+11	D.T. de la regresión	65272.93

Tabla 31. Cuadro resumen regresión MCP. Fuente: elaboración propia con GRETL

4.4.3. Comparativa estimadores MCO frente a estimadores MCP

Un ejemplo claro para comprobar la magnitud de la pérdida de eficiencia bajo heterocedasticidad de los mínimos cuadrados ordinarios en comparación con los mínimos cuadrados ponderados, se muestra a continuación:

Recuperando del resultado de MCO (tabla 10) para el caso homocedástico, el valor del estimador y su varianza

$$\hat{\alpha}_1 = \mathbf{526.628},$$

$$\sqrt{Var(\hat{\alpha}_1)_{hom}} = \mathbf{74.368},$$

para el caso heterocedástico de MCO los mismos valores a partir de (3.1) y (3.6), sustituyendo en este último caso σ_i^2 por s_i^2

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n} x'_i y'_i}{\sum_{i=1}^{n} x'^2_i} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{4.06 \times 10^8}{7.7 \times 10^5} = \mathbf{526.628},$$

$$\sqrt{Var(\tilde{\alpha}_1)_{het}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 s_i^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2}} = \sqrt{\frac{2.06 \times 10^{13}}{(7.7 \times 10^5)^2}} = \mathbf{5.891},$$

por último, el valor del estimador y su varianza con MCP recuperado de la tabla 18 como en el caso anterior

$$\tilde{\alpha}_1 = \mathbf{524.852},$$

$$\sqrt{Var(\tilde{\alpha}_1)} = \mathbf{6.664}$$

En primer lugar, se observa en cada caso que los resultados de los estimadores de mínimos cuadrados de α_1 son prácticamente el mismo en los tres casos, manteniéndose la insesgadez. En segundo lugar, respecto a la varianza de los estimadores, comparando los cocientes de los tres casos

$$\frac{\sqrt{Var(\hat{\alpha}_1)_{hom}}}{\sqrt{Var(\hat{\alpha}_1)_{het}}} = \frac{\mathbf{74.368}}{\mathbf{5.891}} = \mathbf{12.62},$$

$$\frac{\sqrt{Var(\hat{\alpha}_1)_{hom}}}{\sqrt{Var(\tilde{\alpha}_1)}} = \frac{\mathbf{74.368}}{\mathbf{6.664}} = \mathbf{11.16},$$

$$\frac{\sqrt{Var(\hat{\alpha}_1)_{het}}}{\sqrt{Var(\tilde{\alpha}_1)}} = \frac{\mathbf{5.891}}{\mathbf{6.664}} = \mathbf{0.88 \sim 1},$$

por lo que finalmente se comprueba que en condiciones de heterocedasticidad, la varianza de los estimadores MCP son igual de eficientes que los de MCO con σ_i^2 estimada mediante s_i^2 con (2.38).

Capítulo 5. Conclusiones y líneas futuras

La idea de partida de este trabajo ha sido desarrollar el método de mínimos cuadrados ponderados (MCP), dado que se trata de un tema inédito en el currículo del máster.

El inicio del TFM se ha centralizado en la introducción y revisión del método de mínimos cuadrados ordinarios, la validación bajo el supuesto de homocedasticidad (donde la varianza de los errores permanece constante) de un parámetro poblacional para la obtención de la recta de regresión y del coeficiente de correlación lineal mediante técnicas de inferencia estadísticamente. Se ha demostrado que los estimadores y la varianza de los mismos en el modelo de regresión lineal bajo la hipótesis antedicha son insesgados y eficientes, donde en todo momento se ha supuesto que las observaciones del modelo se distribuyen según una distribución normal conforme al Teorema de *Gauss-Markov*.

Considerando como hipótesis que la varianza de los residuos no permanece constante para cada observación, se ha introducido el fenómeno de la heterocedasticidad, las causas que la motivan, su detección gráfica y, sobre todo, el estudio de los estimadores bajo esta suposición. Se comprueba que aunque los estimadores son insesgados la varianza de los mismos deja de ser eficiente, y la cuasivarianza ya no es un estimador insesgado. Para detectar analíticamente la presencia de heterocedasticidad se han enunciado, cronológicamente, los 5 contrastes más conocidos para la inferencia de

ausencia de heterocedasticidad: *Goldfeld-Quandt, Park, Breusch-Pagan-Godfrey, Koenker-Basset* y *White*.

Una vez contextualizada la heterocedasticidad y la revisión del método de MCO bajo este supuesto, se ha introducido el método de MCP como corrección al método MCO, donde se ha formulado con generalidad, la minimización de la varianza residual mediante la inclusión de pesos que introducen una ponderación en las observaciones, transformando el modelo para asegurar la homocedasticidad de los errores, garantizar la aplicabilidad del método MCO para obtener los estimadores de la ecuación de regresión. Con ello, se ha comprobado que la varianza de los estimadores MCP es eficiente (mínima), concluyéndose con un ejemplo ilustrativo (a partir de varianza conocida), donde se compara la eficiencia de un estimador de MCO frente al de MCP. Otro ejemplo adecuado que ilustra la aplicación del método MCP se muestra cuando se desconoce la varianza de cada una de las observaciones, estimándose ésta mediante la varianza muestral. Seguidamente, se ha concluido la exposición del método con un ejemplo de aplicación de MCP, que es el modelo de regresión logística aplicable a variables de carácter categórico o cualitativo en situación heterocedástica, cuya función asociada viene dada por la transformación logística de la probabilidad. Este ejemplo muestra un modelo de regresión que, por construcción, es heterocedástico.

En conclusión, a partir de datos reales se ha confeccionado un ejemplo práctico con la resolución de diversos apartados, donde a partir del análisis de los datos se ha demostrado la presencia de heterocedasticidad, tal y como demuestra la aplicación de los cinco métodos de contraste aplicados. La comparación de resultados con el método de MCO, confirma que el método MCP, corrige al primero en presencia de heterocedasticidad. Además, mediante la estimación de la varianza de la perturbación con la cuasivarianza muestral, a priori. A pesar de que los cálculos se han realizado con 4 tipos de software diferente, y en todos han coincidido los resultados, ha sido con RStudio donde se ha observado una mayor utilidad en la utilización del método MCP al poder comparar los resultados de diversos modelos de ponderación.

Como continuación natural, la generalización del método MCP aplicado al modelo de regresión múltiple sí como la introducción del método de mínimos cuadrados penalizados o *penalized least squares regression (PLSR)* especialmente útiles en la

resolución de problemas de inversión lineal en el tratamiento de señales, como es el caso de los problemas de *denoising* y deconvolución.

Apéndice A. Datos de horas de vuelo y consumo de combustible (2012-2022)

	9ª Escuadrilla (AV-8B PLUS)	
Período	**Horas vuelo (x_i)**	**Consumo de Combustible (y_i)**
ene-12	453.00	189476.00
feb-12	467.00	175439.00
mar-12	473.00	165438.00
abr-12	439.33	178365.00
may-12	443.25	141223.00
jun-12	461.25	156879.00
jul-12	424.67	123678.00
ago-12	233.25	87653.00
sep-12	243.25	98724.00
oct-12	473.00	183266.00
nov-12	454.00	167537.00
dic-12	432.00	125793.00
ene-13	181.00	127906.00
feb-13	272.33	197992.05
mar-13	213.17	155405.00
abr-13	414.25	304987.00
may-13	256.25	177155.00
jun-13	252.42	98640.00
jul-13	239.00	183203.00
ago-13	213.33	170744.00
sep-13	265.08	202383.00
oct-13	311.08	229827.00
nov-13	377.50	292289.00
dic-13	171.25	114017.00
ene-14	298.33	236877.00
feb-14	293.08	231369.00

mar-14	473.17	349781.00
abr-14	309.75	231050.00
may-14	345.92	269591.00
jun-14	392.83	305375.00
jul-14	362.58	261833.00
ago-14	298.67	224355.00
sep-14	447.00	347905.00
oct-14	463.42	348319.00
nov-14	312.00	171579.00
dic-14	278.17	163175.00
ene-15	370.58	267629.00
feb-15	296.67	212648.00
mar-15	415.92	200314.00
abr-15	305.42	164135.00
may-15	432.50	258001.00
jun-15	371.83	272557.00
jul-15	413.00	233103.00
ago-15	280.83	103492.00
sep-15	368.17	136384.00
oct-15	350.92	130896.00
nov-15	495.17	242714.00
dic-15	319.50	116008.00
ene-16	377.50	279572.00
feb-16	409.17	292854.00
mar-16	479.75	371622.00
abr-16	348.92	268536.00
may-16	381.25	281118.00
jun-16	351.58	207401.00
jul-16	334.25	186221.00
ago-16	446.00	275142.00
sep-16	424.67	169093.00
oct-16	446.58	327523.00
nov-16	418.08	130152.00
dic-16	324.00	110350.00
ene-17	430.25	329251.00
feb-17	377.83	257764.00
mar-17	450.22	200196.00
abr-17	338.67	158762.00
may-17	457.67	353827.11
jun-17	453.83	232710.44
jul-17	414.83	158762.00
ago-17	211.58	164006.00
sep-17	449.58	362271.78
oct-17	485.58	376970.00
nov-17	402.25	321181.00
dic-17	219.67	137241.00

ene-18	365.33	275035.00
feb-18	376.08	222922.00
mar-18	406.33	137241.00
abr-18	370.75	265145.68
may-18	410.08	307151.00
jun-18	441.67	271949.00
jul-18	338.50	265145.68
ago-18	262.50	210549.00
sep-18	364.08	269588.88
oct-18	463.67	329688.00
nov-18	369.92	256711.00
dic-18	263.83	101373.00
ene-19	333.58	122879.00
feb-19	313.75	264744.84
mar-19	470.17	277477.13
abr-19	323.92	196300.00
may-19	388.25	156880.00
jun-19	355.25	97626.00
jul-19	365.17	221339.00
ago-19	339.00	181936.37
sep-19	420.08	211470.69
oct-19	413.00	248136.00
nov-19	270.75	207944.00
dic-19	267.67	150702.00
ene-20	337.50	229161.00
feb-20	338.67	248499.00
mar-20	285.58	202739.00
abr-20	252.67	184789.00
may-20	379.25	289187.00
jun-20	346.00	204465.00
jul-20	331.97	260986.00
ago-20	313.50	104952.00
sep-20	387.33	110077.00
oct-20	393.83	128803.00
nov-20	384.08	151081.00
dic-20	421.17	122836.00
ene-21	270.92	236608.55
feb-21	437.63	298696.00
mar-21	435.42	325013.00
abr-21	408.08	326206.00
may-21	450.92	355798.00
jun-21	423.58	313038.00
jul-21	413.00	306479.18
ago-21	389.42	291494.93
sep-21	363.25	281274.00
oct-21	147.57	98524.00

nov-21	435.25	339544.00
dic-21	366.42	274037.53
ene-22	354.42	274871.00
feb-22	282.00	230575.00
mar-22	402.75	307619.00
abr-22	424.17	304103.00
may-22	462.08	369688.00
jun-22	392.58	278535.00
jul-22	354.50	270065.42
ago-22	344.08	268412.00
sep-22	371.58	183765.78
oct-22	461.42	352468.00
nov-22	459.33	356584.74
dic-22	347.17	256719.88

Tabla 32 Datos de horas de vuelo y consumo de combustible Novena Escuadrilla. Fuente: Arma Aérea Armada

Apéndice B. Cálculo de Contrastes de Heterocedasticidad

B.1. Resultados test Goldfeld y Quandt

n1	Horas vuelo (x_i)	Gasto Combustible (y_i)	$\hat{y}_i$	$\lvert e_i \rvert$	$\lvert e_i \rvert / s$	e_i^2
n1	147.57	98524.00	99465.98	941.98	4.22	887332.90302528
n2	171.25	114017.00	114016.79	0.21	0.00	0.04257588
n3	181.00	127906.00	120007.10	7898.90	35.35	62392636.33512240
n4	211.58	164006.00	138797.20	25208.80	112.83	635483484.06307700
n5	213.17	155405.00	139769.99	15635.01	69.98	244453635.15596800
n6	213.33	170744.00	139872.39	30871.61	138.18	953056596.28427300
n7	219.67	137241.00	143763.52	6522.52	29.19	42543316.78174110
n8	233.25	87653.00	152108.99	64455.99	288.49	4154574902.91253000
n9	239.00	183203.00	155641.74	27561.26	123.36	759623263.04535600
n10	243.25	98724.00	158252.89	59528.89	266.44	3543689332.99194000
n11	252.42	98640.00	163886.85	65246.85	292.03	4257151950.65655000
n12	252.67	184789.00	164038.40	20750.60	92.88	430587252.68146800
n13	256.25	177155.00	166239.97	10915.03	48.85	119137906.44171200
n14	262.50	210549.00	170079.91	40469.09	181.13	1637747396.65953000
n15	263.83	101373.00	170899.10	69526.10	311.19	4833877912.68585000
n16	265.08	202383.00	171667.08	30715.92	137.48	943467553.35800500
n17	267.67	150702.00	173254.26	22552.26	100.94	508604340.53607200
n18	270.75	207944.00	175148.63	32795.37	146.79	1075536420.02899000
n19	270.92	236608.55	175251.03	61357.52	274.62	3764745695.87402000
n20	272.33	197992.05	176121.41	21870.64	97.89	478324775.70596400
n21	278.17	163175.00	179705.36	16530.36	73.99	273252672.64225600
n22	280.83	103492.00	181343.73	77851.73	348.45	6060891897.77520000
n23	282.00	230575.00	182060.52	48514.48	217.14	2353654876.87474000
n24	285.58	202739.00	184262.08	18476.92	82.70	341396420.40251600
n25	293.08	231369.00	188870.01	42498.99	190.22	1806164037.31867000
n26	296.67	212648.00	191071.58	21576.42	96.57	465542048.31168600
n27	298.33	236877.00	192095.56	44781.44	200.43	2005377333.59448000
n28	298.67	224355.00	192300.36	32054.64	143.47	1027500127.94442000
n29	305.42	164135.00	196447.49	32312.49	144.62	1044097116.61344000
n30	309.75	231050.00	199109.85	31940.15	142.96	1020173207.75397000
n31	311.08	229827.00	199929.04	29897.96	133.82	893888212.00243500
n32	312.00	171579.00	200492.23	28913.23	129.41	835974739.63026800
n33	313.50	104952.00	201413.81	96461.81	431.74	9304881406.90805000
n34	313.75	264744.84	201567.41	63177.43	282.77	3991387562.91487000
n35	319.50	116008.00	205100.15	89092.15	398.76	7937412078.80179000
n36	323.92	196300.00	207813.71	11513.71	51.53	132565566.73076000
n37	324.00	110350.00	207864.91	97514.91	436.46	9509157927.61425000
n38	331.97	260986.00	212759.55	48226.45	215.85	2325790094.07048000
n39	333.58	122879.00	213752.82	90873.82	406.73	8258050853.92496000
n40	334.25	186221.00	214162.41	27941.41	125.06	780722495.53719400
n41	337.50	229161.00	216159.18	13001.82	58.19	169047315.53069600
n42	338.50	265145.68	216773.57	48372.11	216.50	2339860968.30630000
n43	338.67	158762.00	216875.97	58113.97	260.11	3377233390.30844000
n44	338.67	248499.00	216875.97	31623.03	141.54	1000016091.05511000
n45	339.00	181936.37	217080.77	35144.40	157.30	1235128552.11312000
n46	344.08	268412.00	220203.92	48208.08	215.77	2324019323.25116000
n47	345.92	269591.00	221330.30	48260.70	216.01	2329095297.66056000
n48	346.00	204465.00	221381.50	16916.50	75.71	286167898.20979000
n49	347.17	256719.88	222098.29	34621.59	154.96	1198654737.18812000
n50	348.92	268536.00	223173.47	45362.53	203.03	2057759172.72748000
n51	350.92	130896.00	224402.25	93506.25	418.52	8743418807.41764000
n52	351.58	207401.00	224811.84	17410.84	77.93	303137475.85709300
n53	354.42	274871.00	226552.62	48318.38	216.26	2334666219.55321000
n54	354.50	270065.42	226603.82	43461.60	194.53	1888911081.01803000
n55	355.25	97626.00	227064.61	129438.61	579.34	16754353252.78720000
n56	362.58	261833.00	231570.14	30262.86	135.45	915840884.23214600
n57	363.25	281274.00	231979.73	49294.27	220.63	2429925014.39417000
n58	364.08	269588.88	232491.72	37097.16	166.04	1376198959.31232000
n59	365.17	221339.00	233157.31	11818.31	52.90	139672494.03883300
n60	365.33	275035.00	233259.71	41775.29	186.98	1745174838.49110000
n61	366.42	274037.53	233925.30	40112.23	179.53	1608991021.28636000
n62	368.17	136384.00	235000.48	98616.48	441.39	9725210659.47412000
n63	369.92	256711.00	236075.67	20635.33	92.36	425817021.09598000
n64	370.58	267629.00	236485.26	31143.74	139.39	969932588.25480400
totales	19340.12	12445739.20	12445739.20			154491999446.07

Tabla.33. Datos muestra n1 para realizar MCO GQ1. Fuente: Elaboración propia con Excel

B.2. Resultados test Goldfeld y Quandt

n12	392.58	278535.00	239356.253	39178.75	139.72	1534974189.8976
n13	392.83	305375.00	239475.575	65899.43	235.02	4342734251.9680
n14	393.83	128803.00	239952.86	111149.86	396.39	12354291434.1755
n15	402.25	321181.00	243970.013	77210.99	275.36	5961536441.1773
n16	402.75	307619.00	244208.656	63410.34	226.14	4020871696.5562
n17	406.33	137241.00	245918.929	108677.93	387.58	11810892335.4162
n18	408.08	326206.00	246754.179	79451.82	283.35	6312591850.1159
n19	409.17	292854.00	247271.238	45582.76	162.56	2077788156.1427
n20	410.08	307151.00	247708.75	59442.25	211.99	3533381070.2320
n21	413.00	306479.18	249100.833	57378.35	204.63	3292274713.3679
n22	413.00	233103.00	249100.833	15997.83	57.05	255930658.2101
n23	413.00	248136.00	249100.833	964.83	3.44	930902.5680
n24	414.25	304987.00	249697.44	55289.56	197.18	3056935463.1783
n25	414.83	158762.00	249975.856	91213.86	325.30	8319967598.4591
n26	415.92	200314.00	250492.916	50178.92	178.95	2517923582.8220
n27	418.08	130152.00	251527.034	121375.03	432.86	14731898968.1947
n28	420.08	211470.69	252481.605	41010.92	146.26	1681895184.4620
n29	421.17	122836.00	252998.665	130162.66	464.20	16942319296.2485
n30	423.58	313038.00	254152.105	58885.90	210.00	3467548654.9481
n31	424.17	304103.00	254430.521	49672.48	177.15	2467355135.5067
n32	424.67	123678.00	254670.755	130992.76	467.16	17159101879.3429
n33	424.67	169093.00	254670.755	85577.76	305.20	7323552161.8500
n34	430.25	329251.00	257334.008	71916.99	256.48	5172053691.5769
n35	432.00	125793.00	258169.258	132376.26	472.09	17523473683.0285
n36	432.50	258001.00	258407.901	406.90	1.45	165568.2357
n37	435.25	339544.00	259720.436	79823.56	284.67	6371801373.1744
n38	435.42	325013.00	259799.984	65213.02	232.57	4252737512.3649
n39	437.63	298696.00	260857.966	37838.03	134.94	1431716779.7087
n40	439.33	178365.00	261667.761	83302.76	297.08	6939349980.6963
n41	441.67	271949.00	262783.018	9165.98	32.69	84015223.5928
n42	443.25	141223.00	263538.72	122315.72	436.21	14961135413.6081
n43	446.00	275142.00	264851.255	10290.74	36.70	105899423.7664
n44	446.58	327523.00	265129.672	62393.33	222.51	3892927379.9906
n45	447.00	347905.00	265328.541	82576.46	294.49	6818871587.1777
n46	449.58	362271.78	266561.529	95710.25	341.33	9160452226.2741
n47	450.22	200196.00	266865.4	66669.40	237.76	4444808945.8252
n48	450.92	355798.00	267197.909	88600.09	315.97	7849976073.7103
n49	453.00	189476.00	268192.254	78716.25	280.73	6196248666.7812
n50	453.83	232710.44	268588.401	35877.96	127.95	1287228095.3060
n51	454.00	167537.00	268669.54	101132.54	360.67	10227790581.4421
n52	457.67	353827.11	270419.587	83407.52	297.46	6956814956.0313
n53	459.33	356584.74	271215.063	85369.68	304.45	7287981835.3539
n54	461.25	156879.00	272129.86	115250.86	411.02	13282760678.5109
n55	461.42	352468.00	272209.407	80258.59	286.23	6441441692.2573
n56	462.08	369688.00	272527.598	97160.40	346.50	9440143772.0716
n57	463.42	348319.00	273163.978	75155.02	268.03	5648277268.2347
n58	463.67	329688.00	273283.3	56404.70	201.16	3181490204.0113
n59	467.00	175439.00	274874.252	99435.25	354.62	9887369255.6486
n60	470.17	277477.13	276387.247	1089.88	3.89	1187845.5939
n61	473.00	165438.00	277737.965	112299.96	400.50	12611282084.6208
n62	473.00	183266.00	277737.965	94471.96	336.92	8924952125.2139
n63	473.17	349781.00	277817.512	71963.49	256.64	5178743555.2827
n64	479.75	371622.00	280959.642	90662.36	323.33	8219663141.9134
	27292.02	16352930.00	16352930.00			383283245389.14

Tabla 34. Datos muestra n2 para realizar MCO GQ2. Fuente: Elaboración propia con Excel

NOTA: Elimino las 4 observaciones centrales

MCO GQ1	
614.3902956	**8802.455544**
115.3753074	35419.14866
0.313834231	49917.99718
28.35717426	62
70660589528	1.54492E+11

$$GQ = \frac{SSNEX n_2}{SSNEX n_1} \sim \frac{\chi^2_{n_2-2}}{\chi^2_{n_1-2}} \sim F_{n_2-2,\, n_1-2.}$$

MCO GQ2	
477.2855306	**51981.90879**
327.6493059	140067.307
0.033092625	78625.61815
2.121964102	62
13117956251	3.83283E+11

GQ=	2.48092618
$F_{n_2-2,\, n_1-2;\alpha.}$	1.51328717
$GQ > F_{n_2-2,\, n_1-2;\alpha.}$	2.48>1.51

Como GQ>F7,7;0.05 rechazamos la hipótesis nula de ausencia de heterocedasticidad

Tabla 35. Resultados operación GQ test en las muestras GQ1 y GQ2. Fuente: Elaboración propia con Excel

B.3. Resultados test Breusch-Pagan-Godfrey

n	Horas vuelo (xi)	Gasto Combustible (yi)	$\hat{y}_i$	|ei|	|ei|/s	e_i^2	xi=zi	$p_i = \hat{e}_i^2/\hat{\sigma}^2$
1	453	189476	271911.373	82435.3726	322.681883	6795590652	453	1.61995034
2	467	175439	279284.171	103845.171	406.487582	10783819462	467	2.57067456
3	473	165438	282443.941	117005.941	458.003601	13690390280	473	3.26355037
4	439.33	178365	264712.362	86347.3619	337.994826	7455866909	439.33	1.77734869
5	443.25	141223	266776.745	125553.745	491.4628	15763742975	443.25	3.75780151
6	461.25	156879	276256.057	119377.057	467.285007	14250881772	461.25	3.39716177
7	424.67	123678	256991.989	133313.989	521.839203	17772619689	424.67	4.23668269
8	233.25	87653	156184.775	68531.7746	268.258169	4696604134	233.25	1.11958855
9	243.25	98724	161451.059	62727.059	245.536411	3934683925	243.25	0.93796005
130	461.4166667	352468	276343.829	76124.1715	297.977558	5794889479	461.4166667	1.38140063
131	459.3333333	356584.74	275246.686	81338.054	318.386581	6615879032	459.3333333	1.57711022
132	347.1666667	256719.88	216176.53	40543.3498	158.701345	1643763214	347.1666667	0.39184449
						5.53732E+11		

MCO	
526.6284321	33348.69285
74.36770128	27828.87182
0.278364768	65264.65459
50.14641492	130
2.13597E+11	5.53732E+11

MCO BPG	
24695706.37	**-4851778116**
4485957.014	1678673948
0.189052377	3936849331
30.3062841	130
4.69711E+20	2.01484E+21

$\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{e}_i^2/n = SSNEX/n.$ **4194937636** ok

$\chi^2 = n\hat{\pi}^2$ **24.9549138** ok — Ho: Ausencia de hetero= Presencia de homo

BP= $\Theta = \frac{SSEX}{2} \sim_{as} \chi^2_{m-1}$

p= **p-value** **5.86868E-07** ok

Nivel signif. $\alpha = 0.05$

Como p<0.05 rechazamos la hipótesis nula de ausencia de heterocedasticidad

BP > χ 2 rechazamos la hipótesis nula de ausencia de heterocedasticidad

En resumen: hay HETEROCEDASTICIDAD

Tabla 36. Resultados de operaciones test de Breusch-Pagan-Godfrey. Fuente: Elaboración propia con Excel

B.4. Resultados test Koenker-Basset

n	Horas vuelo (xi)	Gasto Combustible (yi)	$\hat{y}_i$	$\hat{y}_i^2$	$\|e_i\|$	e_i^2	εi
1	453	189476	271911.373	73935794537	82435.3726	6795590652	285614500.2
2	467	175439	279284.171	77999647962	103845.171	10783819462	3828406849
3	473	165438	282443.941	79774579931	117005.941	13690390280	6540428471
4	439.33	178365	264712.362	70072634548	86347.3619	7455866909	1369329335
5	443.25	141223	266776.745	71169831867	125553.745	15763742975	9556942281
6	461.25	156879	276256.057	76317409107	119377.057	14250881772	7479858319
7	424.67	123678	256991.989	66044882460	133313.989	17772619689	12127561531
8	233.25	87653	156184.775	24393683826	68531.7746	4696604134	3616908109
9	243.25	98724	161451.059	26066444437	62727.059	3934683925	2671637643
129	371.5833333	183765.78	229035.041	52457050036	45269.2611	2049305997	-2106398248
130	461.4166667	352468	276343.829	76365911576	76124.1715	5794889479	-981450299
131	459.3333333	356584.74	275246.686	75760738143	81338.054	6615879032	-94128060.4
132	347.1666667	256719.88	216176.53	46732292202	40543.3498	1643763214	-1884453864

MCO		MCO KB	
526.6284321	**33348.69285**	**0.109609382**	**-1594080578**
74.36770128	27828.87182	0.019646609	1092476604
0.278364768	65264.65459	**0.193176909**	3926825035
50.14641492	130	**31.12578014**	130
2.13597E+11	**5.53732E+11**	4.79958E+20	**2.00459E+21**

Al igual que el test BPG, el test de Koenker-Basset (KB) se basa en los residuos cuadráticos $\hat{e}_{i,i}^2$ pero a diferencia de ser regresados en uno o más regresores, dichos residuos cuadráticos son regresados en los valores estimados cuadráticos del regresando. Es decir, si el modelo original es de la forma

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki} + e_i, \qquad (3.29)$$

al estimarse se obtiene $\hat{e}_i$, y a continuación se estima

$$\hat{e}_i^2 = \delta_0 + \delta_1(\hat{y}_i)^2 + \varepsilon_i, \qquad (3.30)$$

donde $\hat{y}_i$ son los valores estimados del modelo (3.29). Si no se rechaza la hipótesis nula es $\delta_1 = 0$, se podría afirmar que no existe heterocedasticidad. Esta hipótesis puede ser comprobada por el test t o el F test, ya que $F_{1,k}=t_k^2$.

$F_{1,130;0.05}$	**3.913989029**	**punto crítico**
F=	**31.12578014**	
$\alpha = 0.05$		
Ho:	**α1=0**	

Como $F>F_{1,130;0.05}$ rechazamos la hipótesis nula de que α1=0,por lo que hay presencia de heterocedasticidad

Tabla 37 Resultados de operaciones test de Koenker-Basset. Fuente: Elaboración propia con Excel

B.5. Resultados test de White

n	Horas vuelo (xi	to Combustible	Horas vuelo (xi)	(xi)^2	$\hat{y}_i$	$\|e_i\|$	e_i^2	
1	453	189476	453	205209	271911.373	-82435.37258	6795590652	
2	467	175439	467	218089	279284.171	-103845.1706	10783819462	
3	473	165438	473	223729	282443.941	-117005.9412	13690390280	
4	439.33	178365	439.33	193010.849	264712.362	-86347.36191	7455866909	
5	443.25	141223	443.25	196470.563	266776.745	-125553.7454	15763742975	
6	461.25	156879	461.25	212751.563	276256.057	-119377.0571	14250881772	
7	424.67	123678	424.67	180344.609	256991.989	-133313.9891	17772619689	
8	233.25	87653	233.25	54405.5625	156184.775	-68531.77463	4696604134	
9	243.25	98724	243.25	59170.5625	161451.059	-62727.05895	3934683925	
129	371.5833333	183765.78	371.5833333	138074.174	229035.041	-45269.26107	2049305997	
130	461.4166667	352468	461.4166667	212905.34	276343.829	76124.17145	5794889479	
131	459.3333333	356584.74	459.3333333	210987.111	275246.686	81338.05402	6615879032	
132	347.1666667	256719.88	347.1666667	120524.694	216176.53	40543.34982	1643763214	

MCO		MCO AUX		
526.6284321	**33348.69285**	**43062.7668**	**-5009467.32**	**0**
74.36770128	27828.87182	14253.3678	5704584.82	#N/A
0.278364768	65264.65459	**0.58322204**	3925878370	#N/A
50.14641492	130	90.9583437	130	#N/A
2.13597E+11	**5.53732E+11**	2.8038E+21	**2.0036E+21**	#N/A

3.3.6. El contraste general de White (1980)

A diferencia del test Goldfeld-Quandt, el test de White no precisa especificar la forma que puede adoptar la heterocedasticidad, por lo que no depende del supuesto de la normalidad[14].(Damodar N. Gujarati, 2004)

Partiendo de un ejemplo ilustrativo, se considera un modelo de regresión de 3 variables

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + e_i, \qquad (3.31)$$

$\chi^2 =$	$nR^2 \sim_{as} \chi^2_{gl}$	76.98530957
Grados libertad		**2 gl**
χ^2 crítico=0		**1.91802E-17**
Nivel signif.		$\alpha = 0.05$

Como χ^2 > χ^2 crítico, entonces hay HETEROCEDASTICIDAD

Como χ^2 crítico < 0.05 rechazamos la hipótesis nula de ausencia de heterocedasticidad

Tabla 38. Resultados de operaciones test de White. Fuente: Elaboración propia con Excel


CENTRO UNIVERSITARIO DE LA DEFEN
UPCT

Referencias

Alfonso Novales Cinca. (1994). *Econometría* (Segunda edición). Mc Graw-Hill.

Andrés Nortes Checa. (1993). *Estadística Teórica y Aplicada* (Segunda). DM y PPU editores.

Damodar N. Gujarati. (2004). *N-Basic Econometrics. Student solutions manual* (Fourth edition). The Mac Graw-Hill Companies.

Daniel Peña Sánchez de Rivera. (1989). *Modelos y métodos 2. Modelos lineales y series temporales* (Segunda). Alianza Universidad Textos.

Ezequiel Uriel. (2019). *Introducción a la Econometría*. Universidad de Valencia.

Gallego Gómez, Jose Luis. (2008). Capítulo 11. Heterocedasticidad. En *Apuntes de Econometría*. Universidad de Cantabria.

Jan Kmenta. (1986). *Elements of Econometrics* (Second Edition). MACMILLAN PUBLISHING COMPANY.

Jeffrey M. Wooldridge. (2010). *Introducción a la econometría. Un enfoque moderno* (4.ª ed.). CENGAGE Learning.

John O. Rawlings, Pantula, Sastry G., & Dickey, David A. (1998). *Applied Regression Analysis: A Research Tool* (Second edition). Springer.

RStudio. (s. f.). *Https://r-coder.com/ayuda-r/*.

Tomás Baenas Tormo. (2022). *Modelos de Regresión (notas de clase)*. Máster Universitario en Técnicas de Ayuda a la Decisión UPCT-CUD (AGA) Curso 2022-23 (v.1.0).

www.statlect.com. (s. f.). *Www.statlect.com*.

PUBLICACIONES
Pd
DE DEFENSA

GOBIERNO DE ESPAÑA
MINISTERIO DE DEFENSA
SUBSECRETARÍA DE DEFENSA
SECRETARÍA GENERAL TÉCNICA
SUBDIRECCIÓN GENERAL DE PUBLICACIONES Y PATRIMONIO CULTURAL

ISBN 978-84-9091-932-3