Modelo de regresión por mínimos cuadrados ordinarios e identificación de los errores del modelo de acuerdo a los supuestos del Teorema Gauss-Markov.

Chávez Huapeo Jacqueline Flores Ochoa Sofia Libertad
Mendoza Esteban Lizzet López Carmona Audrey Carolina
Rosas Moreno Alesi

MCO Introducción

¿Qué es?

Los mínimos cuadrados ordinarios es una regresión lineal común donde se obtiene estimaciones de parámetros, que describe la relación entre una o más variables cuantitativas independiente y una variable dependiente. Su objetivo es minimizar la suma de las diferencias cuadradas entre los valores observados y estimados, considerando posibles problemas de multicolinealidad.

¿Cómo se estima?

Partimos desde una Función de Regresión muestra con sola una variable explicativa:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 Xi + \hat{u}_i$$

Así $Yi = \hat{Y}_i + \hat{u}_i$ donde $\hat{Y}_i$ es el valor estimado de Y_i. Expresamos los residuos $\hat{u}_i$ como

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 Xi$$

Con la expresión anterior, podemos notar que $\hat{u_i}$ son las diferencias entre los valores observados y los valores esperados de una varable. Por ende, nuestro objetivo es determinar que la $\hat{Y_i}$ sea lo más cercana a Y_i, consecuentemente la suma de los residuos $\sum \hat{u_i} = \sum (Y_i - \hat{Y_i})^2$ debe ser lo menor posible, con $\hat{u_i}^2$ para simplificar y darle más peso a residuos más grandes. Recordemos que $\hat{u_i} = Y_i - \hat{\beta}_1 - \hat{\beta}_2 Xi$ entonces deducimos también que

$$\sum \hat{u_i}^2$$

será en algún tipo de función de $\hat{\beta}_1$ y $\hat{\beta}_2$ Como queremos reducir la suma de los residuos, derivamos e igualamos a para encontrar sus puntos débiles y el punto mínimo

$$\frac{\partial (Y_i - \hat{\beta}_1 - \hat{\beta}_2 Xi)^2}{\partial \hat{\beta}_1} = -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = -2 \sum \hat{u}_i = 0$$

Por lo tanto obtenemos para β_1 (hay que tomar en cuenta que se supone que tenemos n observaciones):

$$\sum Y_i - n\hat{\beta}_1 - \beta_2 \sum X_i = 0$$
$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$$

Dividiendo todo por las n observaciones

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$$

despejando obtenemos que:

$$\hat{\beta}_1 = \bar{Y} + \hat{\beta}_2 \bar{X}$$

en el caso de la $beta_2$ con su derivada:

$$\frac{\partial (Y_i - \hat{\beta}_1 - \hat{\beta}_2 Xi)^2}{\partial \hat{\beta}_2} = -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i = -2 \sum \hat{u}_i X_i = 0$$

desarrollamos:

$$\sum (X_i Y_i - \hat{\beta}_1 X_i - \hat{\beta}_2 X_i^2)$$
$$\sum (X_i Y_i) = \hat{\beta}_1 \sum (X_i) + \hat{\beta}_2 \sum (X_i^2)$$

De manera que reemplazamos con la igualdad de $\hat{\beta}_1$

$$\sum (X_i Y_i) = (\bar{Y} + \hat{\beta}_2 \bar{X}) \sum (X_i) + \hat{\beta}_2 \sum (X_i^2)$$
$$\sum (X_i Y_i) - \bar{Y} \sum (X_i) = \hat{\beta}_2 (\sum (X_i^2) - \bar{X} \sum (X_i)$$
$$\hat{\beta}_2 = \frac{\sum X_i Y_i - \bar{Y} \sum (X_i)}{\sum (X_i^2) - \bar{X} \sum (X_i)}$$

¿Por qué se llama así?

Como podemos observar en el la estimación de la función de regresión, el modelo le debe su nombre por su objetivo, al buscar el modelo de mejor ajuste que minimice la suma de las diferencias al cuadrado entre los valores observados y los valores esperados. El término "ordinarios" se refiere a las condiciones que el modelo aplica. Para que los estimadores (parámetros) sean óptimos y se consideren insesgados y de varianza mínima.

Teorema de Gauss-Markov

Dados los supuestos del modelo clásico de regresión lineal, los estimadores de mínimos cuadrados, dentro de la clase de estimadores lineales insesgados, tienen varianza mínima, es decir,son MELI.

Supuestos del Teorema Gauss-Markov

1. Linealidad en los parámetros

Establece que la relación entre las variables independientes y la variable dependiente es lineal en los parámetros del modelo. Es decir, implica que el modelo de regresión es una combinación lineal de los coeficientes de las variables independientes. Este supuesto es fundamental en la regresión lineal ordinaria, ya que el método MCO busca estimar los coeficientes de manera que la suma de los cuadrados de los residuos sea mínima, asumiendo una relación lineal entre las variables. Si este supuesto no se cumple, es decir, si la relación entre las variables no es lineal, los estimadores obtenidos mediante MCO pueden ser sesgados y poco confiables. La relación lineal implica que un cambio en una unidad en una variable independiente produce un cambio constante en la variable dependiente, manteniendo todas las demás variables independientes constantes.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

2. Los valores de la variable vectorial X son fijos en diferentes muestras.

Establece que los valores de las variables explicativas se consideran fijos en muestras repetidas y son independientes del término de error o perturbación. En otras palabras, la matriz de variables explicativas se compone de un conjunto de números constantes.

3. La media condicional de los residuales ϵ_i es cero

Significa que, en promedio, los errores de la regresión son cero. Esto implica que la regresión no tiene un sesgo hacia arriba o hacia abajo en la predicción de los valores observados. Lo que asegura que no existe correlación entre las variables explicativas y los errores, evitando errores de especificación en el modelo.

$$E(\varepsilon_i \mid X) = 0$$

4. La varianza del modelo es homoscedástica.

Esto indica que la varianza de los errores es constante en todos los niveles de las variables independientes. Lo que significa que la dispersión de la variable dependiente alrededor de la línea de regresión es uniforme a lo largo de todo el rango de las variables explicativas.

$$\mathrm{Var}(u_i) = \sigma^2$$

5. Los errores son no autocorrelacionados

Se supone que los errores son independientes entre sí, de manera que la covarianza entre dos términos de perturbación cualesquiera es cero. Esto garantiza que no existen patrones sistemáticos en los errores a lo largo de las observaciones

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j \mid X) = 0$$

6. Los errores son no correlacionados con las variables explicativas

(Exogeneidad / media condicional cero)

$$E(\varepsilon \mid X) = 0$$

De esto se desprende que, para cada regresor X_k:

$$\mathrm{Cov}(\varepsilon, X_k) = 0$$

7. El número de observaciones debe ser mayor al número de parámetros a estimar

(Identificación y grados de libertad)

Si tienes p parámetros (incluyendo el intercepto) y n observaciones, necesitas al menos $n \geq p$ para que el sistema no esté subdeterminado y, en la práctica, $n > p$ para poder estimar la varianza del error y hacer inferencia:

$$\text{grados de libertad} = n - p > 0$$

- Si $n < p$: hay infinitas soluciones; OLS no es único.
- Si $n = p$: hay solución exacta, pero sin residuos $\Rightarrow$ no puedes estimar σ^2 ni SE.
- Con n apenas mayor que p: errores estándar enormes (inferencia débil).

8. Los valores de cada vector en X no son iguales

(No es constante / no todos sus valores son iguales)

Para cada columna de X, debe haber variabilidad.
Si un regresor es constante (o casi), su efecto no se puede identificar (con un intercepto, una constante adicional es redundante).

- Sin variación: el coeficiente es inencontrable (columna colineal con el intercepto).
- Con variación mínima: el coeficiente tendrá varianza enorme.

9. El modelo está bien especificado

(Forma funcional y variables relevantes/irrelevantes)

La forma funcional es adecuada (lineal en parámetros; puedes incluir X^2, $\log X$, interacciones) y están incluidas las variables relevantes.
Los errores capturan ruido, no estructura sistemática.

Omisión de variables relevantes sesgo de omisión

$$\text{Bias}(\hat{\beta}_1) \approx \beta_2 \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)}$$

10. No hay multicolinealidad perfecta

(Rango completo de X)

Ningún regresor es combinación lineal exacta de otros; $X'X$ debe ser invertible.

- **Perfecta:** MCO no puede calcular $\hat{\beta}$.
- **Alta:** $\hat{\beta}$ sigue insesgado pero con SE enormes, signos inestables y sensibilidad a pequeñas perturbaciones.

Errores del modelo de acuerdo a los supuestos del Teorema Gauss-Markov

Multicolinealidad

La **multicolinealidad** ocurre cuando dos o más variables explicativas de un modelo de regresión están fuertemente correlacionadas entre sí. Esto no sesga los coeficientes, pero aumenta su *varianza*, lo que hace que las estimaciones sean imprecisas y dificulta determinar qué variables realmente influyen en la variable dependiente.

Ejemplo: Un modelo que predice el precio de una casa usando el tamaño en metros cuadrados y el número de habitaciones. Si ambos están fuertemente relacionados, puede ser difícil separar el efecto de cada uno sobre el precio.

Identificación

- **R^2 alta pero coeficientes no significativos**: El modelo explica bien la variable dependiente, pero los coeficientes individuales tienen t-estadísticos pequeños.
- **Altas correlaciones entre regresoras**: Correlación mayor a 0.8 puede ser indicativa de multicolinealidad severa.
- **Regresiones auxiliares (Regla de Klein)**: Cada variable se regresa sobre el resto. R^2 alta sugiere fuerte dependencia lineal.

- **Factor de Inflación de la Varianza (FIV)**: Mide cuánto se inflan los errores estándar por colinealidad. FIV > 10 indica problema serio. TOL, el inverso del FIV, cercano a 0 confirma esto.

Consecuencias y corrección

La multicolinealidad puede inducir errores de interpretación. Aunque los signos de los coeficientes sean correctos, los intervalos de confianza se amplían y las decisiones basadas en significancia estadística pueden ser incorrectas.

Opciones de corrección: - No intervenir si el objetivo es solo predicción y los coeficientes son razonables. - Usar **información a priori** para imponer restricciones. - Eliminar variables altamente correlacionadas, cuidando de no omitir variables relevantes. - Transformaciones como primeras diferencias o variables per cápita. - Incrementar el tamaño de la muestra para reducir la varianza de los estimadores.

Heterocedasticidad

La **heterocedasticidad** viola el supuesto de varianza constante de los errores en un modelo de regresión (homoscedasticidad). Cuando la varianza de los errores cambia según el valor de las variables explicativas, los intervalos de confianza y pruebas t pueden ser incorrectos, llevando a conclusiones erróneas.

Ejemplo: En un estudio sobre ingreso y ahorro, personas con ingresos altos tienden a mostrar mayor variabilidad en su ahorro que personas con ingresos bajos, generando heterocedasticidad.

Identificación

Métodos informales: - Gráficos de residuos al cuadrado vs valores estimados o regresoras. Patrones sistemáticos (cono, abanico) indican heterocedasticidad.

Métodos formales: - Prueba de Park: Regresión de $\ln(\hat{u}_i^2)$ sobre $\ln(X_i)$. - Prueba de Glejser: Regresión de $|\hat{u}_i|$ sobre X_i. - Prueba de White: Regresión de $\hat{u}_i^2$ sobre regresoras originales, sus cuadrados y productos cruzados.

Consecuencias y corrección

Errores estándar incorrectos pueden llevar a rechazar hipótesis verdaderas o aceptar hipótesis falsas.

Opciones de corrección: - Mínimos Cuadrados Ponderados (MCP) si se conoce la estructura de la varianza. - Errores estándar robustos (White) cuando la varianza es desconocida. - Transformaciones logarítmicas o de razón de variables para estabilizar la varianza.

Autocorrelación

La **autocorrelación** ocurre cuando los errores de un modelo de regresión están correlacionados entre sí, especialmente en series de tiempo. Esto viola el supuesto de independencia de los errores y puede dar lugar a estimaciones ineficientes y pruebas de hipótesis inválidas.

Ejemplo: En predicciones de ventas mensuales, un error positivo en un mes suele acompañarse de un error positivo en el mes siguiente, indicando autocorrelación positiva.

Identificación

- Método gráfico: Residuos vs tiempo o residuos rezagados.
- Prueba de las rachas: Analiza secuencias de signos en los residuos.
- Prueba d de Durbin-Watson: $D \approx 2$ indica ausencia de autocorrelación.
- Prueba de Breusch-Godfrey: Permite rezagos de mayor orden y regresoras estocásticas.

Consecuencias y corrección

La autocorrelación puede inflar la significancia aparente de los coeficientes.

Opciones de corrección:

- Revisar especificación del modelo (variables omitidas, forma funcional incorrecta).
- Mínimos Cuadrados Generalizados (MCG) o Factibles (MCGF).
- Método de Newey-West para muestras grandes.
- Conservar MCO si $\rho < 0.3$ en muestras pequeñas.

Error de Especificación

Un **error de especificación** es una violación al **Supuesto 9 del MCRL**, que establece que *"el modelo está correctamente especificado, por lo que no hay sesgo de especificación"*.
Cuando ocurre un error de especificación:
- Se violan los supuestos del MCRL.
- El estimador de **MCO pierde la propiedad de ser MELI** (insesgado y de varianza mínima).

Las consecuencias específicas dependen del **tipo de error de especificación**.

Tipos de Errores de Especificación

Los errores más comunes, que invalidan las propiedades óptimas del estimador MCO, incluyen:

1. Omisión de una Variable Relevante (Subajuste)

Este es el error de especificación más grave. Ocurre cuando se excluye una variable explicativa importante que sí influye en la variable dependiente.

Consecuencias
- Los estimadores de Mínimos Cuadrados Ordinarios (MCO) de los coeficientes incluidos están sesgados e inconsistentes.
- La varianza y los errores estándar de los coeficientes se estiman incorrectamente, invalidando las pruebas de hipótesis habituales (como las pruebas t).

Solución y Detección
1. **Detección por Residuos:** Se debe examinar la gráfica de los residuos (u_i) del modelo inicial. Si existe un error de omisión, los residuos suelen mostrar un patrón distinguible (giros cíclicos o parabólicos), lo que sugiere que la variable omitida está capturada en el término de error.
2. **Detección por Durbin-Watson (d):** Un valor muy bajo de d puede indicar "correlación" positiva en los residuos, que en este contexto refleja un error de especificación (forma funcional incorrecta o variable omitida).
3. **Remedio:** Identificar e incluir la variable relevante omitida o corregir la forma funcional si el error se debe a haber elegido una forma demasiado simple (ejemplo: lineal en lugar de cuadrática o cúbica).

2. Inclusión de una Variable Irrelevante (Sobreajuste)

Ocurre cuando se incluye en el modelo una variable explicativa que no tiene un impacto significativo sobre la variable dependiente.

Consecuencias
- Los estimadores de MCO siguen siendo insesgados y consistentes.
- El principal problema es la **ineficiencia**: las varianzas de los estimadores aumentan, debilitando las pruebas de hipótesis.
- El R^2 aumenta, pero el R^2 ajustado ($\bar{R}^2$) puede disminuir o no mejorar significativamente.

Solución y Detección
1. **Detección:** Realizar pruebas t sobre el coeficiente de la variable sospechosa. Si no es estadísticamente significativo (p-valor alto), la variable es irrelevante.
2. **Prueba F:** Para verificar la significancia conjunta de un grupo de variables irrelevantes, se puede usar la prueba F o comparar el R^2 del modelo restringido (sin variables) con el no restringido (con variables).
3. **Remedio:** Eliminar la variable irrelevante y optar por un modelo más parsimonioso.

3. Adopción de la Forma Funcional Equivocada

Implica seleccionar una relación matemática que no se ajusta a los datos (ejemplo: usar un modelo lineal cuando la verdadera relación es log-lineal o polinomial).

Solución y Detección
- **Detección:** El análisis de residuos es fundamental para identificar patrones que indiquen una mala especificación.
- **Remedio:** Probar formas funcionales alternativas (log-lineales, semi-logarítmicas, recíprocas,

polinomiales) y utilizar criterios de selección como:
- Criterio de Información Akaike (CIA)
- Criterio de Información Schwarz (CIS)
- Criterio Cp de Mallows
- R^2 ajustado ($\bar{R}^2$), que penaliza la inclusión de variables innecesarias

Supuestos que se infringen con los errores

Error de Especificación

Es cuando el modelo que estás usando no representa adecuadamente la realidad o el fenómeno que deseas analizar. Este comete una violación al supuesto 1 de linealidad en los parámetros y a su vez el 9 que nos dice que el modelo está bien especificado. El error de especificación puede violar el supuesto de linealidad en los parámetros directamente, si los parámetros aparecen en formas no lineales, o indirectamente, al usar una forma funcional incorrecta se distorsiona la relación entre variables y parámetros. por ejemplo, si se omite una variable que tiene una relación no lineal con Y, el modelo mal especificado puede inducir una relación compleja y no lineal (implícita) con los parámetros estimados, violando de nuevo la linealidad de forma práctica, aunque no en la estructura de la ecuación como tal.

Error de Multicolinealidad

La multicolinealidad se refiere a la violación del supuesto 10 del modelo clásico de regresión. La principal implicación es ver que quizá el modelo logrado tiene un grado de explicación alto pero la mayoría de las regresoras pueden ser no significativas (diferentes de cero). Este error se da cuando en los modelos de regresión dos o más variables explicativas están altamente correlacionadas entre sí. Es decir, que una variable independiente puede ser explicada en gran parte por otra(s), lo cual genera problemas al estimar los coeficientes del modelo. Si hay multicolinealidad perfecta, pues es una violación directa al supuesto, sin embargo si la multicolinealidad es alta, pero no perfecta, produce coeficientes inestables, errores estándar grandes y mala precisión en las inferencias

Error de Heterocedasticidad

Este error viola directamente el supuesto 4. Pues la heteroscedasticidad ocurre cuando la varianza del término de error no es constante a lo largo de las observaciones. Es decir $Var(\epsilon_i) \neq \sigma^2$ lo que nos indica que los errores (residuos) del modelo tienen una dispersión diferente dependiendo del valor de alguna variable independiente. Y aunque la heteroscedasticidad no sesga los coeficientes estimados por MCO (es decir, siguen siendo insesgados), sí afecta su eficiencia: ya que los estimadores MCO ya no son los de varianza mínima (Pierden la propiedad de ser BLUE), Los errores estándar de los coeficientes son incorrectos (afecta los intervalos de confianza y los tests de hipótesis) y se pueden cometer errores al interpretar significancia estadística (creer que una variable no es significativa cuando sí lo es y viceversa)

Error de Autocorrelación

La autocorrelación se presenta cuando los errores del modelo están correlacionados entre sí a lo largo de las observaciones. Es decir, el error en un periodo está relacionado con el error en otro lo que viola el supuesto 6 del Teorema de Gauss-Markov que nos dice que los errores son no correlacionados con las variables explicativas. Esto ocasiona que los MCO dejen de ser eficientes ya que los errores estándar suelen estar mal estimados lo que afecta directamente la significancia estadística e intervalos de confianza. En pocas palabras, se pierde precisión y confiabilidad en la inferencia estadística.