Examen de programación

Juan Alvaro Díaz Raimond Kedilhac

Examen de programación

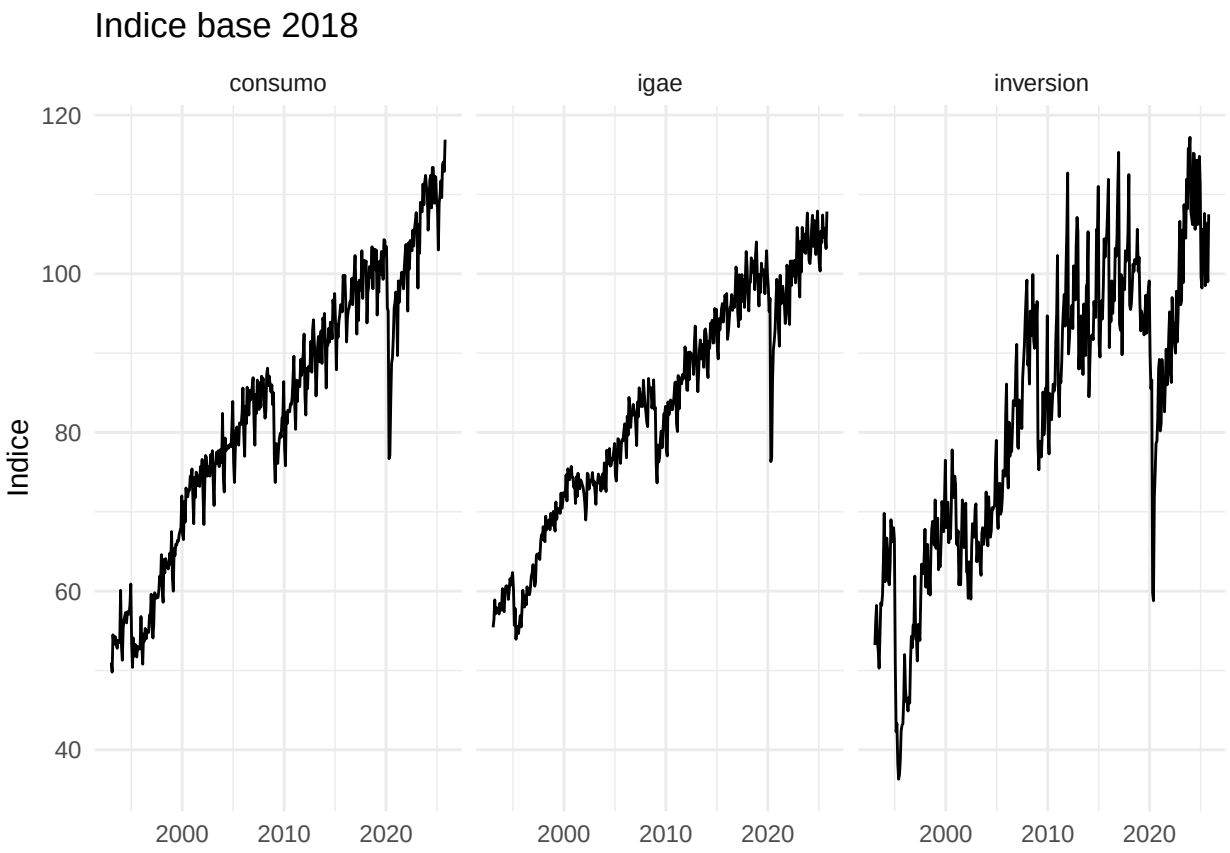
La información teórica fue obtenida del libro *Time Series Econometrics: Learning Through Replication. Springer. (2023)* de Levendis, J. D.

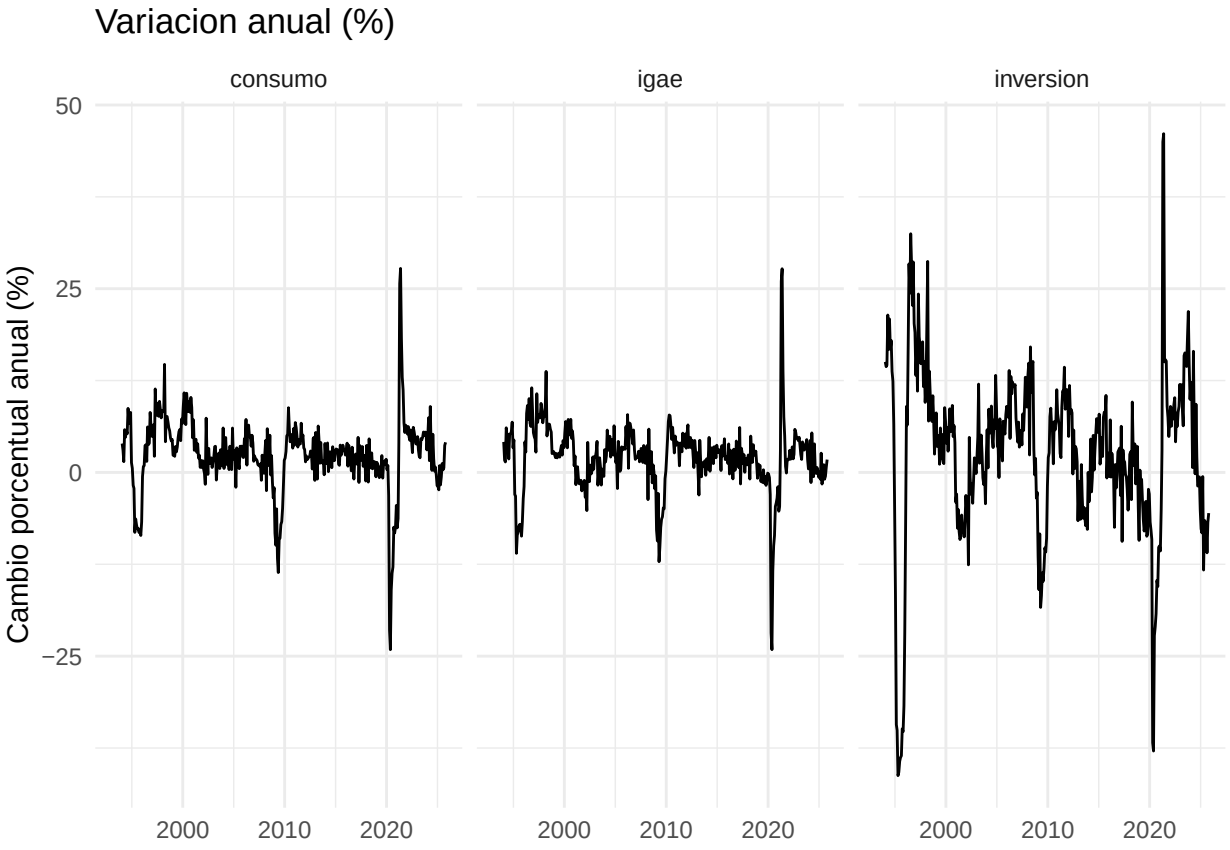
Limpieza de datos

El método preferente para extraer los datos de INEGI es por medio de la API usando el código token y la librería INEGIpy, sin embargo, dado que ha presentado inestabilidad, decidí descargarlas de forma individual y procesarlas con código para unirlas. La base de datos fue limpiada y después exportada para su uso para el script de R, cuya libreria de inegiR también estaba presentando dificultades.

A pesar de estas limitaciones técnicas, toda la información utilizada en el estudio fue obtenida directamente del Banco de Información Económica (BIE), empleando los índices en cifras originales, es decir, sin ajuste estacional.

Análisis exploratorio





Pruebas de estacionariedad

El test de ADF (Dickey-Fuller Aumentado por sus siglas en inglés) es una prueba de cointegración de la forma

$$\Delta y_t = \alpha + \gamma y_{t-1} + \sum_{k=1}^{p-1} \delta_k \Delta y_{t-k} + \varepsilon_t$$

Donde $\gamma = (\sum_{k=1}^{p} \beta_k) - 1$, por lo tanto si $\beta \to 1$ entonces y_t es una serie de raíz unitaria. La H_0 es que el proceso es de raíz unitaria, mientras que H_1 es que la serie no es un random walk. Los términos $\delta_k \Delta y_{t-p}$ se incluyen para ajustar por la autocorrelación. El estadístico sigue un distribución Dickey-Fuller, similar a la t-Student, pero con un ajuste en los valores críticos.

El test KPSS (Kwiatkowski-Phillips-Schmidt-Shin) es una prueba LM (multiplicador de Lagrange) que tiene forma

$$LM = \frac{(\text{score})^2}{\text{information}} = \frac{\sum_{t=0}^{T} S_t^2}{\hat{\sigma}_\varepsilon^2}$$

Partimos de que el modelo se descompone en los componentes: tendencia deterministica (βt), componente estocástico (r_t) y error (ε_t).

$$y_t = \beta t + r_t + \varepsilon_t$$

El componente estocástico evoluciona como:

$$r_t = r_{t-1} + u_t, \quad u_t \sim iid(0, \sigma_u^2)$$

La $H_0 : \sigma_u^2 = 0$, implica que $u_t = 0$ y $r_t = r_0$. El modelo se reduce a

$$y_t = \beta t + r_0 + \varepsilon_t$$

Esto es un modelo estacionario alrededor de una tendencia determinística. Bajo H_1 la serie y_t tiene una tendencia estocástica (caminata aleatoria).

unique_id	adf_stat	p_value	n_lags
consumo	-3.528909	0.0399396	7
igae	-3.647127	0.0285724	7
inversion	-3.030894	0.1418548	7

unique_id	kpss_stat	kpss_p	kpss_lags
consumo	0.5011787	0.01	5
igae	0.3544670	0.01	5
inversion	0.4684660	0.01	5

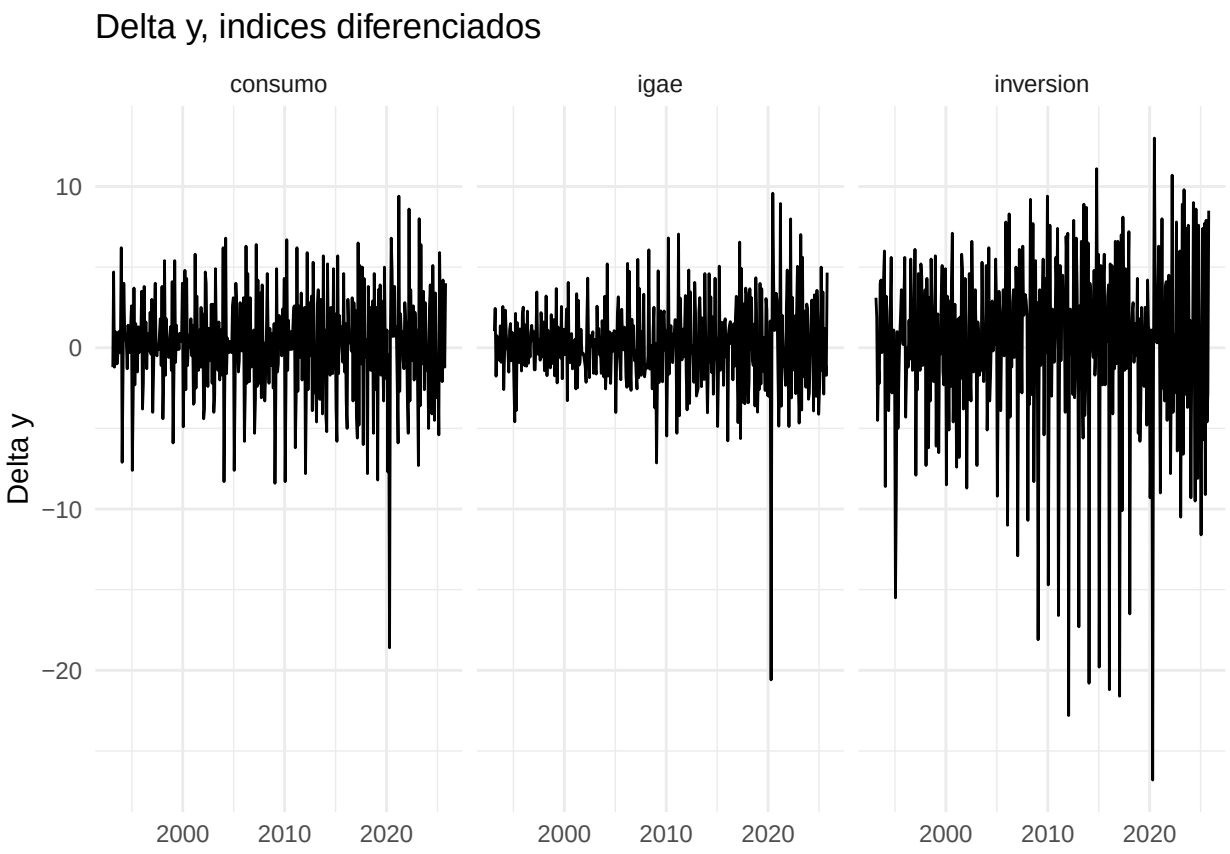
La prueba de Dickey–Fuller Aumentada (ADF) fue implementada tanto en Python como en R, con diferencias en la determinación del número de rezagos. En Python, el número óptimo de rezagos se seleccionó de manera endógena utilizando el criterio de información de Akaike (AIC), el cual penaliza la inclusión de parámetros adicionales y busca un balance entre ajuste y parsimonia. Bajo esta especificación, los resultados del test ADF indican que, para todas las series, no se rechaza la hipótesis nula de presencia de raíz unitaria incluso al nivel de significancia del 1%, lo que sugiere que las series no son estacionarias en nivel.

En contraste, en R la implementación del test ADF utiliza una selección de rezagos fija, lo que conduce a resultados ligeramente distintos. En particular, para las series de consumo y del IGAE se rechaza la hipótesis nula de raíz unitaria al nivel de significancia del 5%, aunque no al 1%, mientras que para la inversión no se encuentra evidencia suficiente para rechazar la hipótesis nula. Estos resultados constituyen evidencia mixta respecto a la estacionariedad en nivel de algunas series.

De manera complementaria, se aplicó la prueba KPSS, cuya hipótesis nula establece que la serie es estacionaria en nivel o alrededor de una tendencia determinística. En todos los casos, el estadístico KPSS permite rechazar la hipótesis nula al nivel de significancia convencional, lo que indica la presencia de tendencias estocásticas en las series.

En conjunto, considerando la sensibilidad de la prueba ADF a la especificación del número de rezagos y la evidencia robusta proporcionada por la prueba KPSS, se concluye que las series presentan comportamiento compatible con procesos integrados de orden uno. Por lo tanto, a efectos del análisis posterior, todas las series se tratan como no estacionarias en nivel.

Primera diferencia



unique_id	adf_stat	p_value	n_lags
consumo	-9.637344	0.01	7
igae	-9.067185	0.01	7
inversion	-9.892597	0.01	7

unique_id	kpss_stat	kpss_p	kpss_lags
consumo	0.0196341	0.1	5
igae	0.0170070	0.1	5
inversion	0.0150415	0.1	5

El mismo procedimiento se aplicó a la primera diferencia de cada serie. A diferencia del análisis en niveles, los resultados del test ADF para las series diferenciadas permiten rechazar la hipótesis nula de presencia de raíz unitaria, lo que indica que las primeras diferencias son estacionarias en nivel.

De manera complementaria, se aplicó la prueba KPSS a las series en primera diferencia. Dado que la diferenciación elimina cualquier componente de tendencia determinística, la prueba se especificó únicamente con constante. En todos los casos, no se rechaza la hipótesis nula de estacionariedad, lo que confirma que las series diferenciadas son estacionarias.

En conjunto, los resultados de las pruebas ADF y KPSS indican que todas las series son integradas de orden uno, ya que no son estacionarias en nivel, pero sí lo son después de aplicar la primera diferencia.

Pruebas de cointegración

Existen diversas metodologías para realizar pruebas de cointegración. En sistemas bivariados es común utilizar el procedimiento de Engle–Granger, el cual consiste en un enfoque de dos etapas. En la primera etapa se estima la relación de largo plazo

$$Y_t = \beta' X_t + \varepsilon_t,$$

y en la segunda se aplica una prueba de raíz unitaria, como el test ADF, a los residuos estimados. Si los residuos resultan estacionarios, se concluye que existe cointegración entre las variables.

Sin embargo, para sistemas con más de dos variables resulta más apropiado emplear el test de Johansen, el cual es un enfoque multivariado basado en un modelo de corrección de error (VECM). Este procedimiento permite determinar el número de relaciones de cointegración mediante dos pruebas distintas: el test del máximo eigenvalor y el test de la traza. Ambos son pruebas de razón de verosimilitud construidas a partir de los eigenvalores de la matriz Π, la cual contiene toda la información sobre las relaciones de largo plazo del sistema.

El modelo VECM puede escribirse como:

$$\Delta Y_t = \sum_{i=1}^{k-1} \Gamma_i \Delta Y_{t-i} + \Pi Y_{t-1} + (\gamma + \tau t) + \varepsilon_t.$$

La matriz Π determina la existencia de cointegración. Si el sistema contiene n variables, puede haber a lo sumo $n-1$ relaciones linealmente independientes de cointegración. Cuando Π no es de rango completo, es decir, es singular, su determinante es igual a cero, lo cual implica que al menos uno de sus eigenvalores es igual a cero. El número de relaciones de cointegración es igual al número de eigenvalores distintos de cero de Π.

Dado que los verdaderos eigenvalores de Π no son observables, el test de Johansen utiliza sus estimaciones muestrales y procede de manera secuencial. En el test del máximo eigenvalor se contrasta la hipótesis nula de que el rango de Π es r frente a la alternativa de que es $r+1$, utilizando el estadístico:

$$LR(r, r+1) = -T \ln(1 - \lambda_{r+1}).$$

El procedimiento inicia con $r = 0$ y se incrementa sucesivamente hasta que no se rechaza la hipótesis nula.

Por su parte, el test de la traza contrasta la hipótesis nula de que el rango de Π es menor o igual a r frente a la alternativa de que es mayor que r, utilizando el estadístico:

$$LR(r, n) = -T \sum_{i=r+1}^{n} \ln(1 - \lambda_i).$$

A diferencia del test de máximo eigenvalor, el test de la traza evalúa de manera conjunta la contribución de todos los eigenvalores restantes, permitiendo identificar el número total de relaciones de cointegración en el sistema.

```
## AIC(n)  HQ(n)  SC(n) FPE(n)
##     12     12     12     12
```

AIC(n)	HQ(n)	SC(n)	FPE(n)
12	12	12	12

```
##
## ######################
## # Johansen-Procedure #
## ######################
```

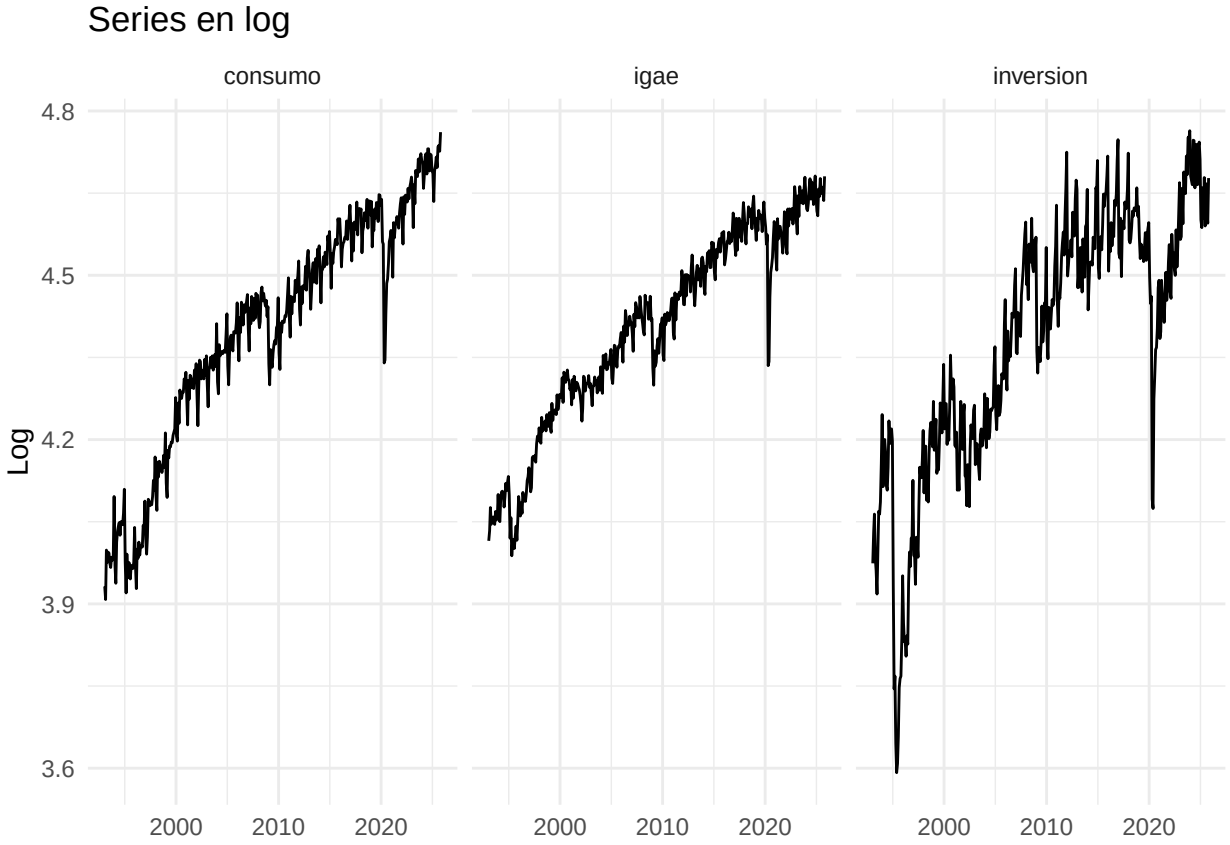
```
##
## Test type: trace statistic , without linear trend and constant in cointegration
##
## Eigenvalues (lambda):
## [1] 4.520103e-02 1.898242e-02 7.729540e-03 1.115185e-16
##
## Values of teststatistic and critical values of test:
##
##           test 10pct  5pct  1pct
## r <= 2 |  2.97  7.52  9.24 12.97
## r <= 1 | 10.31 17.85 19.96 24.60
## r = 0  | 28.03 32.00 34.91 41.07
##
## Eigenvectors, normalised to first column:
## (These are the cointegration relations)
##
##                 igae.l11 consumo.l11 inversion.l11   constant
## igae.l11       1.0000000   1.0000000     1.0000000  1.0000000
## consumo.l11   -0.8939366   0.8557536    -0.7648572 -0.8869817
## inversion.l11 -0.0543606  -1.6315078    -0.0313105 -0.1195202
## constant       1.6866868 -20.1702525   -17.0000306 -0.5322236
##
## Weights W:
## (This is the loading matrix)
##
##               igae.l11 consumo.l11 inversion.l11      constant
## igae.d      0.07153643  0.01228901   -0.02122161 -2.131512e-15
## consumo.d   0.07292571  0.01720804    0.02967870 -7.962503e-15
## inversion.d 0.06654869  0.05333585   -0.04750563  4.966221e-15
##
## ######################
## # Johansen-Procedure #
## ######################
##
## Test type: maximal eigenvalue statistic (lambda max) , without linear trend and constant in cointegra
##
## Eigenvalues (lambda):
## [1] 4.520103e-02 1.898242e-02 7.729540e-03 1.115185e-16
##
## Values of teststatistic and critical values of test:
##
##           test 10pct  5pct  1pct
## r <= 2 |  2.97  7.52  9.24 12.97
## r <= 1 |  7.34 13.75 15.67 20.20
## r = 0  | 17.72 19.77 22.00 26.81
##
## Eigenvectors, normalised to first column:
## (These are the cointegration relations)
##
##                 igae.l11 consumo.l11 inversion.l11   constant
## igae.l11       1.0000000   1.0000000     1.0000000  1.0000000
## consumo.l11   -0.8939366   0.8557536    -0.7648572 -0.8869817
## inversion.l11 -0.0543606  -1.6315078    -0.0313105 -0.1195202
```

```
## constant       1.6866868 -20.1702525   -17.0000306 -0.5322236
##
## Weights W:
## (This is the loading matrix)
##
##               igae.l11 consumo.l11 inversion.l11      constant
## igae.d      0.07153643  0.01228901   -0.02122161 -2.131512e-15
## consumo.d   0.07292571  0.01720804    0.02967870 -7.962503e-15
## inversion.d 0.06654869  0.05333585   -0.04750563  4.966221e-15
```

De acuerdo con los resultados obtenidos a partir de las pruebas de cointegración de Johansen, se concluye que el sistema no presenta cointegración. Tanto el test de la traza como el test del máximo eigenvalor se evalúan de manera secuencial. En ambos casos, no se rechaza la hipótesis nula de ausencia de cointegración para $r = 0$. Dado que no se rechaza para $r = 0$, no es necesario continuar con las siguientes hipótesis nulas.

En consecuencia, no se identifica una combinación lineal estacionaria entre el IGAE, el consumo y la inversión. A pesar de que las series individuales son integradas de orden uno, no comparten una tendencia común de largo plazo.

Transformación logarítmica y regresión




```
## Ajuste de errores estándar HAC

##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.754107   0.073725 10.2286  < 2e-16 ***
```

```
## consumo     0.738154   0.043006 17.1640  < 2e-16 ***
## inversion   0.091926   0.037519  2.4501  0.01472 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Resumen del modelo con SE iid
```

```
##
## Call:
## lm(formula = igae ~ consumo + inversion, data = df_wide_ly)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.071474 -0.016278 -0.001982  0.018409  0.060193
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.75411    0.02523  29.891  < 2e-16 ***
## consumo      0.73815    0.01446  51.050  < 2e-16 ***
## inversion    0.09193    0.01297   7.085 6.52e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02446 on 391 degrees of freedom
## Multiple R-squared:  0.9819, Adjusted R-squared:  0.9818
## F-statistic: 1.062e+04 on 2 and 391 DF,  p-value: < 2.2e-16
```

El modelo de regresión propuesto captura relaciones promedio entre las variables, aunque no es adecuado para realizar inferencia causal ni para fines predictivos. Bajo esta especificación en logaritmos, los coeficientes estimados se interpretan como elasticidades; es decir, miden el cambio porcentual promedio en la variable dependiente ante un cambio de 1% en la variable independiente. En particular, la estimación sugiere que un aumento de 1% en el consumo se asocia, en promedio, con un incremento aproximado de 0.74% en el IGAE, que actúa como proxy del PIB. De manera análoga, un aumento de 1% en la inversión se asocia con un incremento promedio cercano a 0.09% en el IGAE.

Con el fin de corregir posibles problemas de heterocedasticidad y autocorrelación en los residuos, se emplearon errores estándar robustos de tipo HAC (Newey-West). Tras este ajuste, los coeficientes de interés resultaron estadísticamente significativos al nivel del 1%, o cercanos a dicho umbral, de acuerdo con los valores p obtenidos. El uso de errores HAC incrementa la incertidumbre asociada a las estimaciones al ampliar los errores estándar. Asimismo, pueden observarse ligeras discrepancias respecto a estimaciones realizadas en Python, lo cual se explica porque la librería *statsmodels* utiliza aproximaciones asintóticas para la inferencia, mientras que en R se emplea la distribución t de Student, resultando en inferencia más conservadora. Recordemos que el valor p representa la probabilidad, bajo la hipótesis nula, de observar un estadístico al menos tan extremo como el obtenido, y está directamente relacionado con la probabilidad de cometer un error tipo I (α).

El estadístico F rechaza la hipótesis nula de que, conjuntamente, los coeficientes de los regresores (excluyendo la constante) sean iguales a cero, lo que indica que el modelo proporciona un mejor ajuste que una especificación que solo incluya la media del logaritmo del IGAE. El coeficiente de determinación R^2 y su versión ajustada indican una elevada proporción de la varianza explicada, lo cual es consistente con que las series tienen una tendencia positiva a pesar de que no estén cointegradas. No obstante, este buen ajuste no debe interpretarse como evidencia de causalidad ni de capacidad predictiva fuera de muestra, ya que el modelo no incorpora explícitamente la dinámica de corto plazo ni el mecanismo de corrección de error.

Finalmente, los estadísticos de Jarque-Bera y Omnibus se utilizan para evaluar la normalidad de los residuos a partir del sesgo y la curtosis. En este caso, no se rechaza la hipótesis nula de normalidad, lo que sugiere que los residuos no presentan desviaciones significativas respecto a una distribución normal.