218

EL COEFICIENTE DE CORRELACIÓN Y CORRELACIONES ESPÚREAS

Erick Lahura

Enero, 2003

DOCUMENTO DE TRABAJO 218
http://www.pucp.edu.pe/economia/pdf/DDD218.pdf

EL COEFICIENTE DE CORRELACIÓN Y CORRELACIONES ESPÚREAS

Erick Lahura

RESUMEN

En este ensayo se presenta y analiza el ***coeficiente de correlación***, una herramienta estadística elemental e importante para el estudio econométrico de ***relaciones lineales bivariadas*** que involucran el uso de datos de ***corte transversal*** o ***series de tiempo***. En particular, se analiza su relación con las denominadas ***correlaciones espúreas o sin sentido***. Asimismo, se muestran aplicaciones utilizando datos para la economía peruana.

ABSTRACT

An important statistical tool for the econometric study of linear bivariate relationship that involves the use of cross-section or time series data is presented and analyzed in this essay: the ***correlation coefficient***. In particular, its relationship with ***spurious or non-sense correlations*** is analyzed. Likewise, empirical applications based on Peruvian data are shown.

EL COEFICIENTE DE CORRELACIÓN Y CORRELACIONES ESPÚREAS[1]

Erick Lahura[2]

1. INTRODUCCIÓN

La Econometría es el campo de la economía que se ocupa de la medición empírica (estimación, inferencia y predicción) de las relaciones entre variables que establece la teoría económica, a través de la aplicación de métodos estadísticos, matemáticos y computacionales. El propósito fundamental es proporcionar contenido empírico a las relaciones teóricas.

Una manera elemental de llevar a cabo este propósito consiste en analizar relaciones entre dos variables. Si bien es cierto existen muchas relaciones económicas de naturaleza no lineal y/o que involucran más de dos variables (***relaciones multivariadas***), existen otras relaciones relevantes lineales y bivariadas.

Como primer ejemplo, considérese el modelo clásico de demanda por dinero real, que relaciona linealmente la demanda por dinero y el ingreso reales a través de la siguiente ecuación:

$$\left(\frac{M}{P}\right)_t = \beta_1 + \beta_2 Y_t \qquad (1.1)$$

donde $\beta_1 > 0$ y $\beta_2 > 0$. Si se asume que la demanda por dinero real (*M/P*) y el ingreso real (*Y*) pueden ser representados por las series de tiempo *circulante real* y *PBI real*[3] medidas

---

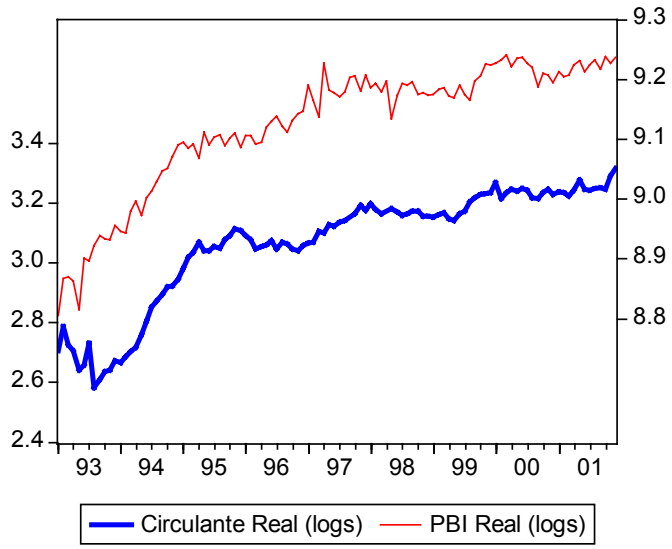
1 Este ensayo forma parte de uno de los capítulos del libro "Econometría Básica: Teoría y Aplicaciones" que actualmente se encuentra en elaboración.

2 Profesor del Departamento de Economía de la Pontificia Universidad Católica del Perú y del Departamento de Ciencias - Sección Matemáticas. El autor agradece el apoyo de Magrith Mena, Ana M. Whittembury y Manuel Barrón por su eficiente labor como asistentes de investigación. Asimismo, agradece a Gisella Chiang, Kristian López, Julio Villavicencio, Luis Orezzoli, Martín Tello, Carla Murguía, Caroline Postigo, Donita Rodriguez y al arbitro anónimo, por sus valiosos comentarios y sugerencias.

3 Más adelante se detallará la forma de obtener cada uno de estos datos.

mensualmente, se obtiene el siguiente gráfico que muestra la evolución de los valores de cada una de ellas (eje vertical) entre enero de 1993 y diciembre de 2001 (eje horizontal):

Figura 1: Gráfico del Circulante y PBI reales
(enero1993-diciembre 2001)



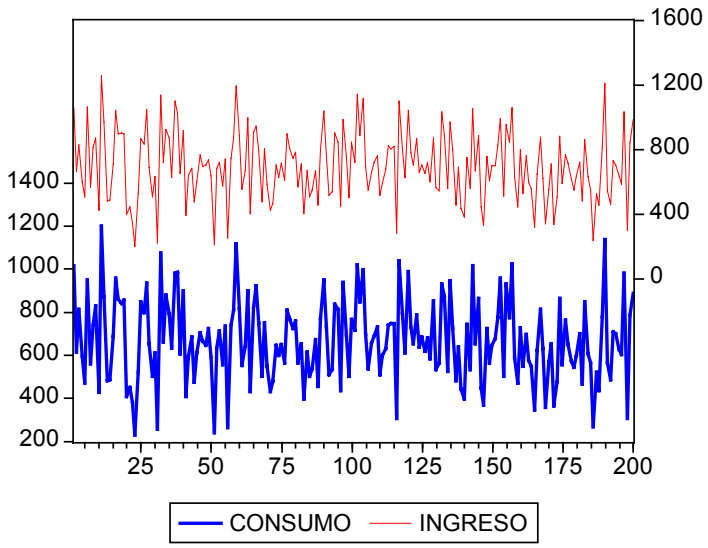
La figura 1 muestra que tanto el circulante como el *PBI* real tienden a crecer a lo largo del período estudiado; es decir, crecen con el tiempo. Esta información, si bien es útil, no es suficiente para dar sustento empírico al modelo teórico de demanda por dinero planteado: no es posible saber exactamente que tan fuerte es la relación entre la demanda por dinero y el ingreso reales.

Como segundo ejemplo, considérese un modelo de consumo de tipo *keynesiano* con el que se intenta explicar el consumo de un grupo de familias representativas de una región para un año determinado:

$$C_i = \beta_1 + \beta_2 Y_i \qquad (1.2)$$

donde $\beta_1 > 0$, $0 < \beta_2 < 1$, C_i es el consumo de la *i-ésima* familia e Y_i el ingreso de la *i-ésima* familia. La figura 2 muestra el comportamiento de las cantidades de consumo e ingreso reales (eje vertical) para 200 observaciones generadas artificialmente (eje horizontal), que constituyen datos de corte transversal:

Figura 2: Gráfico del Consumo e Ingreso
(200 observaciones)



En este caso, a diferencia del modelo de demanda por dinero real, las series no presentan una tendencia clara a crecer o decrecer, sino más bien parecen revertir a un valor promedio constante a lo largo de todas las observaciones. De este modo, no es posible concluir fácilmente a partir del gráfico si las series se mueven juntas, en el mismo sentido o en sentidos opuestos.

De esta manera, tanto en el contexto del modelo de demanda por dinero (series de tiempo) como en el modelo de consumo (corte transversal), se hace necesario un instrumento que permita determinar la fuerza y el sentido de la posible relación lineal existente entre los pares de variables mencionados. Éste se denomina ***coeficiente de correlación***.[4]

El coeficiente de correlación es una herramienta estadística elemental e importante para el estudio econométrico de ***relaciones lineales bivariadas*** que involucran el uso de datos de ***corte transversal*** o ***series de tiempo.*** Sin embargo, este instrumento puede fallar en algunas ocasiones al sugerir la presencia de una relación estadísticamente significativa entre dos variables que en verdad no tienen sentido o no poseen relación lineal alguna, es decir, que presentan una ***correlación espúrea***.

[4] El coeficiente de correlación es solamente uno de los estadísticos que existen para medir el grado de asociación entre variables, lo cual depende de la clase de variables analizadas (categórica, continua, etc.). Una referencia más amplia de los diversos estadísticos existentes es Liebetrau (1983).

En este ensayo se analiza el ***coeficiente de correlación*** y su relación con las denominadas ***correlaciones espúreas o sin sentido***. En la sección 2 se examina estadísticamente el ***coeficiente de correlación***. En la sección 3 se define el problema de ***correlaciones espúreas*** o ***sin sentido***. En la sección 4 se discute la presencia de correlaciones espúreas en un ***contexto de corte transversal***. En la sección 5 se analiza la presencia de correlaciones espúreas en un ***contexto de series de tiempo***. Finalmente, en la sección 6 se presentan aplicaciones utilizando datos de la economía peruana.

2. EL COEFICIENTE DE CORRELACION

El coeficiente de correlación es un estadístico que proporciona información sobre la ***relación lineal*** existente entre ***dos variables*** cualesquiera. Básicamente, esta información se refiere a dos características de la relación lineal: la ***dirección*** o ***sentido*** y la ***cercanía*** o ***fuerza***.

Es importante notar que el uso del coeficiente de correlación sólo tiene sentido si la relación bivariada a analizar es del tipo ***lineal***. Si ésta no fuera ***no lineal***, el coeficiente de correlación ***sólo indicaría la ausencia de una relación lineal más no la ausencia de relación alguna***. Debido a esto, muchas veces el coeficiente de correlación se define - de manera más general - como un instrumento estadístico que mide el ***grado de asociación lineal*** entre dos variables.

2.1. Desviaciones y gráfico de dispersión

Sea una ***muestra de n observaciones*** o ***muestra de tamaño n*** para dos variables *X e Y*, denotada por:

$$M = [(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)] \qquad (2.1)$$

donde cada par (X_i, Y_i) representa los valores de cada variable para la *i-ésima* observación, con $i = 1, 2, \ldots, n$. Asimismo, sea X_i la *i-ésima* observación de la variable *X* y $\overline{X}$ el valor promedio de las *n* observaciones de la misma. Con esto, se define la ***desviación de la i-ésima observación de la variable X respecto de su valor promedio observado***, o simplemente ***desviación de X_i***, como:

$$x_i = X_i - \overline{X} \tag{2.2}$$

La variable x_i puede tomar valores positivos o negativos dependiendo del valor de cada observación, es decir, si es mayor o menor que el valor promedio observado. Cuando $x_i > 0$ se dice que la desviación de la variable X_i es positiva, mientras que si $x_i < 0$ se dice que la desviación es negativa. De manera análoga, se define la ***desviación de Y_i*** como:

$$y_i = Y_i - \overline{Y} \tag{2.3}$$

De esta forma, es posible escribir la muestra en términos de desviaciones como:

$$m = [(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)] \tag{2.4}$$

El gráfico de todos los pares de observaciones (X_i, Y_i) en el plano *X*-*Y* se denomina ***gráfico de dispersión*.** La Figura 2a representa el gráfico de dispersión de las variables *X* e *Y* para *n=2* observaciones. El plano se ha dividido en cuatro cuadrantes determinados por el punto *O*, el cual representa el par ordenado de los valores promedios de las observaciones de las variables *X* e *Y*, $(\overline{X}, \overline{Y})$. Nótese que los valores promedio no necesariamente son observaciones de la muestra, sino simplemente un par ordenado que sirve como referencia.

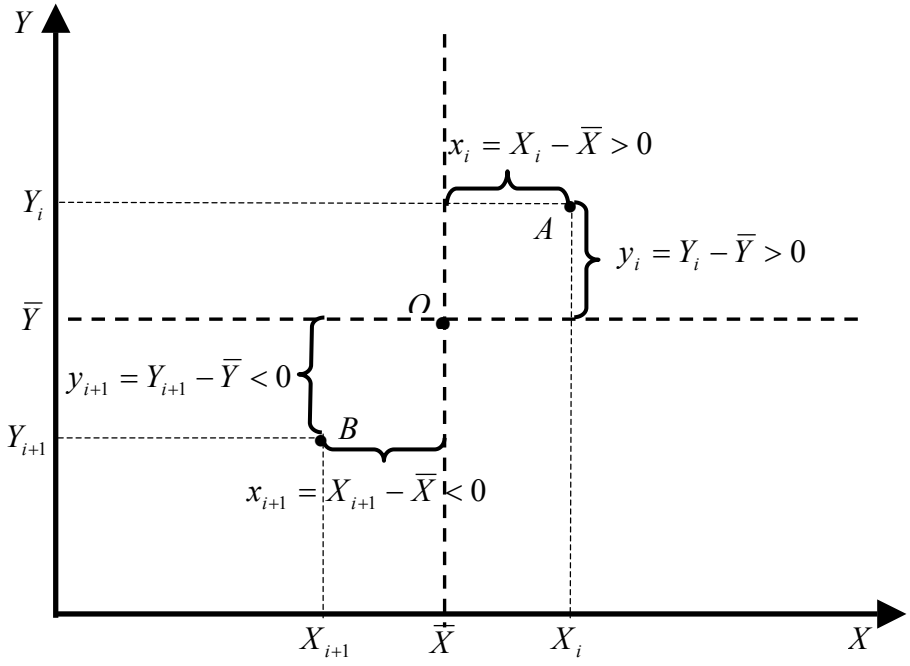


Figura 2a: Desviaciones de X e Y en la misma dirección.

El punto *A*, situado en el primer cuadrante de la Figura 2a, representa los valores de las variables *X* e *Y* correspondientes a la *i-ésima* observación de la muestra. En este punto, el valor de cada variable es mayor que sus correspondientes valores promedios, es decir, las ***desviaciones de ambas variables son positivas***. De esta forma, las variables *X* e *Y* varían conjuntamente y en el mismo sentido, es decir, ***covarían positivamente***. En este caso, se dice que existe una ***relación lineal y positiva*** entre ambas variables.

El punto *B*, situado en el tercer cuadrante de la Figura 2a, representa los valores de las variables *X* e *Y* correspondientes la (*i+1*)-*ésima* observación de la muestra. En este punto, las ***desviaciones de ambas variables son negativas***. Así, se tiene que *X* e *Y* varían conjuntamente y en el mismo sentido, es decir, ***covarían positivamente***. En este caso, se dice que un punto como *B* implica la existencia de una ***relación lineal y positiva*** entre estas variables.

Si la relación entre las variables *X* e *Y* estuviera representada sólo por las dos observaciones de la Figura 2a (puntos *A* y *B*), se dice que la relación entre estas variables es ***lineal y positiva*** o que las variables ***covarían positivamente***.

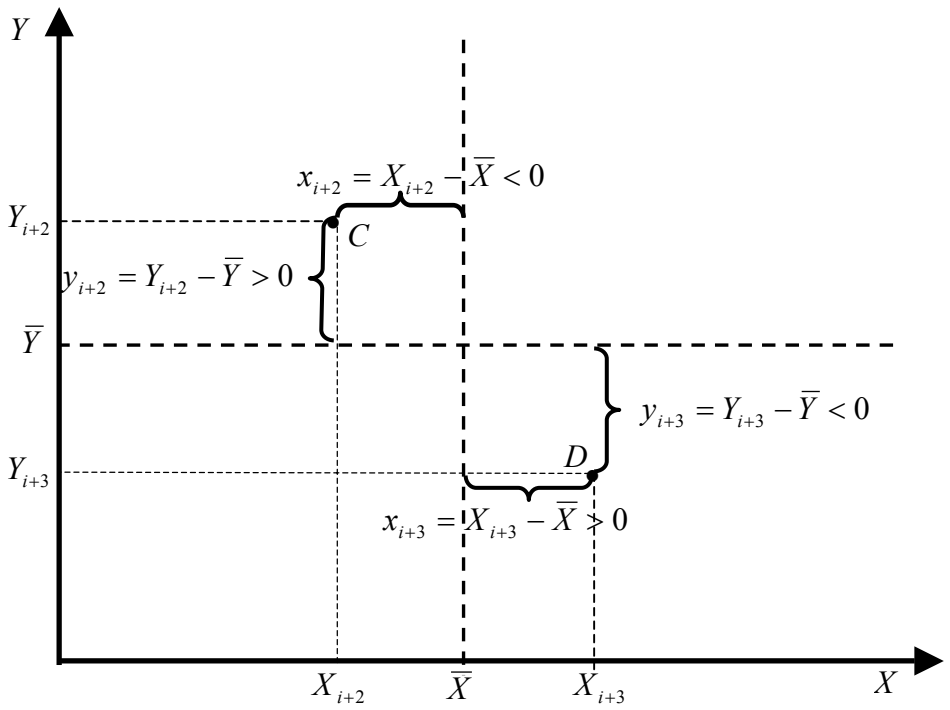


Figura 2b: Desviaciones de X e Y en direcciones opuestas.

El punto *C*, situado en el segundo cuadrante de la Figura 2b, representa los valores de las variables *X* e *Y* correspondientes a la (*i+2*)-*ésima* observación de la muestra. En este punto, las ***desviación de X es negativa*** y la ***desviación de Y es positiva***. Así, las variables *X* e *Y* varían conjuntamente y en sentidos opuestos; es decir, ***covarían negativamente***. En este caso, se dice que existe una ***relación lineal negativa*** entre ambas variables.

El punto *D*, situado en el cuarto cuadrante de la Figura 2b, representa los valores de *X* e *Y* correspondientes a la (*i+3*)-*ésima* observación de la muestra. De manera análoga al caso anterior, el análisis de los signos de las desviaciones permite afirmar que existe una ***relación lineal negativa*** entre *X* e *Y*.

Si las variables *X* e *Y* están representadas sólo por las dos observaciones de la Figura 2 (puntos *C* y *D*), entonces la relación entre estas variables sería ***lineal y negativa***.

Para el caso de un gráfico en el cual las variables *X* e *Y* estuvieran representadas por cuatro observaciones iguales a los puntos *A, B, C* y *D,* tales que las desviaciones positivas y negativas se compensaran entre sí, entonces se concluye que ***no existe relación lineal*** entre las variables.

2.2. La Covarianza Muestral

Si los *n* pares de observaciones se ubicaran en el primer y tercer cuadrante (es decir, si la relación entre *X e Y* fuera positiva), la multiplicación de sus desviaciones, $x_i y_i$, tendría signo positivo. Por lo tanto, la suma de las *n* desviaciones también sería positiva:

$$\sum_{i=1}^{n} x_i y_i > 0 , \qquad (2.5)$$

De esta forma, el signo de (2.5) indicaría que la dirección o sentido de la relación es positiva. Si se trazara una línea tal que represente aproximadamente la distribución de los pares ordenados, el signo de (2.5) indicaría el signo de la pendiente de esa línea, como se muestra en Figura 3.

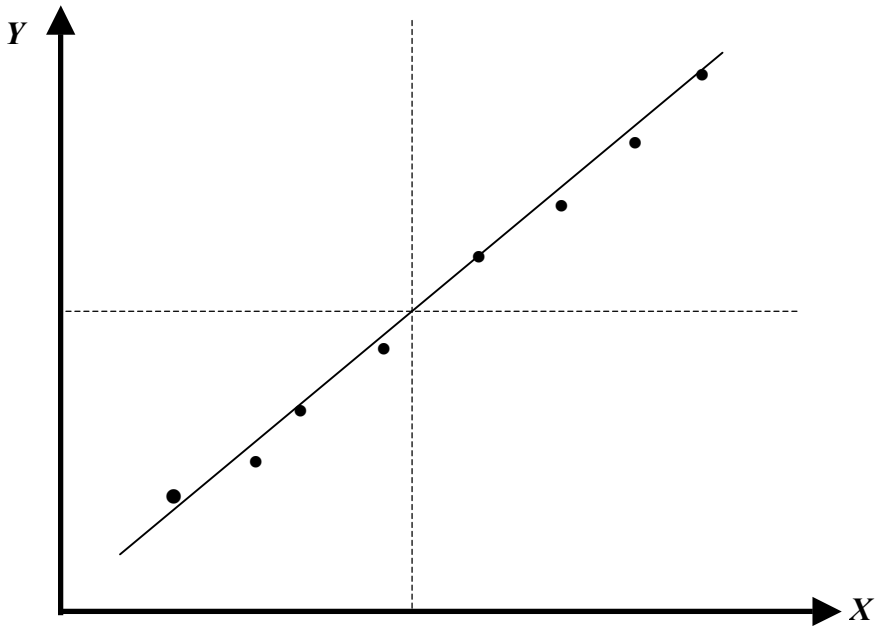


Figura 3: Relación lineal y positiva entre las variables X e Y .

A través de un análisis análogo al anterior, se tiene que si los *n* pares de observaciones se ubicaran en el segundo y cuarto cuadrante (es decir, si la relación lineal entre las variables *X e Y* fuera negativa), tendríamos:

$$\sum_{i=1}^{n} x_i y_i < 0 \qquad (2.6)$$

Si las *n* observaciones de las variables se ubicaran en los cuatro cuadrantes de manera simétrica respecto de sus valores promedios (es decir, si no existiera relación lineal alguna entre las variables), la suma del producto de cada par de desviaciones tomaría valores positivos o negativos muy cercanos a cero:

$$\sum_{i=1}^{n} x_i y_i \approx 0 \qquad (2.7)$$

Dadas las desigualdades (2.5), (2.6) y (2.7) y una muestra de $n = n_1 + n_2$ observaciones para un par de variables X *e* Y cuyo gráfico de dispersión consta de n_1 puntos ubicados en el primer y tercer cuadrante:

$$\sum_{n_1} x_i y_i > 0$$

y n_2 puntos ubicados en el segundo y cuarto cuadrante:

$$\sum_{n_2} x_i y_i < 0$$

es posible establecer las siguientes afirmaciones:

(1) Si los puntos ubicados en el primer y tercer cuadrante son más importantes que los ubicados en el segundo y cuarto cuadrante, es decir:

$$\left|\sum_{n_1} x_i y_i\right| - \left|\sum_{n_2} x_i y_i\right| > 0$$

o equivalentemente:

$$\sum_{n_1} x_i y_i + \sum_{n_2} x_i y_i > 0$$

entonces la relación lineal predominante entre las variables es ***positiva***. En este caso, se dice que las variables X *e* Y ***covarían lineal y positivamente***.

(2) Si los puntos ubicados en el segundo y cuarto cuadrante son más importantes que los ubicados en el primer y tercer cuadrante, es decir:

$$\left|\sum_{n_1} x_i y_i\right| - \left|\sum_{n_2} x_i y_i\right| < 0$$

o equivalentemente:

$$\sum_{n_1} x_i y_i + \sum_{n_2} x_i y_i < 0$$

entonces la relación lineal predominante entre las variables es ***negativa***. En este caso, se dice que las variables X *e* Y ***covarían lineal*** y ***negativamente***.

(3) Si los puntos ubicados en el segundo y cuarto cuadrante son tan importantes como los ubicados en el primer y tercer cuadrante, es decir, si:

$$\left|\sum_{n_1} x_i y_i\right| - \left|\sum_{n_2} x_i y_i\right| \approx 0$$

o equivalentemente:

$$\sum_{n_1} x_i y_i + \sum_{n_2} x_i y_i \approx 0$$

entonces no predomina ningún tipo de ***relación lineal*** entre las variables; es decir, ***no covarían linealmente.*** Sin embargo, esto último no implica que no pueda existir algún tipo de relación no lineal entre las variables.

Estas tres afirmaciones se resumen en la Figura 4:

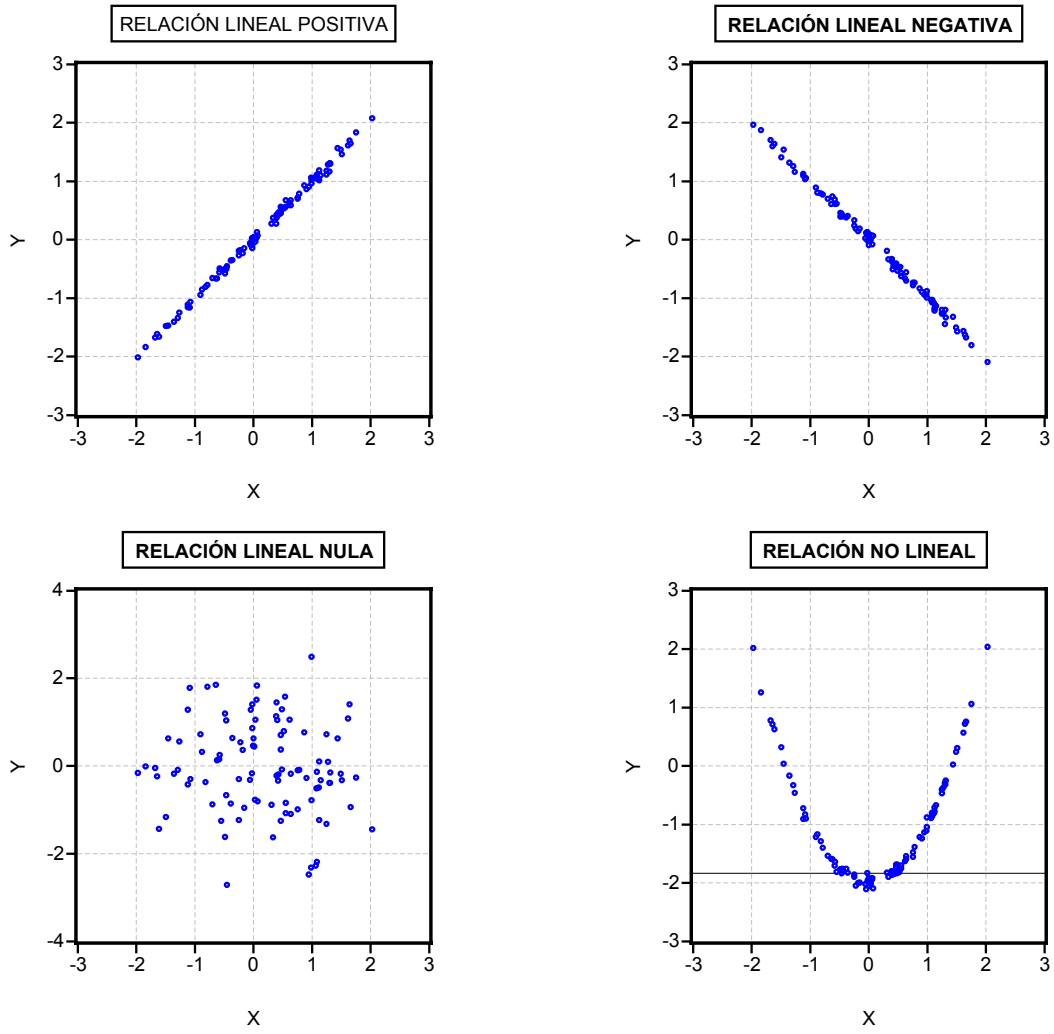


Figura 4: Relaciones Lineales y No Lineales

Si el número de observaciones, o tamaño de muestra, fuera muy grande y si las variables presentaran algún tipo de comovimiento lineal (positivo o negativo), la expresión $\sum_{i=1}^{n} x_i y_i$ crecería con el tamaño de muestra. Debido a esto, es mejor promediar $\sum_{i=1}^{n} x_i y_i$ considerando la información que proporciona cada desviación, obteniéndose de esta manera el estadístico conocido como ***covarianza muestral***:

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i \qquad (2.8)$$

El promedio de la suma de desviaciones se obtiene a través de un factor igual a (n-1) porque basta con la información (valor) de las primeras *n-1* desviaciones para conocer la información (valor) incorporada en la última desviación. Por ejemplo, si *n* fuera igual a 2:

$$\frac{(X_1 + X_2)}{2} = \overline{X}$$

Entonces, bastaría con conocer la primera desviación (el valor de X_1 y $\overline{X}$) para conocer el valor de X_2. Dado que (2.8) depende de $\sum_{i=1}^{n} x_i y_i$, el análisis precedente implica que la ***covarianza muestral*** permite ***identificar*** la ***dirección*** o ***sentido*** de la relación lineal entre las variables, a través de su signo. Esta es la ***única información relevante*** que proporciona la covarianza muestral para el análisis de la relación entre dos variables.

2.3. El coeficiente de correlación

Intuitivamente, la fuerza o cercanía de la relación entre dos variables podría medirse a través de la covarianza muestral: mientras más grande sea el valor de la covarianza muestral, más fuerte será la relación entre las variables. Sin embargo, los valores que puede tomar la covarianza muestral dependen de las unidades de medida de las variables involucradas, lo cual podría conducir a interpretaciones equivocadas acerca de la fuerza de la relación.

Para ilustrar este problema, considérese las variables ***X = tasa de interés activa*** e ***Y= tasa de interés pasiva***, para las cuales se cuenta con una muestra ficticia de 10 observaciones:

$$M = [(10,5);(20,10);(30,15);(40,20);(50,25);(60,30);(70,35);(80,40);(90,45);(100,50)]$$

donde las variables están expresadas como porcentajes en la escala del 0 al 100 (por ejemplo, "10" representa "10 por ciento"). La covarianza muestral entre *X* e *Y*, dados estos valores muestrales, es igual a:

$$Cov(X,Y) = 412{,}5$$

Este resultado indica que existe una relación lineal positiva entre la tasa de interés activa y pasiva. Si se divide todos los valores de la muestra por 100, se obtiene la siguiente muestra:

$$M' = [(0{,}10\ ;\ 0{,}05); (0{,}20;\ 0{,}10); (0{,}30\ ;\ 0{,}15); (0{,}40\ ;\ 0{,}20); (0{,}50\ ;\ 0{,}25),$$
$$0{,}60\ ;\ 0{,}30); (0{,}70\ ;\ 0{,}35); (0{,}80\ ;\ 0{,}40); (0{,}90\ ;\ 0{,}45); (1{,}00\ ;\ 0{,}50)]$$

donde las variables están expresadas como porcentajes en la escala del 0 al 1 (por ejemplo, "0,10" representa "10 por ciento"). En este caso, la covarianza muestral entre X e Y, dados estos valores muestrales, es igual a:

$$Cov(X,Y) = 0{,}04125$$

Este resultado confirma que la relación entre las tasas de interés activa y pasiva es positiva. Así, el ***sentido*** de una ***relación lineal*** medido por la ***covarianza muestral*** es ***invariante a cambios en las unidades de medida***.

Sin embargo, luego de reducir la escala de las variables, el valor de la covarianza disminuye (se hace prácticamente cero) respecto del caso original. De esta forma, si se utilizara el valor absoluto de la covarianza para medir la fuerza de la relación lineal entre las variables, se podrían obtener conclusiones equivocadas: en el primer caso se afirmaría que la relación es muy fuerte, mientras que en el segundo caso que la relación es muy débil, lo cual es inconsistente pues se está analizando la misma relación en ambos casos. Así, la ***fuerza*** de una ***relación lineal*** medida por la ***covarianza muestral*** es ***sensible a cambios en las unidades de medida***[5].

Para obtener un indicador de la fuerza de la relación lineal entre dos variables que no dependa de las unidades de medida de las mismas, se deberá expresar las desviaciones en ***unidades de desviación estándar***. La ***covarianza muestral estandarizada*** se denomina ***coeficiente de correlación muestral***, y se denota usualmente como ***r***:

$$Corr(X,Y) \equiv r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i}{S_X}\right)\left(\frac{y_i}{S_Y}\right) \qquad (2.9)$$

[5] La demostración matemática de este resultado se muestra en el apéndice (demostración 1).

donde:

$$S_X = \sqrt{\frac{\sum {x_i}^2}{n-1}} \qquad (2.9)'$$

$$S_Y = \sqrt{\frac{\sum {y_i}^2}{n-1}} \qquad (2.9)''$$

Es fácil observar que el coeficiente de correlación muestral no es otra cosa que el cociente entre la covarianza muestral y los desvíos estándar muestrales de cada variable:

$$r = \frac{Cov(X,Y)}{S_X S_Y} \qquad (2.10)$$

Alternativamente, el coeficiente de correlación puede ser expresado como:

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \qquad (2.11)$$

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{\sqrt{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}\sqrt{n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2}} \qquad (2.12)$$

2.4. Interpretación del Coeficiente de Correlación

El coeficiente de correlación muestral, además de ser independiente de las unidades de medida de las variables, se caracteriza por tomar valores dentro del intervalo cerrado [-1,1][6]:

$$-1 \le r \le 1$$

6 La demostración matemática se presenta en el apéndice (demostración 2).

o equivalentemente:

$$| r | \leq 1$$

La ***interpretación*** del coeficiente de correlación muestral depende del valor y del signo que tome y de las características de la muestra analizada. Para los propósitos del presente ensayo, se asume que el número de observaciones de la muestra (tamaño de muestra), es tal que la ***muestra es representativa***: presenta las mismas características de la población. De esta manera, las conclusiones que puedan extraerse a partir del análisis del coeficiente de correlación serán válidas para la relación poblacional.

A partir de la expresión (2.9), dado que S_X y S_Y solamente pueden tomar valores no negativos, se tiene que el signo del coeficiente de correlación muestral dependerá del signo de la suma del producto de todas las desviaciones, $\sum_{i=1}^{n} x_i y_i$. Así, el ***signo*** de ***r*** indica la ***dirección de la relación lineal*** (al igual que la covarianza muestral): valores positivos indican una relación directa y valores negativos una relación inversa entre las variables involucradas.

Por otro lado, el ***valor absoluto*** del coeficiente de correlación indica ***la fuerza de la relación lineal***. Un coeficiente de correlación muy cercano a uno en valor absoluto indica que la relación entre las variables es muy fuerte, mientras que si es muy cercano a cero, indica que la relación es muy débil. El cuadro 1 muestra las posibles interpretaciones del coeficiente de correlación muestral.

Cuadro 1: Interpretación del Coeficiente de Correlación Muestral

VALOR DEL COEFICIENTE	*INTERPRETACIÓN*
$0 < r < 1 \;\; y \;\; r \to 1$	*relación lineal positiva y fuerte.*
$0 < r < 1 \;\; y \;\; r \to 0$	*relación lineal positiva y débil.*
$r = 0$	*no existe relación lineal.*
$-1 < r < 0 \;\; y \;\; r \to -1$	*relación lineal negativa y fuerte.*
$-1 < r < 0 \;\; y \;\; r \to 0$	*relación lineal negativa y débil.*

El coeficiente de correlación muestral, a diferencia de la covarianza muestral, no solamente mide el sentido de la relación entre las variables sino también la ***fuerza*** de la ***relación lineal*** o ***grado de asociación lineal***. La figura 5 relaciona el grado de asociación lineal con diversos valores del coeficiente de correlación muestral ***r***.

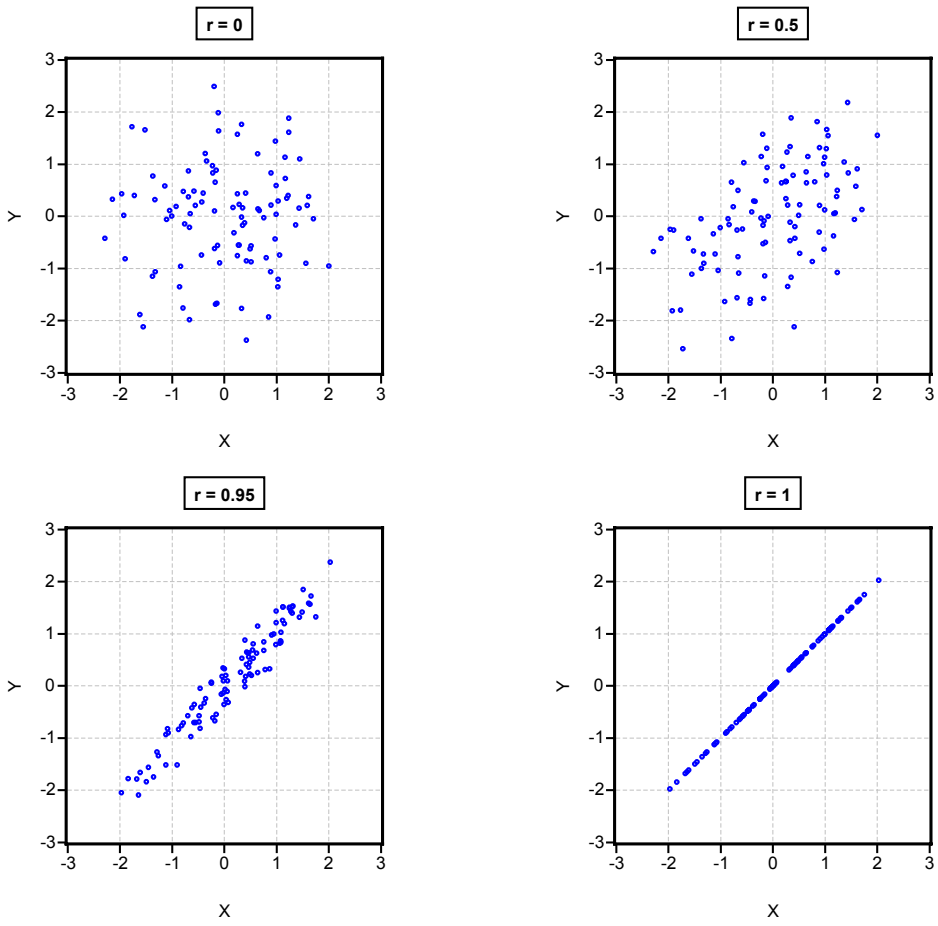


Figura 5: Coeficiente de Correlación Muestral y Grado de Asociación Lineal

Es importante observar que un coeficiente de correlación bajo no significa que ***no existe relación alguna*** entre las variables, sino simplemente que ***no existe relación lineal*** entre ellas.

A partir del análisis precedente sobre las desviaciones muestrales de las variables, se puede concluir que si la relación es ***no lineal***, la expresión $\sum_{i=1}^{n} x_i y_i$ puede tener un valor muy cercano a cero, lo cual implica que el coeficiente de correlación muestral "*r*" también tendrá un valor muy cercano a cero. Así, el ***coeficiente de correlación muestral no***

proporciona información adecuada sobre la existencia de una relación no lineal entre dos variables.

Como nota adicional, es importante saber que el coeficiente de correlación ***no proporciona información sobre la causalidad entre las series***. Lo único que permite identificar es ***co-movimientos significativos***. Existen pruebas estadísticas que permiten determinar en cierta medida la causalidad entre variables, como por ejemplo la prueba de causalidad a la Granger (1969). Sin embargo, a este nivel básico de econometría la única forma de determinar causalidad será a través de la teoría económica.

2.5. Uso del coeficiente de correlación: Un modelo simulado

Para finalizar, considérese el caso de la relación entre el consumo y el ingreso de una muestra simulada de 200 familias representativas de una región, que se presentó en la sección 1. La simulación se realizó de tal forma que exista una relación lineal significativa entre el consumo y el ingreso, como se muestra en la Figura 6.

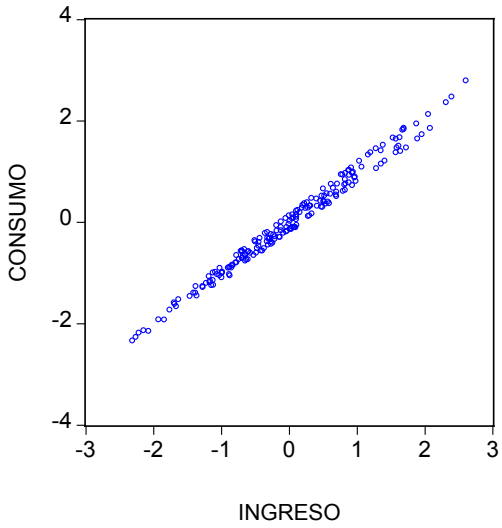


Figura 6: Gráfico de Dispersión entre Consumo e Ingreso

Al aplicar la fórmula (2.11), se obtiene un coeficiente de correlación igual a 0.99, de donde se deduce que existe una ***fuerte relación lineal positiva*** entre el consumo y el ingreso. Usualmente, los paquetes econométricos permiten mostrar el coeficiente de correlación a

través de una matriz de correlaciones, donde los elementos de la diagonal son siempre iguales a 1 (pues muestran la correlación entre cada variable consigo misma) y los que están fuera de la diagonal miden la correlación entre cada par de variables.

Matriz de Correlaciones: Consumo e Ingreso

	INGRESO	CONSUMO
INGRESO	1.000000	0.993350
CONSUMO	0.993350	1.000000

3. CORRELACIONES ESPÚREAS

En esta sección, se define el concepto de correlación espúrea. Además, se presentan las características más importantes que pueden presentar los datos empíricos utilizados para representar las variables económicas, las cuales serán elementos importantes para analizar las causas de la presencia de correlaciones espúreas

3.1. Definición de Correlaciones Espúreas

El coeficiente de correlación muestral permite establecer ***estadísticamente*** el grado de asociación lineal entre dos variables a partir de una muestra o conjunto de observaciones representativas para cada una de ellas. Esto significa que el coeficiente de correlación permite establecer la ***fuerza*** y el ***sentido*** de una posible ***relación lineal entre dos variables***, a partir de una muestra representativa.

Sin embargo, muchas veces es posible encontrar un ***elevado coeficiente de correlación entre dos variables que no tienen relación alguna***, es decir, variables que no presentan relación justificada a través de alguna teoría específica (biología, astronomía, economía, entre otras). Cuando sucede esto, se dice que la correlación estadística existente entre estas variables es una ***correlación espúrea*** o ***sin sentido***. De esta forma, es posible hablar de correlación espúrea entre variables relacionadas a la economía, a la biología, a la astronomía, entre otras.

Formalmente, se dice que un alto coeficiente de correlación entre dos variables es ***espúreo*** si éste se explica por la ***presencia de un tercer factor*** y ***no debido a la existencia de una relación con sentido entre las variables analizadas***. En este caso, la correlación estadísticamente significativa entre las variables es una ***correlación espúrea*** o ***sin sentido***.

Karl Pearson (1897) fue el primero en utilizar el término ***correlación espúrea*** para ilustrar el origen de una correlación sin sentido entre ratios, a través del siguiente caso. Considérese un grupo de esqueletos que son desarmados en cada uno de sus huesos componentes y que se vuelven a armar unos nuevos utilizando huesos elegidos aleatoriamente de los distintos esqueletos originales. Si para verificar que los huesos de cada nuevo esqueleto corresponden al mismo individuo (lo cual no es cierto), se correlacionan la longitud de varios huesos de cada nuevo esqueleto divididos por la longitud del nuevo esqueleto al cual pertenecen, se obtiene un coeficiente de correlación muy alto y estadísticamente significativo. Si bien es cierto este resultado sugiere que los huesos de cada esqueleto analizado (los nuevos esqueletos) corresponden a los mismos individuos, esta no es una conclusión cierta. En este caso, se dice que existe una correlación espúrea pues la alta correlación se explica por la presencia de un tercer componente: la división de la longitud de los huesos que se correlacionaban por la longitud de cada nuevo esqueleto al cual pertenecen. Este caso será estudiado en detalle en la sección 4.1.

Durante el siglo *XX* se estudiaron muchos casos de ***correlaciones espúreas*** entre variables medidas a través de datos de ***corte transversal*** y ***series de tiempo***. El caso más anecdótico de una correlación espúrea en un contexto de corte transversal fue presentado por J. Neyman en 1952. Neyman analizó la relación entre la tasa de nacimientos y la población de cigüeñas en varias regiones, y encontró un alto coeficiente de correlación entre estas variables.

Entre los casos más conocidos de correlaciones espúreas en un contexto de series de tiempo se tienen los analizado por G. Udny Yule (1926) y Ploser y Schwer (1978). Por un lado, utilizando datos anuales para el período 1866-1911, G. Udny Yule encontró que el coeficiente de correlación entre la tasa de mortalidad en Inglaterra y Gales y el porcentaje de matrimonios en la iglesia de Inglaterra era de 0.95. Por otro lado, utilizando datos anuales para el período 1897-1958, Ploser y Schwert encontraron que el coeficiente de correlación

entre el logaritmo del ingreso nominal de Estados Unidos y logaritmo de la acumulación de manchas solares era de 0.91.

Estos casos sugieren que no siempre es posible asociar un coeficiente de correlación alto a la existencia de una relación lineal con significado (económico, biológico, o algún otro) entre dos variables. Entonces, lo ***único seguro*** es que el coeficiente de correlación permite determinar la ***fuerza*** y ***sentido*** de una ***relación lineal estadística*** entre dos variables, más ***no necesariamente de una relación lineal con sentido*** entre las mismas.

Dado este problema, es importante analizar las causas por las cuales pueden surgir correlaciones espúreas. Como se muestra en las siguientes secciones, las razones por las cuales surgen correlaciones espúreas en un contexto de corte transversal y en uno de series de tiempo pueden ser distintas. Sin embargo, antes de realizar este análisis, será importante conocer la estructura de una serie que puede representar a una variable económica.

3.2. Estructura de una Serie Económica

En general, las series económicas pueden presentan los siguientes componentes:

a. Un ***componente tendencial***, que puede ser determinístico (lineal o no lineal) o estocástico.
b. Un ***componente estacional***; es decir, patrones de comportamiento recurrentes para determinados períodos de tiempo.
c. Un ***componente irregular*** o ***modelable***[7].

Es importante señalar que no todas las series económicas presentan necesariamente los tres componentes. Por ejemplo, si los valores de las variables en cuestión están representados por datos de corte transversal, es usual que no presenten componentes tendenciales ni estacionales. Sin embargo, con datos de series de tiempo, es muy probable que las variables presenten los tres componentes.

[7] Existe un cuarto componente denominado cíclico que muchas veces –como en este caso– se asume como parte del componente irregular o modelable.

En general, el componente más importante de una serie económica es el ***componente irregular*** o ***modelable***, ya que contiene la mayor parte de la información económicamente relevante. Sin embargo, existen situaciones en las que los componentes tendenciales determinísticos y/o estocásticos poseen interpretación económica.

Un caso muy conocido en el que el componente tendencial determinístico (lineal o no lineal) tiene sentido económico lo constituye la tendencia determinística del PBI real[8]. Esta tendencia representa el PBI potencial, el cual crece a una tasa igual a la pendiente de la tendencia. Así, la diferencia entre la serie observada del PBI real y la tendencia (lineal o no lineal), permite obtener una aproximación de las fluctuaciones del PBI o ciclo económico. En la Figura 6 se muestra la serie mensual del PBI real de la economía peruana para el período enero 1933 - diciembre 2001, el PBI potencial (representado por una tendencia determinística no lineal) y el ciclo económico.

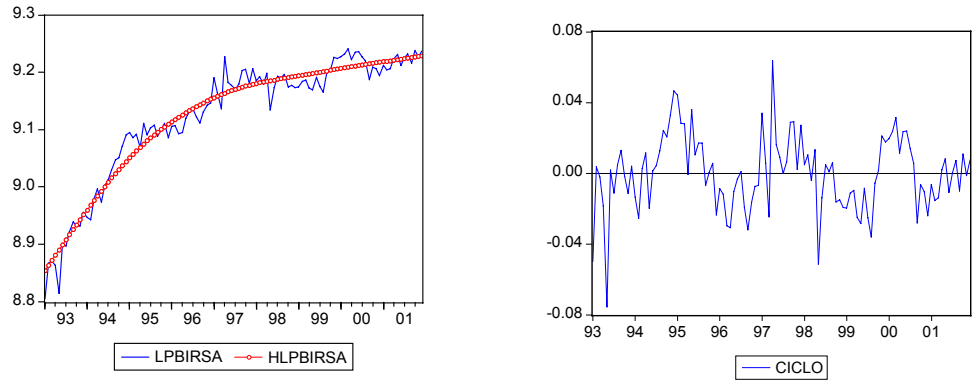


Figura 6: El PBI real, el PBI potencial y el ciclo económico

Asimismo, las tendencias estocásticas de dos o más series económicas podrían presentar una relación con sentido económico. Si dos series presentan ***tendencias estocásticas*** relacionadas entre sí (es decir, comparten una tendencia estocástica), se dice que las variables ***cointegran***. En términos económicos, si la relación tiene sustento teórico, se dice que estas variables presentan una ***relación de largo plazo económicamente significativa***. En este caso, la tendencia estocástica que comparten las variables se interpreta como un camino común del cual pueden desviarse temporalmente (corto plazo), pero no permanentemente (largo plazo).

[8] Aunque también tiene sentido la tenencia estocástica.

4. CORRELACIONES ESPÚREAS Y DATOS DE CORTE TRANSVERSAL

En esta sección se muestran tres posibles formas en las que pueden presentarse correlaciones espúreas en un contexto de corte transversal: el uso de ratios, observaciones atípicas o extraordinarias y grupos no relacionados.

4.1. Correlaciones Espúreas y el uso de Ratios

K. Pearson (1898) y R. Kronmal (1993) muestran que las correlaciones espúreas pueden surgir entre dos variables que se miden como cocientes o ratios respecto de una tercera variable.

Pearson, a través de su ejemplo de los esqueletos construidos aleatoriamente, muestra la existencia de un coeficiente de correlación significativo entre dos ratios cuyos componentes variables no presentan relación alguna. Para entender esto, considérese las variables *W, X*, *Y* y *Z*, tales que se cumplen las siguientes condiciones:

a. *Y* y *X* son independientes, por lo que no presentan relación significativa alguna.
b. *Z* es igual a la suma de *Y*, *X* y *W*

En el contexto del caso analizado por Pearson, *X e Y* representan las longitudes de diferentes huesos que fueron correlacionados. Estos huesos pertenecen al mismo esqueleto aleatoriamente construido, por lo cual no están relacionados. *Z* representa la longitud total del *esqueleto aleatorio* y *W* la longitud de los huesos de cada *esqueleto aleatorio* que no fueron utilizados en la correlación. Si se contara con 200 esqueletos construidos aleatoriamente a partir de 200 esqueletos originales, entonces se tienen los siguientes pares de observaciones:

$$\begin{matrix} (Y_1, X_1) \\ (Y_2, X_2) \\ \vdots \\ (Y_{200}, X_{200}) \end{matrix}$$

donde Y_1 representa la longitud de un primer grupo de huesos del esqueleto artificial 1, X_1 representa la longitud de un segundo grupo de huesos del esqueleto artificial 1; Y_2 representa la longitud de un primer grupo de huesos del esqueleto artificial 2 y X_2 representa la longitud de un segundo grupo de huesos del esqueleto artificial 2; y así sucesivamente. Si se dividen cada una de estas observaciones por la longitud total del esqueleto artificial al cual pertenecen, se obtienen las siguientes observaciones:

$$\begin{gathered}(Y_1/Z_1, X_1/Z_1) \\ Y_2/Z_2, X_2/Z_2 \\ \vdots \\ (Y_{200}/Z_{200}, X_{200}/Z_{200})\end{gathered}$$

donde Y_1/Z_1 representa la longitud de un primer grupo de huesos del esqueleto artificial 1 como porcentaje de la longitud total de éste esqueleto, X_1/Z_1 representa la longitud de un segundo grupo de huesos del esqueleto artificial 1 como porcentaje de la longitud total de este esqueleto; y así sucesivamente.

Para simular los resultados del caso presentado por Pearson, se construyeron artificialmente las variables W, X, Y y Z, con las características del problema que han sido mencionadas. Al analizar el coeficiente de correlación entre las observaciones de *Y* y *X*, el resultado es un coeficiente de correlación cercano a cero consistente con la realidad analizada:

Matriz de Correlaciones entre X e Y

	X	Y
X	1.000000	-0.006123
Y	-0.006123	1.000000

Esto puede observarse claramente en el gráfico de dispersión de las mismas (Figura 7):

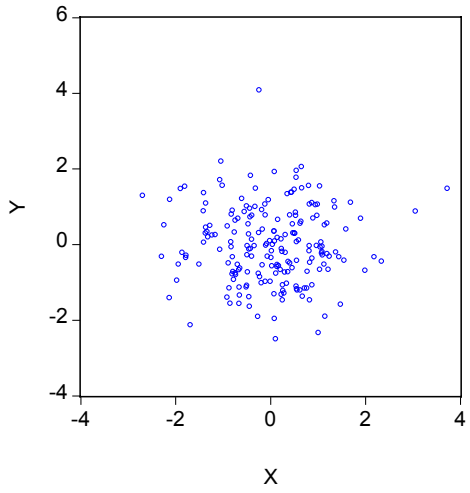


Figura 7: Relación entre X e Y

Sin embargo, al analizar la correlación entre *Y* y *X* utilizando los ratios *X/Z* e *Y/Z* (cuyos componentes variables son *X* e *Y*), se obtiene un alto coeficiente de correlación:

Matriz de Correlaciones entre X/Z e Y/Z

	X/Z	Y/Z
X/Z	1.000000	-0.965746
Y/Z	-0.965746	1.000000

lo cual también puede observarse claramente en el gráfico de dispersión de las mismas. Como se muestra en la Figura 8:

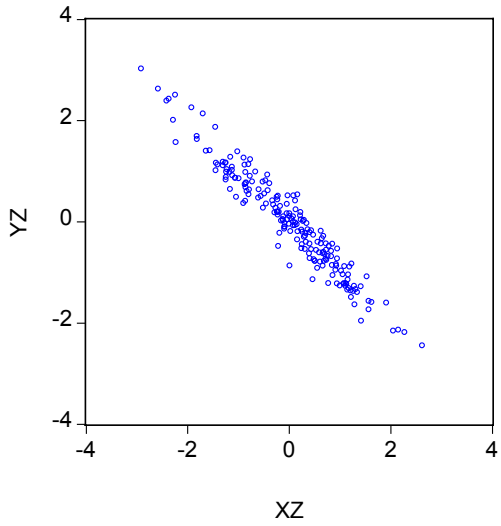


Figura 8: Relación entre X/Z e Y/Z

De esta manera, la alta correlación entre *X/Z* e *Y/Z* sería espúrea o sin sentido. La alta correlación está explicada por un tercer componente, *Z*, que esta relacionado a cada uno de los componentes variables, *X* e *Y*, que son independientes entre sí. En el contexto del caso analizado por Pearson, dado que se sabe que no existe relación entre las longitudes de los huesos de los esqueletos artificiales, un alto coeficiente de correlación entre las longitudes de los huesos como porcentaje de la longitud total del esqueleto artificial al cual pertenecen es espúreo o sin sentido. La explicación de este alto coeficiente de correlación está en la división de las longitudes de los huesos por la longitud total del esqueleto artificial al cual pertenecen.

Una forma simple de detectar la presencia de correlaciones espúreas cuando se utilizan ratios es analizar el gráfico de dispersión y el coeficiente de correlación entre los componentes variables de los mismos (cuando sea posible obtenerlos).

4.2. Presencia de Observaciones Atípicas (Out layers)

Un segundo caso en el cual puede surgir correlaciones espúreas se presenta cuando existen observaciones atípicas (*out layers)* tan importantes en magnitud que pueden generar un alto coeficiente de correlación entre dos variables que no tienen relación alguna.

Para mostrar este caso, se crearon dos series artificiales *S1* y *S2* independientes entre sí, como se muestra en la matriz de correlaciones:

Matriz de Correlaciones entre S1 y S2

	S1	S2
S1	1.000000	0.024656
S2	0.024656	1.000000

y en el gráfico de dispersión correspondiente:

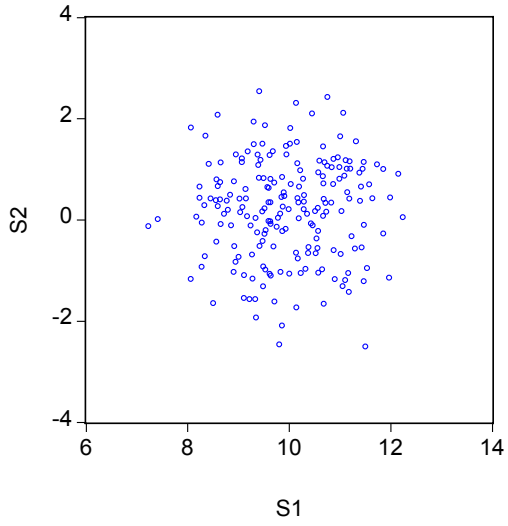


Figura 9: Relación entre S1 y S2

Si una de las observaciones de esta muestra ésta compuesta por valores muy diferentes a los usuales (relativamente muy grandes o muy pequeños), denominados atípicos u *out layers*, se obtendría la siguiente la matriz de correlaciones:

Matriz de Correlaciones entre S1 y S2 con un "out layer"

	S1	S2
S1	1.000000	0.906860
S2	0.906860	1.000000

y el siguiente gráfico de dispersión:

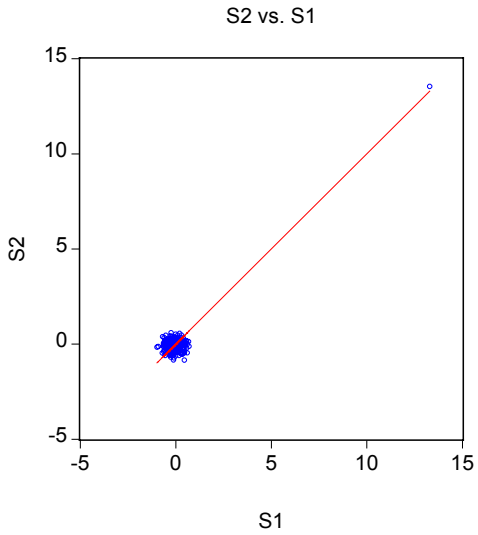


Figura 10: Relación entre S1 y S2

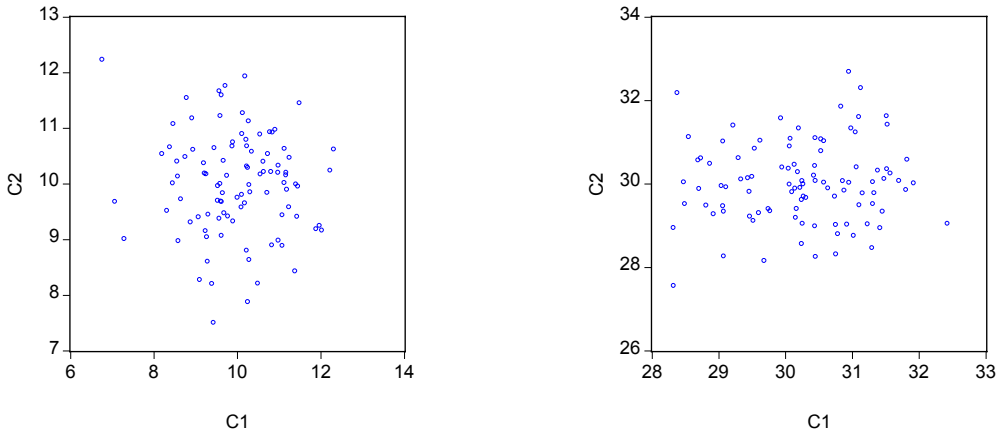
Claramente, el alto coeficiente de correlación (0.91) entre las variables *S1* y *S2* se debe a la presencia de un *out layer*, el cual fuerza la existencia de una relación lineal entre las mismas (como se muestra en la Figura 10), a través de una línea que representa los datos. De manera natural, este resultado se puede generalizar a más de un *out layer*.

La identificación de la existencia de correlaciones espúreas entre dos variables debida a la presencia de uno o más *out layers*, puede ser identificada fácilmente analizando el gráfico de dispersión de las mismas.

4.3. Grupos No Relacionados

Un tercer caso en el cual pueden surgir correlaciones espúreas se presenta cuando existen dos o más grupos de observaciones que relacionan dos variables, tales que el coeficiente de correlación es bajo en cada grupo, pero alto cuando se analizan todos los grupos simultáneamente. A estos grupos de observaciones se les denomina grupos no relacionados.

Para mostrar este caso, se crearon dos grupos de observaciones para dos variables ficticias *C1* y *C2* tales que estas no tienen relación alguna, como se puede apreciar en sus respectivos gráficos de dispersión:



Grupo 1 ***Grupo 2***

Figura 11: Grupos No Relacionados

y en sus matrices de correlaciones:

Matriz de Correlaciones entre C1 y C2 en el Grupo 1

	C1	C2
C1	1.000000	-0.066483
C2	-0.066483	1.000000

Matriz de Correlaciones entre S1 y S2 en el Grupo 2

	C1	C2
C1	1.000000	0.062029
C2	0.062029	1.000000

Sin embargo, al analizar de manera conjunta ambos grupos de observaciones, se obtiene un gráfico de dispersión donde se muestra que las observaciones pueden ser representadas por una línea recta:

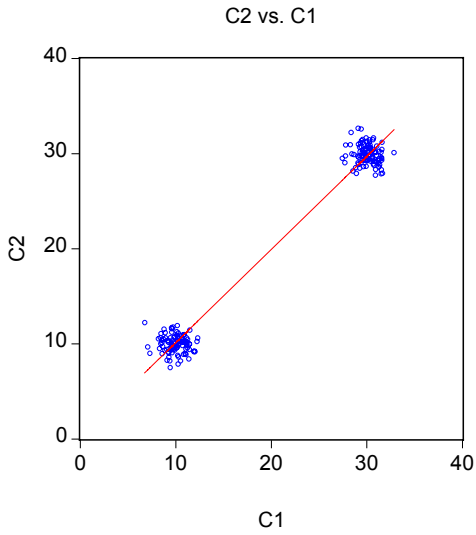


Figura 12: Correlación entre Conglomerados y Correlación Espúrea

La matriz de correlaciones confirma la existencia de una relación lineal estadísticamente significativa:

Matriz de Correlaciones entre C1 y C2

	C1	C2
C1	1.000000	0.988784
C2	0.988784	1.000000

De esta manera, es posible encontrar un coeficiente de correlación alto entre dos variables no relacionadas entre sí, o correlación espúrea, si se analizan de manera conjunta dos o más grupos no relacionados de observaciones.

Este caso de correlación espúrea entre grupos no relacionados puede ser considerado como una generalización del caso de correlaciones espúreas cuando existen *out-layers*: un grupo estaría constituido por el *out-layer* y el segundo por las observaciones restantes.

El caso de correlación espúrea presentado por Neyman, a través de su ejemplo de las cigüeñas, puede ser considerado como un caso de grupos no relacionados. Los datos utilizados por Neyman que se presentan en la cuadro 2 corresponden al número de mujeres (por cada 10 mil), cigüeñas y nacimientos, para una muestra de 54 localidades. El número de mujeres está aproximado a números enteros; es decir, si en una localidad existen 14 mil mujeres, entonces el número de mujeres por cada 10 mil es 1.

Cuadro 2: Información sobre Mujeres, Nacimientos y Cigüeñas

Localidad	Mujeres	Cigüeñas	Nacimientos
1	1	2	10
2	1	2	15
3	1	2	20
4	1	3	10
5	1	3	15
6	1	3	20
7	1	4	10
8	1	4	15
9	1	4	20
10	2	4	15
11	2	4	20
12	2	4	25
13	2	5	15
14	2	5	20
15	2	5	25
16	2	6	15
17	2	6	20
18	2	6	25
19	3	5	20
20	3	5	25
21	3	5	30
22	3	6	20
23	3	6	25
24	3	6	30
25	3	7	20
26	3	7	25
27	3	7	30
28	4	6	25
29	4	6	30
30	4	6	35
31	4	7	25
32	4	7	30
33	4	7	35
34	4	8	25
35	4	8	30
36	4	8	35
37	5	7	30
38	5	7	35
39	5	7	40
40	5	8	30
41	5	8	35
42	5	8	40
43	5	9	30
44	5	9	35
45	5	9	40
46	6	8	35
47	6	8	40
48	6	8	45
49	6	9	35
50	6	9	40
51	6	9	45
52	6	10	35
53	6	10	40
54	6	10	45
		Corr (C,N) =	**0,83**

Si se analiza toda la muestra, se observa que el coeficiente de correlación entre el número de cigüeñas y el número de nacimientos es 0.83, lo cual indicaría la existencia de una relación lineal estadísticamente significativa entre estas variables. Sin embargo, no existe una relación con algún sentido diferente del estadístico, pues no es posible afirmar a partir de este

resultado que las cigüeñas traen a los bebés (a mayor número de cigüeñas, mayor número de nacimientos).

Cuadro 3: Información por Grupos

	Localidad	Mujeres	Cigüeñas	Nacimientos
GRUPO 1	1	1	2	10
	2	1	2	15
	3	1	2	20
	4	1	3	10
	5	1	3	15
	6	1	3	20
	7	1	4	10
	8	1	4	15
	9	1	4	20
			Corr (C,N) =	**0.00**
GRUPO 2	10	2	4	15
	11	2	4	20
	12	2	4	25
	13	2	5	15
	14	2	5	20
	15	2	5	25
	16	2	6	15
	17	2	6	20
	18	2	6	25
			Corr (C,N) =	**0.00**
GRUPO 3	19	3	5	20
	20	3	5	25
	21	3	5	30
	22	3	6	20
	23	3	6	25
	24	3	6	30
	25	3	7	20
	26	3	7	25
	27	3	7	30
			Corr (C,N) =	**0.00**
GRUPO 4	28	4	6	25
	29	4	6	30
	30	4	6	35
	31	4	7	25
	32	4	7	30
	33	4	7	35
	34	4	8	25
	35	4	8	30
	36	4	8	35
			Corr (C,N) =	**0.00**
GRUPO 5	37	5	7	30
	38	5	7	35
	39	5	7	40
	40	5	8	30
	41	5	8	35
	42	5	8	40
	43	5	9	30
	44	5	9	35
	45	5	9	40
			Corr (C,N) =	**0.00**
GRUPO 6	46	6	8	35
	47	6	8	40
	48	6	8	45
	49	6	9	35
	50	6	9	40
	51	6	9	45
	52	6	10	35
	53	6	10	40
	54	6	10	45
			Corr (C,N) =	**0.00**

La existencia de un coeficiente de correlación alto se explica por la presencia de grupos de datos no relacionados en la muestra analizada. El cuadro 3 muestra la información utilizada por Neyman en seis grupos de localidades de acuerdo al número de mujeres de las mismas.

Al analizar el coeficiente de correlación y el gráfico de dispersión entre el número de cigüeñas y de nacimientos para cada grupo de localidades (Figura 13), se observa que en cada grupo de localidades el número de cigüeñas y de nacimientos no están correlacionados, como era de esperarse.

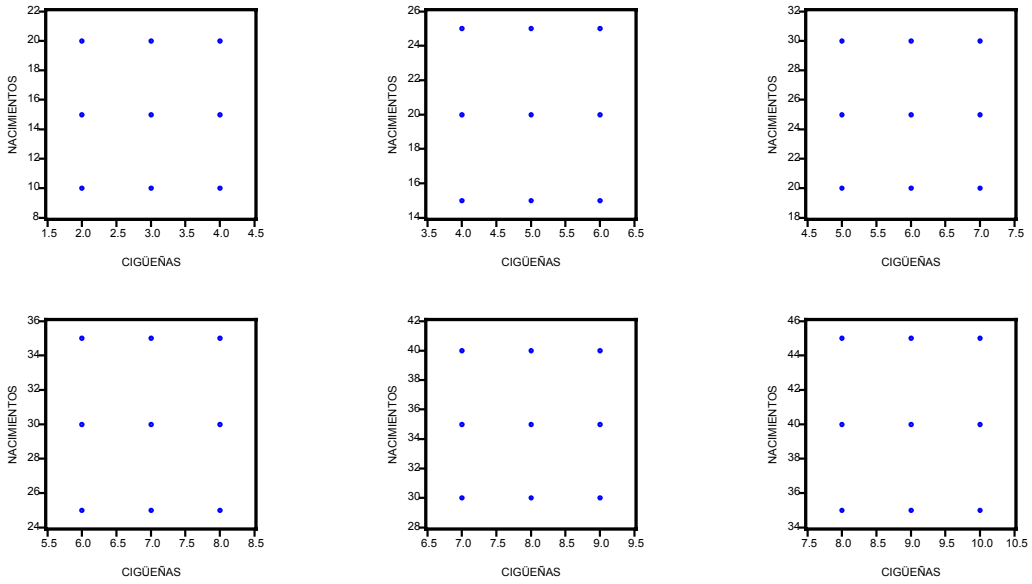


Figura 13: Relación entre nacimientos y cigüeñas para cada grupo de localidades.

Sin embargo, al considerar (por ejemplo) los grupos 1 y 6 en una sola muestra, se obtiene un coeficiente de correlación alto (0.92), al igual que cuando se consideran todos los grupos simultáneamente (0.82), lo cual se muestra a través de los gráficos de dispersión correspondientes de la Figura 14. De esta forma, se concluye que el alto coeficiente de correlación entre las cigüeñas y los nacimientos es una correlación espúrea, explicada por la presencia de conglomerados.

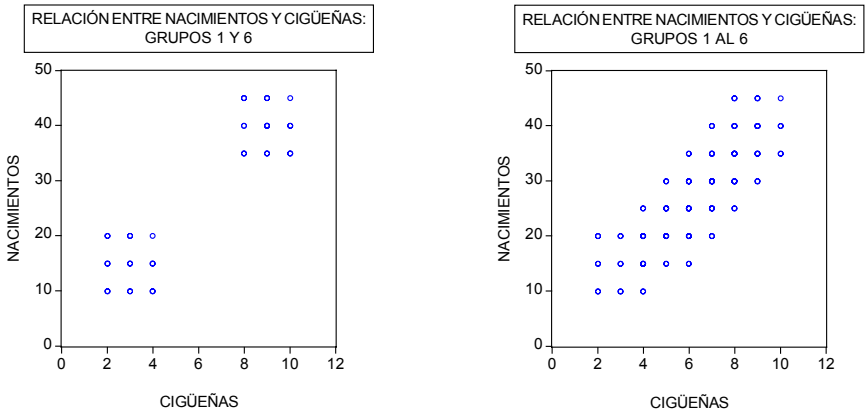


Figura 14:Relación lineal entre grupos no relacionados

5. CORRELACIONES ESPÚREAS Y SERIES DE TIEMPO

En el contexto de series de tiempo, las correlaciones espúreas pueden surgir por las mismas razones consideradas para el contexto de corte transversal: el uso de ratios, la presencia de *out layers* y grupos no relacionados. Sin embargo, las tendencias determinísticas o estocásticas - propias de la mayoría de series de tiempo – también pueden generar correlaciones espúreas entre variables que no tienen sentido alguno.

5.1. Simulación de Series de Tiempo

Para demostrar que la presencia de componentes tendenciales pueden generan coeficientes de correlación significativamente altos, se utilizan dos series de tiempo de 100 observaciones cada una que representan a dos variables X e Y. Estas series se construyen artificialmente de tal forma que no presentan relación lineal alguna (considerando alguna teoría específica), como lo muestra la matriz de correlaciones:

Matriz de Correlaciones para X e Y

	X	Y
X	1.000000	0.032141
Y	0.032141	1.000000

A partir de ésta matriz se observa que la correlación entre *X* e *Y* es prácticamente cero (0.03), lo cual es consistente con la construcción de las dos series. La Figura 15 muestra el comportamiento de estas variables a lo largo de las 100 observaciones.

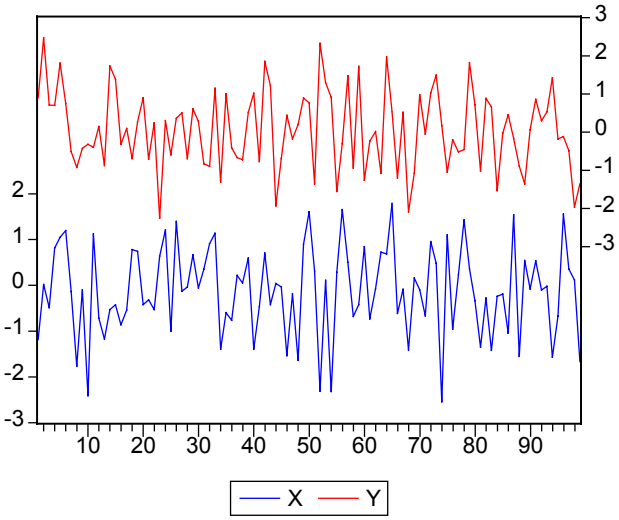


Figura 15: Series Artificiales X e Y

A partir de esta figura, no es posible afirmar que si una de las variables aumenta, la otra también aumenta, disminuye o no se mueve. Es decir, no es posible inferir la existencia de una relación positiva, negativa o simplemente que no exista relación. Sin embargo, el gráfico de dispersión para *X* e *Y* - en el cual se han normalizado las unidades de medida de las series[9] - sí permite afirmar que ***no existe relación lineal*** entre estas variables, como se muestra en la Figura 16. En este ejemplo, el coeficiente de correlación funciona bien, pues es posible afirmar que no existe relación lineal estadística ni significativa.

[9] En adelante, los gráficos de dispersión que se presenten utilizarán datos normalizados.

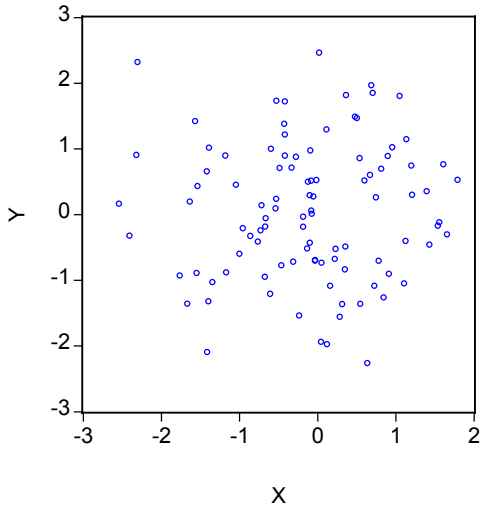


Figura 16: Gráfico de dispersión entre X e Y .
No existe relación lineal significativa.

Si a cada una de estas dos series construidas artificialmente se les añade una ***tendencia determinística*** (lineal) ***que crece con el tiempo***[10], pero sin significado alguno:

$$XT_t = X_t + t^1$$

$$YT_t = Y_t + t^2$$

se obtiene una matriz de correlaciones como la siguiente:

Matriz de Correlaciones para XT e YT

	XT	YT
XT	1.000000	0.997099
YT	0.997099	1.000000

Como se puede observar a partir de la matriz de correlaciones, el simple hecho de añadir una tendencia lineal creciente a cada una de las variables originales, genera una fuerte relación lineal positiva entre las nuevas variables *XT* e *YT* (*r = 0,99*).

[10] También podría ser una tendencia lineal decreciente, los resultados son similares en términos de la existencia de una relación lineal significativa estadísticamente, pero negativa.

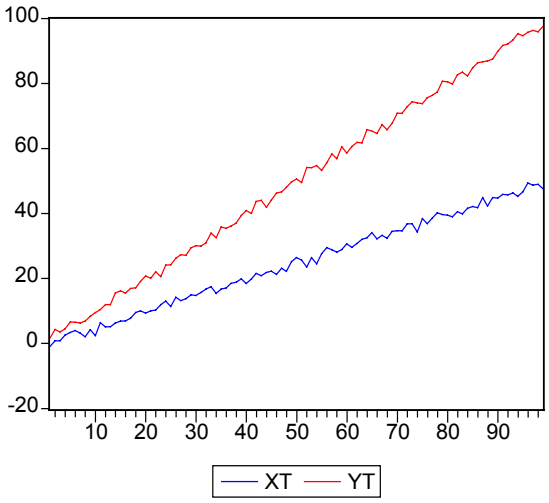


Figura 17: Series Artificiales XT e YT

A partir de la Figura 17, se puede observar que ambas series crecen a lo largo del tiempo, lo cual muestra la posible existencia de una ***relación positiva entre las variables XT e YT***, cuya fuerza y linealidad pueden inferirse a partir del gráfico de dispersión (Figura 18). Además, el alto grado de correlación se refleja por el hecho de que los puntos que representan las observaciones están muy cercanos entre sí y alrededor de una recta.

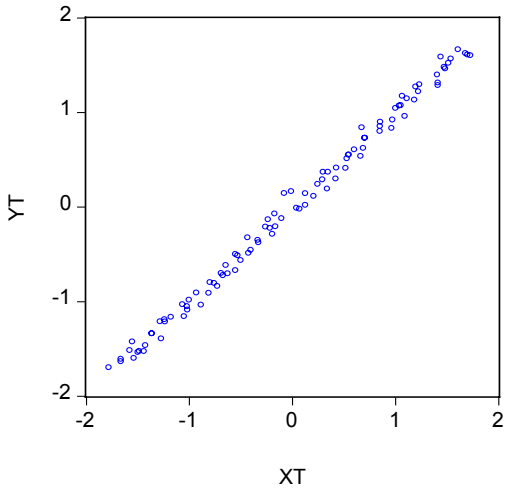


Figura 18: Gráfico de dispersión entre XT e YT.

Estos mismos resultados se obtienen si se añade una ***tendencia estocástica***[11] **(**sin ningún significado especial) a cada una de las series originalmente creadas, *X* e *Y*, creándose de esta forma dos nuevas variables, *XRW* e *YRW*:

$$XRW_t = X_t + Z_t^1$$

$$YRW_t = Y_t + Z_t^2$$

Las nuevas variables, como en el caso anterior, muestran un coeficiente de correlación muy cercano a uno:

Matriz de Correlaciones para XRW e YRW

	XRW	YRW
XRW	1.000000	0.989578
YRW	0.989578	1.000000

Al observar el gráfico de las series (Figura 19), se puede apreciar que ambas crecen, lo cual induciría a pensar (al menos estadísticamente) que la correlación entre las variables es positiva y aparentemente fuerte.

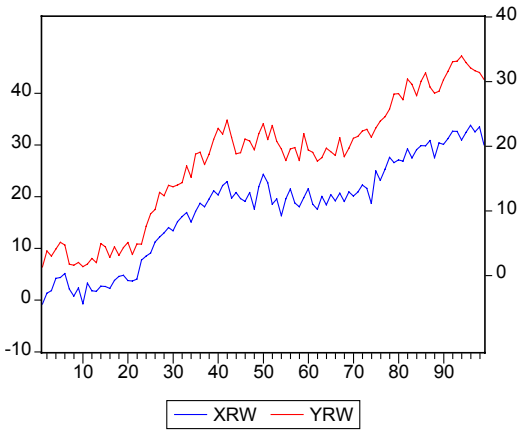


Figura 19: Series Artificiales XW e YW

[11] Una tendencia estocástica es de la forma: $Z_t = Z_{t-1} + \varepsilon_t$, donde $\varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$. También se le conoce como paseo aleatorio o random walk.

En efecto, al observar el gráfico de dispersión de las variables *XW* e *YW* (Figura 20), se puede apreciar que la correlación es significativamente alta (fuerte) y positiva.

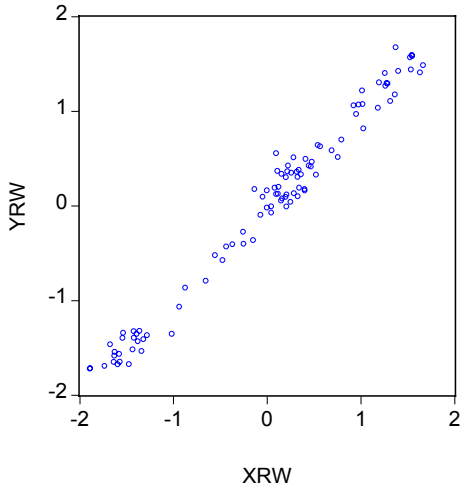


Figura 20: Relación lineal significativa entre XW e YW.

Estos ejemplos, que utilizan series artificiales, muestran que un coeficiente de correlación alto entre dos variables que no tienen relación alguna puede ser producto de la existencia de una ***tendencia lineal*** (***determinística***) o una ***tendencia estocástica***.

5.2. Tendencias y Correlaciones Espúreas

Considérense las variables *XT* e *YT* tales que cada una de ellas contiene un componente modelable (*X* e *Y*) y una tendencia lineal determinística (t^1 y t^2), como las simuladas anteriormente:

$$XT_t = X_t + t^1$$

$$YT_t = Y_t + t^2$$

Dada la estructura de las series, podrían presentarse cuatro situaciones diferentes:

a. Relación con sentido entre los componentes irregulares o modelables (X e Y) y entre los componentes tendenciales (t^1 y t^2).
b. Relación con sentido entre los componentes irregulares o modelables (X e Y), pero no entre los componentes tendenciales (t^1 y t^2).
c. No existe relación con sentido entre los componentes irregulares o modelables (X e Y), pero sí entre los componentes tendenciales (t^1 y t^2).
d. No existe relación con sentido entre los componentes irregulares o modelables (X e Y) ni entre los componentes tendenciales (t^1 y t^2).

En todos los casos, a excepción de la situación *d*, se afirma que existe una relación con sentido entre *XT* e *YT*, pues existe relación con sentido entre al menos uno de sus componentes. Sin embargo, ***en todos los casos existe una relación estadísticamente significativa*** entre *XT* e *YT* medida a través del ***coeficiente de correlación*** entre las variables, lo cual se explica por la presencia de las tendencias lineales (pesar de que no estén relacionadas significativamente). Este resultado sería el mismo si el componente tendencial de las variables fuera estocástico.

Así, cuando se analizan dos series temporales que presentan tendencias que no tienen significado alguno, el coeficiente de correlación no siempre esta asociado a una relación con sentido. Cuando esto sucede, como en el caso *d*, se dice que la ***correlación es espúrea o sin sentido***: el factor que genera la correlación espúrea es el componente tendencial. En otro caso, el coeficiente de correlación si estaría asociado a una relación con sentido entre las variables.

5.3. Identificación de Correlaciones Espúreas

Dado que es posible la existencia de correlaciones espúreas en series de tiempo que presentan tendencias (determinísticas o estocásticas), que no tienen algún significado especial, es importante tratar de establecer alguna metodología para identificar este problema.

Para este propósito, es necesario entender el concepto de ***variables en niveles*** y en ***primeras diferencias***, pues la metodología de identificación consistirá en el análisis de los coeficientes de correlación de las mismas.

5.3.1. Variables en niveles y primeras diferencias

Una ***variable en niveles*** es cualquier variable que se toma como punto de partida para cualquier transformación posterior. Por ejemplo, el nivel del PBI nominal en 1999 es el valor del PBI nominal en 1999, el nivel del gasto público en 1999 es el saldo del gasto público en 1999, el nivel de la tasa de crecimiento anual del PBI en 1999 es la tasa de crecimiento registrada entre 1999 y 1998, entre otros ejemplos. Una variable está expresada en ***primeras diferencias*** en el período *t* si se obtiene como la diferencia entre su valor actual y su valor pasado:

$$\Delta X_t = X_t - X_{t-1} \tag{4.1}$$

Así, por ejemplo, la primera diferencia del PBI nominal en 1999 es la ***diferencia*** entre el valor del PBI nominal en 1999 y el valor del PBI nominal en 1998; la primera diferencia del gasto público en 1999 es la ***diferencia*** entre el saldo del gasto público en 1999 y el saldo del gasto público en 1998; la primera diferencia de la tasa de crecimiento anual del PBI en 1999 es la ***diferencia*** entre la tasa de crecimiento entre 1999 y 1998 y la tasa de crecimiento entre 1998 y 1997.

Los ejemplos planteados muestran que la definición de variable en niveles y en primeras diferencias es relativa. Así, la primera diferencia de X_t podría considerarse como una ***variable en niveles***, mientras que la segunda diferencia de X_t, definida como:

$$\Delta^2 X_t = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \tag{4.2}$$

podría considerarse como la ***primera diferencia*** de "la primera diferencia de X_t".

5.3.2. Primeras Diferencias y Eliminación de Tendencias

Dadas las definiciones de variables en niveles y primeras diferencias, es fácil mostrar que ***si una variable en niveles presenta una tendencia lineal, su primera diferencia ya no presenta ese componente tendencial***. Para verificar esto, considérense las variables simuladas *XT* e *YT*:

$$XT_t = X_t + t^1$$

$$YT_t = Y_t + t^2$$

donde las tendencias no tienen significado alguno. La primera diferencia de la serie *X* está dada por:

$$\Delta XT_t \equiv XT_t - XT_{t-1} = X_t + t^1 \text{ - } [X_{t-1} + (t^1 - 1)]$$

$$\Delta XT_t = X_t + t^1 \text{ - } X_{t-1} \text{ - } t^1 + 1$$

$$\Delta XT_t = 1 + (X_t \text{ - } X_{t-1})$$

$$\Delta XT_t \equiv 1 + \Delta X_t \qquad (3.3)$$

De manera análoga, la primera diferencia de *Y* está dada por:

$$\Delta YT_t \equiv YT_t - YT_{t-1} = Y_t + t^1 \text{ - } [Y_{t-1} + (t^1 - 1)]$$

$$\Delta YT_t = 1 + \Delta Y_t \qquad (3.5)$$

Así, ***la primera diferencia de una variable elimina todo componente tendencial determinístico*** existente en su expresión en niveles. Sin embargo, las primeras diferencias de los componentes irregulares o modelables permanecen. Si las variables en niveles presentaran un componente tendencial estocástico, la primera diferencia también lo eliminaría[12].

Dado el supuesto que X_t y Y_t no están relacionados, si se calcula el coeficiente de correlación entre las primeras diferencias de las variables, éste sería bajo a pesar de que el coeficiente en niveles es alto, cómo se mostró en la sección 5.1.

[12] Este es un ejercicio que el lector debería resolver.

Matriz de Correlaciones para las primeras diferencias de XT e YT

	XT	YT
XT	1.000000	0.002450
YT	0.002450	1.000000

De esta manera, la ***relación espúrea entre X e Y*** presenta un ***coeficiente de correlación alto para los niveles de las variables*** y uno ***bajo para las primeras diferencias de las variables***. Sin embargo, si existiera una relación con sentido entre X_t e Y_t, entonces el ***coeficiente de correlación entre los niveles de las variables sería alto, al igual que el coeficiente de correlación para sus primeras diferencias***.

El análisis precedente permite establecer la siguiente metodología de identificación de correlaciones espúreas en un contexto de series de tiempo, bajo el supuesto de que los componentes tendenciales (determinísticos o estocásticos) no tienen sentido alguno:

a. *Si el **coeficiente de correlación de los niveles de las variables es significativamente alto**, pero el **coeficiente de correlación de las primeras diferencias de las mismas es bajo**, entonces la correlación entre los niveles de las variables es una **correlación espúrea o sin sentido alguno**.*
b. *Si el **coeficiente de correlación de los niveles y de las primeras diferencias de las variables es significativamente alto**, entonces la correlación entre los niveles de las variables es una **correlación con sentido**.*

5.3. Correlaciones Espúreas y Correlaciones con Sentido Económico

En muchas ocasiones, la teoría económica establece que es posible encontrar relaciones con sentido entre pares de variables bajo ciertas circunstancias. Por ejemplo, no siempre existe una relación con sentido económico entre los precios de todos los bienes de la economía, pero sí entre los precios de los bienes producidos por dos empresas que compiten en el mismo mercado. De hecho, si los precios de dos bienes cualesquiera presentan un componente tendencial sin sentido económico, es posible encontrar un coeficiente de correlación alto entre ellos sin que eso implique que guarden alguna relación con sentido.

En casos como estos, donde es posible que exista relación con sentido económico y se intenta encontrar evidencia sobre ello, es posible aplicar la metodología presentada en la sección anterior. Así, bajo el supuesto de que los componentes tendenciales de cada una de las variables económicas relacionadas no tiene sentido:

a. Si el ***coeficiente de correlación de los niveles de las variables es significativamente alto***, pero el ***coeficiente de correlación de las primeras diferencias de las mismas es bajo***, entonces la correlación entre los niveles de las variables es una ***correlación espúrea o sin sentido económico***.

b. Si el ***coeficiente de correlación de los niveles y de las primeras diferencias de las variables es significativamente alto***, entonces la correlación entre los niveles de las variables es una ***correlación con sentido económico***.

Esta metodología para identificar correlaciones espúreas a partir de los coeficientes de correlación, fue utilizada por Stigler y Sherwin (1985). La hipótesis principal del artículo en el cual ponen en práctica la metodología es la siguiente:

> " ***Si dos bienes o servicios compiten en el mismo mercado, sus precios deberían estar muy correlacionados***."

Sin embargo, como la mayoría de precios muestran movimientos tendenciales en el tiempo (que se asumen que no tienen interpretación económica en muchos casos), los autores analizaron la correlación entre los niveles de los precios y sus primeras diferencias (cambio en lo precios), para poder identificar la existencia de una correlación con sentido económico. Los resultados del análisis mostraron que el valor del coeficiente de correlación no disminuye drásticamente al utilizar las variables en niveles y en primeras diferencias. De esta manera, los autores concluyen que los bienes estudiados compiten en el mismo mercado.

6. APLICACIONES: IDENTIFICACIÓN DE RELACIONES ESPÚREAS ENTRE VARIABLES ECONÓMICAS

En esta sección se presentan aplicaciones del coeficiente de correlación para el análisis de relaciones bivariadas y la identificación de posibles correlaciones espúreas. Para este fin, se analizan diferentes relaciones bivariadas entre variables económicas empleando datos mensuales para el Perú (series de tiempo) que abarcan el período enero 1993 - diciembre 2001.

6.1. Descripción de los datos utilizados

Los datos empleados para realizar las aplicaciones han sido obtenidos del Banco Central de Reserva del Perú (Boletines Semanales, Memorias Anuales y página web[13]), y corresponden a ocho variables importantes para la economía peruana.

Las series escogidas como variables *proxy* (aproximadas) para cada una de las variables económicas utilizadas fueron: la emisión primaria (*EMISION*), el índice de precios al consumidor con base 1994 (*IPC94*), el circulante nominal (CIR), el producto bruto interno real (*PBIR*), el tipo de cambio informal promedio (*TCIP*), el crédito de las empresas bancarias al sector privado en moneda extranjera (*CREME*), la tasa de interés pasiva en moneda nacional (*TIPMN*) y la tasa de interés activa en moneda extranjera (*TAMEX*).

Para obtener series que permitan representar relaciones bivariadas con sentido económico, se realizaron algunas transformaciones. En primer lugar, se eliminaron los componentes cíclicos o estacionales de las series *EMISIÓN*, *CIR* y *PBIR*, a través del procedimiento de desestacionalización *ratio moving average-multiplicative* del programa Eviews. Las series desestacionalizadas se denominaron *EMISIONSA*, *CIRSA* y *PBIRSA*.

Los componentes estacionales dependen de la economía analizada. Por ejemplo, la emisión monetaria en el caso peruano presenta un componente estacional representado por *picos* en los meses de julio y diciembre, que se observan en la figura 21.

[13] http://www.bcrp.gob.pe

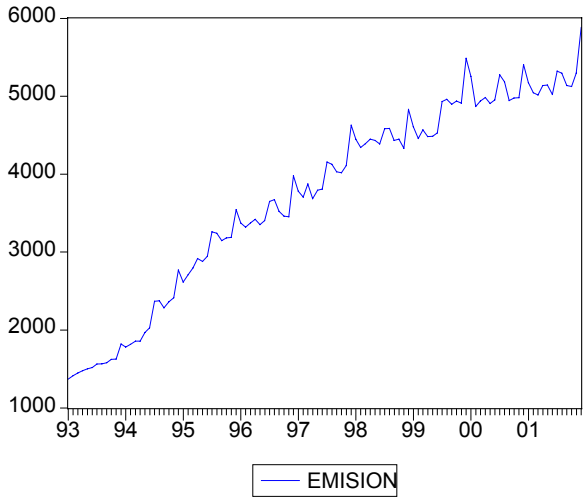


Figura 21: La Emisión en el Perú y su estacionalidad

Luego de aplicar a la serie *EMISIÓN* el proceso *ratio moving average-multiplicative*, la serie desestacionalizada (*EMISIONSA*) ya no presenta el componente cíclico, como se muestra en la Figura 22:

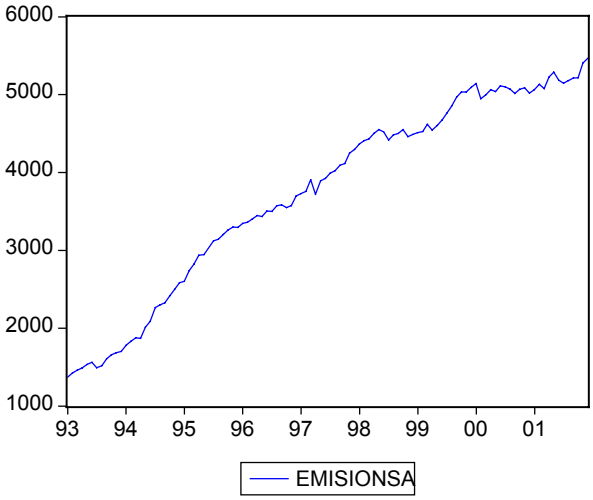


Figura 22: Emisión desestacionalizada

Como segundo las variables nominales *CIR*, *CREME*, *TAMEX* y *TCIP* fueron transformadas a reales (*CR*, *CRERME*, *TARMEX* y *TCIPR*), utilizando el *IPC94* y el *TCIP* para el caso de las variables denominadas en moneda extranjera. Finalmente, para

homogenizar la escala de las variables se aplicaron logaritmos a todas las series con excepción de las tasas de interés, generando así las series finales *LEMISIONSA*, *LCR*, *LPBIRSA*, *LIPC94*, *LCRERME* y *LTCIPR*.

6.2. Relaciones Económicas y Correlaciones Espúreas

Para aplicar los conceptos de correlación y correlaciones espúreas, se analizarán seis relaciones bivariadas importantes propuestas por la teoría macroeconómica convencional: dinero y precios, demanda por dinero real e ingreso, demanda por dinero real y tasa de interés nominal en moneda nacional, producción y tasa de interés real en moneda extranjera, producción y crédito real en moneda extranjera, producción y tipo de cambio real.

6.2.1. Dinero y Precios

Para analizar la relación entre dinero y precios se utilizaron como variables *proxy* *LEMISIONSA* y *LIPC94*. La teoría macroeconómica establece la existencia de una relación positiva entre estas variables. Por un lado, la ecuación cuantitativa del dinero, definida como:

$$MV = PQ$$

establece que en el largo plazo – bajo el supuesto que la velocidad de circulación del dinero (V) es estable y que el producto está en su nivel potencial o de pleno empleo - un aumento de la cantidad de dinero tendrá efectos únicamente sobre el nivel de precios; así, un aumento de 10 por ciento en la cantidad de dinero generará en el largo plazo un incremento de 10 por ciento en el nivel de precios. Este resultado es uno de los supuestos que subyacen a la nueva macroeconomía clásica.

Por otro lado, considerando el modelo de la síntesis neoclásica (que asume una oferta agregada con pendiente positiva), un aumento de la cantidad de dinero genera un aumento en la demanda agregada y por lo tanto un aumento en el producto (fenómeno denominado *demanda efectiva*), el cual va acompañado de un aumento del nivel de precios, como se muestra en la Figura 23:

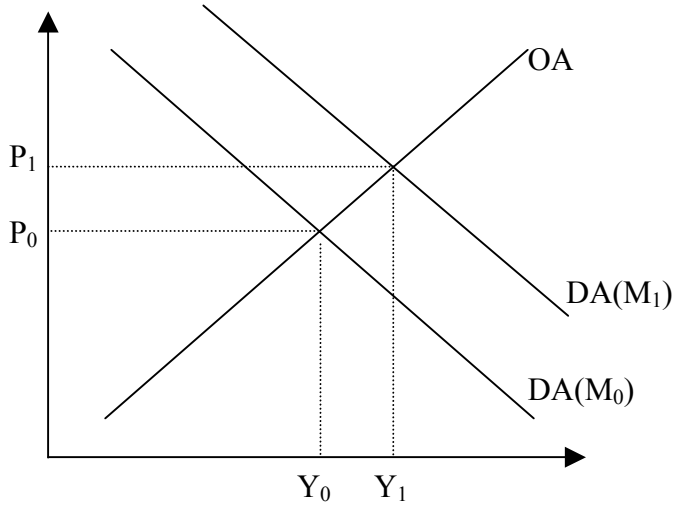


Figura 23: Demanda, Oferta, Producción y Precios

Para verificar la existencia de un sustento empírico de esta relación teórica entre dinero y precios, se grafican (Figura 24) las series *LEMISIONSA* e *IPC94* para el periodo enero 1993 – diciembre 2001:

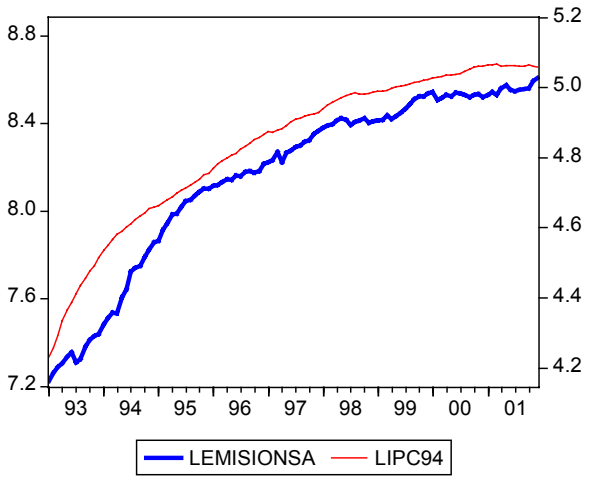


Figura 24: Dinero y Precios

El gráfico muestra que ambas series crecen a lo largo del tiempo y de forma parecida. Para verificar la existencia de una relación lineal se utiliza el gráfico de dispersión de ambas series:

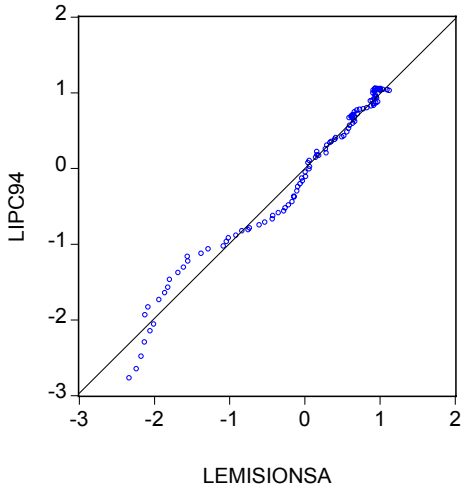


Figura 25: Gráfico de dispersión entre los niveles de Emisión y Precios

Como se puede apreciar en la Figura 25, la relación entre *LEMISIONSA* y *LIPC94* es aproximadamente lineal (los puntos pueden ser representados por una línea recta) y positiva. Para determinar la fuerza de la relación entre estas variables se utiliza la matriz de correlaciones:

	LEMISIONSA	LIPC94
LEMISIONSA	1.000000	0.988957
LIPC94	0.988957	1.000000

Como era de esperar por lo establecido por la teoría económica, la correlación entre los niveles de las variables es positiva y fuerte (0.99). Para comprobar si la relación lineal entre las variables analizadas es o no ***espúrea***, es necesario analizar la correlación entre las primeras diferencias de las variables.

	DLEMISIONSA	DLIPC94
DLEMISIONSA	1.000000	0.224576
DLIPC94	0.224576	1.000000

La matriz de correlaciones entre las primeras diferencias de las variables muestra una relación lineal positiva y débil entre *DLEMISIONSA* y *DLIPC94* (0.22), lo cual también puede observarse en el gráfico de dispersión:

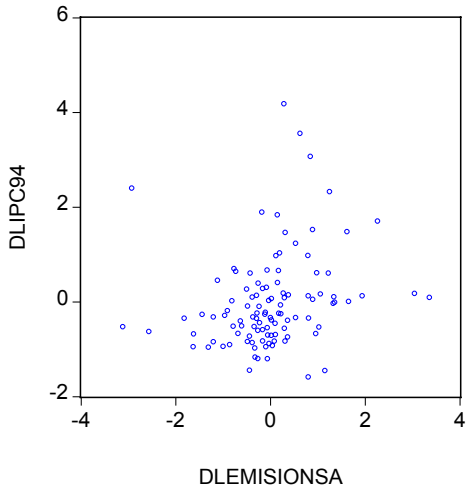


Figura 26: Gráfico de dispersión entre las primeras diferencias de la Emisión y los Precios

Bajo el supuesto de que las tendencias de las series no tienen sentido alguno, este resultado permitiría concluir que la relación entre el dinero y el nivel de precios de la economía es ***espúrea***. Sin embargo, si se considera la teoría monetarista de la determinación de los precios, las tendencias que presentan cada una de las series se interpretan como la senda de largo plazo que siguen; así, en este caso la correlación sería una correlación con sentido económico: el dinero determina los precios en el largo plazo.

6.2.2. Demanda por Dinero Real e Ingreso Real.

Como segundo ejemplo, se analiza la relación entre la demanda por dinero real y el ingreso real, relacionados teóricamente a través del modelo clásico de demanda por dinero real. En este modelo, el dinero se utiliza solamente para realizar transacciones.

Para verificar empíricamente la existencia de ésta relación teórica se utilizaron como variables *proxy LCR* (para la demanda por dinero real) y *LPBIRSA* (para el ingreso real).

El gráfico de la demanda por dinero real y el ingreso real (figura 27), permite apreciar un comportamiento creciente en ambas series, para el periodo enero 1993 – diciembre 2001:

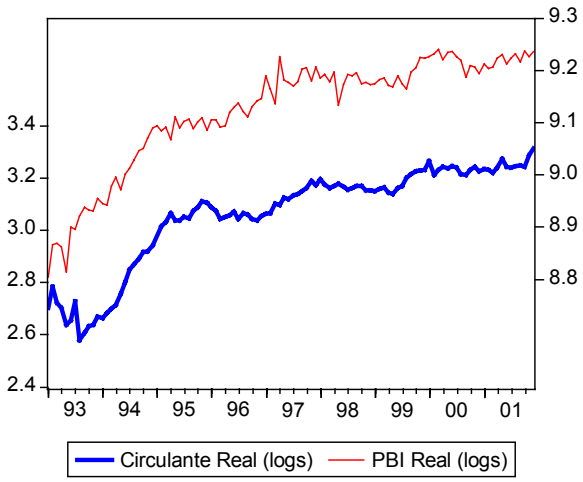


Figura 27: Dinero e Ingreso reales

El gráfico de dispersión (Figura 28), sugiere la existencia de una relación aproximadamente lineal positiva entre las variables *LCR* y *LPBIRSA*.

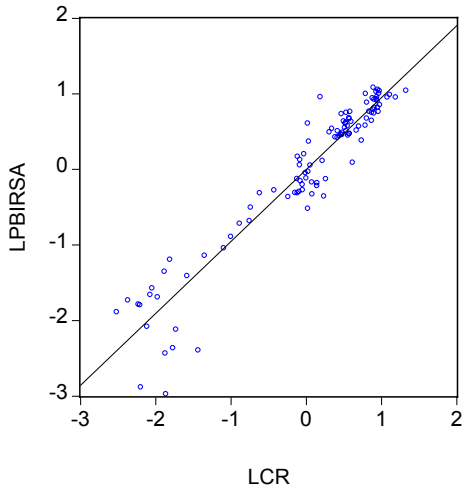


Figura 28: Gráfico de dispersión entre los niveles de Dinero e Ingreso reales

La matriz de correlaciones indica que la existencia de un grado de asociación lineal entre las variables positivo y bastante alto (0.95), corroborándose de esta forma lo establecido por la teoría económica y lo observado preliminarmente en el gráfico de dispersión:

Matriz de Correlaciones en Niveles: Dinero e Ingreso Reales

	LCR	LPBIRSA
LCR	1.000000	0.953381
LPBIRSA	0.953381	1.000000

Sin embargo, para determinar si la correlación entre estas variables es una correlación espúrea, es necesario analizar la matriz de correlaciones entre las primeras diferencias de las mismas:

Matriz de Correlaciones en Primeras Diferencias: Dinero e Ingreso Reales

	DLCR	DLPBIRSA
DLCR	1.000000	0.036931
DLPBIRSA	0.036931	1.000000

La matriz muestra una relación lineal débil entre las primeras diferencias de las variables. Por lo tanto, se concluye que la correlación entre la demanda por dinero real y el ingreso es ***espúrea*** o sin sentido, lo cual implica que la correlación entre las variables está explicada por la presencia de componentes tendenciales.

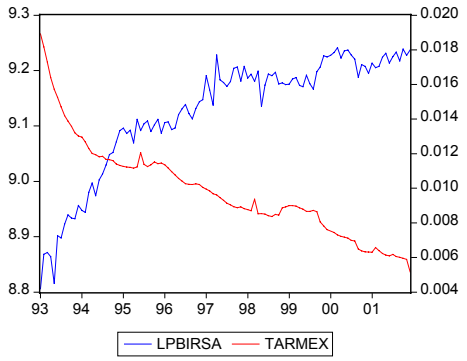
Nuevamente, bajo el supuesto de que las tendencias de las series no tienen sentido alguno, este resultado permitiría concluir que la relación entre el dinero real y el ingreso real es ***espúrea***. Sin embargo, si se considera como relevante la relación de largo plazo entre el dinero real y el ingreso real, entonces las tendencias que presentan cada una de las series se interpretan como la senda de largo plazo que siguen cada una de ellas; en este caso, la correlación sería una ***correlación con sentido económico***: en el largo plazo la demanda real por dinero esta íntimamente relacionada con el ingreso real.

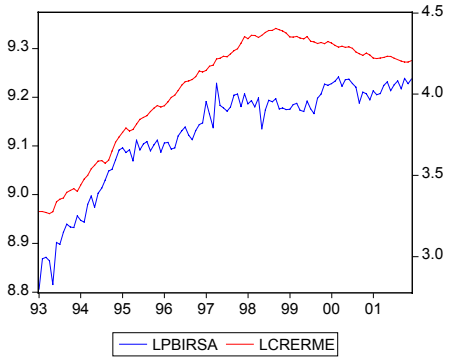
6.2.3 Producción, Tasa de Interés, Crédito y Tipo de Cambio reales

La teoría macroeconómica tradicional establece relaciones económicamente significativas entre la producción, la tasa de interés, el crédito y tipo de cambio - todas expresadas en términos reales.

Por un lado, un aumento de la tasa de interés real encarece el crédito y por lo tanto desincentiva la inversión en términos reales, generándose finalmente un efecto adverso sobre la producción. Por ello, existe una relación directa o positiva entre el producto real y el crédito real, y una relación inversa o negativa entre el producto real y la tasa de interés real.

Por otro lado, la relación entre el producto real y el tipo de cambio real no es única y depende de las características de la economía. Por ejemplo, en una economía como la peruana donde aproximadamente el 80 por ciento de los activos monetarios están denominados en dólares y la carga de la deuda externa –también denominada en dólares– constituye más del 20 por ciento del producto, un aumento del tipo de cambio tendría efectos negativos sobre el producto. Sin embargo, en otras economías, la relación entre el producto y el tipo de cambio reales es positiva: un incremento del tipo de cambio real incentiva las exportaciones netas, generándose un aumento de la demanda agregada y en consecuencia –por demanda efectiva– un aumento del producto.





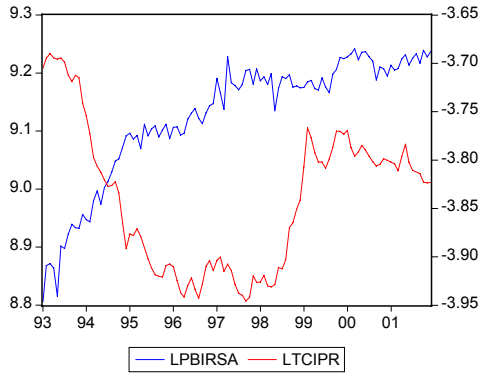
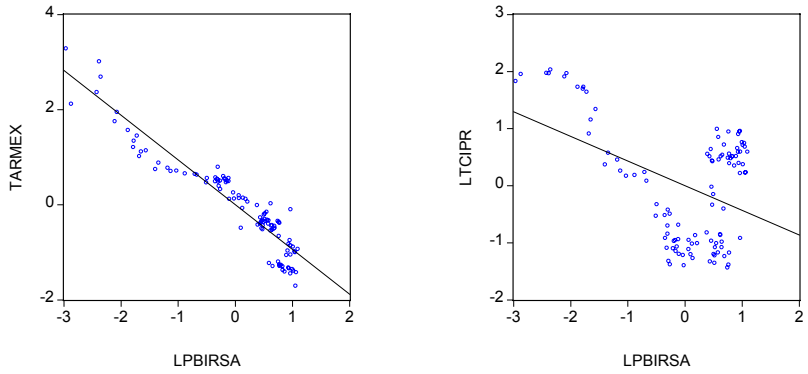


Figura 29: Producto, Tasa de Interés, Crédito y Tipo de Cambio reales

Las variables *proxy* utilizadas para medir la producción, la tasa de interés, el crédito y el tipo de cambio reales fueron *LPBIRSA*, *TARMEX*, *CRERME* y *LTCIPR*, respectivamente. Para el caso de la tasa de interés y el crédito, se utilizaron las series correspondientes a moneda extranjera, dada la importancia de la dolarización de la economía peruana para el periodo estudiado.

La contrapartida empírica para las relaciones bivariadas establecidas por la teoría económica se aprecia en los gráficos de las series de la Figura 29, verificándose el sentido de las relaciones establecidas por la teoría.



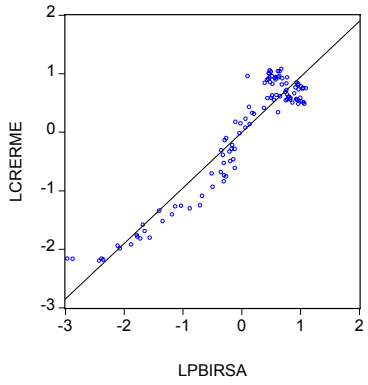


Figura 30: Relaciones Bivariadas entre Producto, Tasa de Interés, Crédito y Tipo de Cambio Reales.

Los gráficos de dispersión de la Figura 30, confirman la existencia de una relación lineal inversa, tanto para el caso del producto real vs. la tasa de interés real como para el del producto real vs. el tipo de cambio real, y una relación directa entre el producto y el crédito reales.

La matriz de correlaciones de las variables permite determinar la fuerza de estas relaciones bivariadas lineales:

	LPBIRSA	TARMEX	LCRERME	LTCIPR
LPBIRSA	1.000000	-0.940297	0.947747	-0.431767
TARMEX	-0.940297	1.000000	-0.871744	0.256799
LCRERME	0.947747	-0.871744	1.000000	-0.440914
LTCIPR	-0.431767	0.256799	-0.440914	1.000000

Bajo el supuesto de que los componentes tendenciales que presentan cada una de las series no tienen sentido alguno, la matriz de correlaciones en primeras diferencias de las mismas permite concluir que las relaciones bivariadas entre el producto y la tasa de interés, el crédito y el tipo de cambio reales son ***espúreas***:

	DLPBIRSA	DTARMEX	DLCRERME	DLTCIPR
DLPBIRSA	1.000000	-0.178665	0.248468	0.077523
DTARMEX	-0.178665	1.000000	-0.016953	0.090657
DLCRERME	0.248468	-0.016953	1.000000	-0.131397
DLTCIPR	0.077523	0.090657	-0.131397	1.000000

El gráfico de dispersión de las variables en primeras diferencias (Figura 31), refuerza la conclusión, pues muestra la ausencia de relación alguna entre las variables:

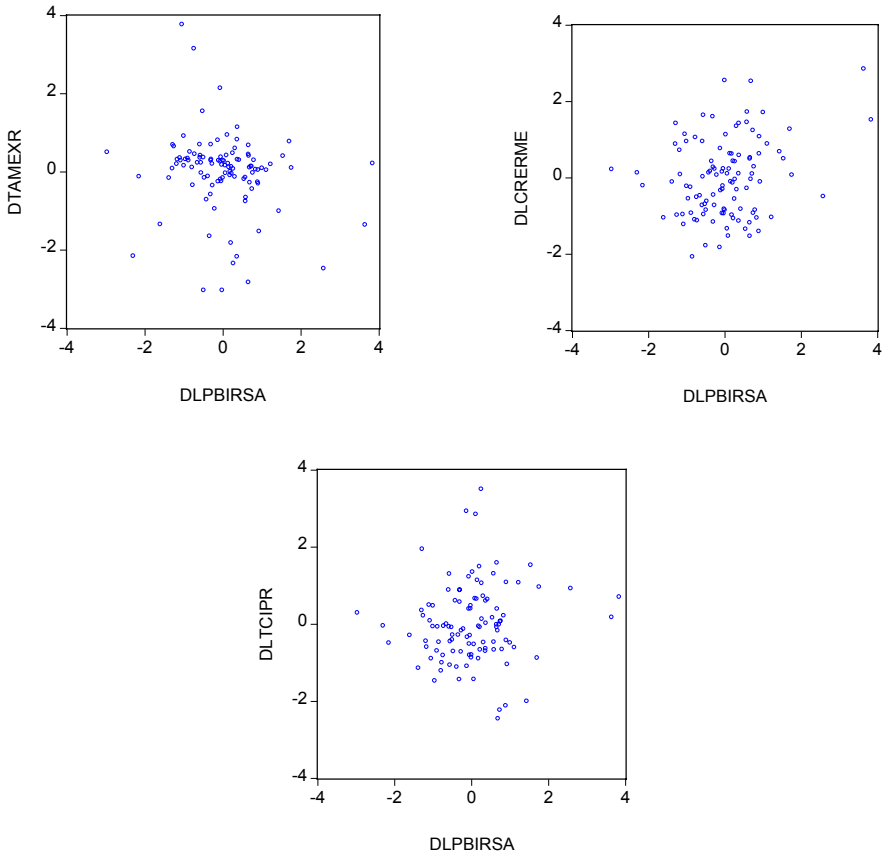


Figura 31: Gráficos de dispersión de las primeras diferencias de las series.

Sin embargo, una vez más, si la teoría económica establece la posibilidad de una relación de largo plazo entre cualquiera de los pares de variables analizados, entonces los componentes tendenciales tendrían sentido (en este caso económico) y por lo tanto las correlaciones no serían espúreas.

7. CONCLUSIONES

(1) El coeficiente de correlación es un instrumento estadístico que permite establecer la ***fuerza*** y ***dirección*** de una relación lineal estadística entre dos variables a partir de una muestra determinada, bajo el supuesto de que ésta es representativa.

(2) Existen casos en los que un coeficiente de correlación significativo entre dos variables es consecuencia de un ***tercer factor*** diferente de ellas y no de la existencia de una relación lineal con algún sentido entre las mismas bajo la consideración de alguna ***teoría conocida*** (por ejemplo, biológica, física, económica, entre otras). Cuando esto sucede, se dice que la ***correlación*** es ***espúrea***. La correlaciones espúreas pueden presentarse cuando las variables analizadas son medidas a través de datos de corte transversal o series de tiempo.

(3) Entre las principales causas de las correlaciones espúreas en un contexto de corte transversal figuran el uso de ratios, la presencia de datos atípicos (*out layers*) y de grupos no relacionados. Para el caso de correlaciones espúreas causadas por el uso de ratios, la detección implica el análisis de la correlación y sentido de los componentes variables (numeradores) de los mismos. Para el caso de datos atípicos el gráfico de dispersión es un instrumento muy importante para identificarlos. Para el caso de grupos no relacionados, es posible detectar la presencia de correlaciones espúreas analizando el gráfico de dispersión y la estructura de la muestra.

(4) Para el caso de series temporales, además de las causas mencionadas para el contexto de corte transversal, la presencia de tendencias (determinísticas o estocásticas) que carecen de sentido alguno también pueden generar correlaciones espúreas. Bajo estas circunstancias, es posible detectar la presencia de correlaciones espúreas analizando el coeficiente de correlación de los niveles y las primeras diferencias de las series. Si el coeficiente de correlación es considerado significativamente alto en niveles y en primeras diferencias, entonces la correlación no es espúrea. Si el coeficiente de correlación es considerado significativamente alto en niveles mas no así en primeras diferencias, entonces la correlación es espúrea.

Referencias Bibliográficas

Banco Central de Reserva del Perú
Memoria Anual. Varios números.

Banco Central de Reserva del Perú
Boletín Semanal. Varios números.

Granger, C.W.
1969 Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. Econométrica, Vol. 37.

Kronmal, Richard
1993 Spurious Correlation and the Fallacy of the Standard Ratio Revisited. Journal of the Royal Statistical Society. Vol. 156, parte 3, p. 379-392.

Liebetrau, Albert M.
1983 *Mesures of Association.* Newbury Park: Sage.

Lima, Elon Lages
1997 Análisis Real. Lima: Perú Offset, Vol. 1.

Mendenhall, William; Scheaffer, Richard L. y Dennis D. Wackerly
1994 Estadística Matemática con Aplicaciones. México: Grupo Editorial Iberoamerica. 2da. Edición.

Neyman, Jerzy
1952 Lectures and Conferences on Mathematical Statistics and Probability. Washington DC: US Department of Agriculture. 2da, ed., p. 143-154. 1952.

Pearson, Karl
1897 Mathematical contributions to the theory of evolution – on the form of spurious correlation which may arise when indices are used in the measurements of organs. Proceedings of the Royal Society of London. Vol. 60, p. 268-286. 1897.

Stigler, George J. y Robert A. Sherwin
1985 Extent of the Market. Journal of Political Economics, Vol. 28, 1985.

APÉNDICE

Demostración 1:

La covarianza es sensible a las unidades de medida de las variables

Sean dos variables *X* e *Y* para las cuales se cuenta con *n* valores observados. Sus valores muestrales promedio están dados por:

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

$$\overline{Y} = \frac{Y_1 + \ldots + Y_n}{n}$$

y su covarianza por:

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

Si todos los valores de *X* e *Y* fueran multiplicados por α y β, tales que $\alpha > 1$ y $\beta > 1$, las unidades de medida de ambas variables se incrementarían porcentualmente en dichos factores, al igual que sus respectivas medias muestrales:

$$\frac{\alpha X_1 + \ldots + \alpha X_n}{n} = \frac{\alpha(X_1 + \ldots + X_n)}{n} = \alpha\overline{X}$$

$$\frac{\beta Y_1 + \ldots + \beta Y_n}{n} = \frac{\beta(Y_1 + \ldots + Y_n)}{n} = \beta\overline{Y}$$

Dado esto, tenemos que:

$$Cov(\alpha X, \beta Y) = \frac{1}{n-1}\sum_{i=1}^{n}(\alpha X_i - \alpha\overline{X})(\beta Y_i - \beta\overline{Y}) = \frac{1}{n-1}\sum_{i=1}^{n}\alpha\beta(X_i - \overline{X})(Y_i - \overline{Y})$$

$$Cov(\alpha X, \beta Y) = \frac{\alpha\beta}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \alpha\beta\, Cov(X,Y)$$

$$Cov(\alpha X, \beta Y) = \alpha\beta Cov(X, Y)$$

Es decir, el valor de la covarianza entre las variables se altera en una proporción igual a $\alpha\beta$. Sin embargo, el signo no se altera.

<u>Demostración 2</u>: -1< r < 1

Considérese la siguiente tautología:

$$\sum_{i=1}^{n}(y - cx)^2 \geq 0$$

donde *y* y *x* son variables que son afectadas por la suma, y *c* es una constante. Dado esto, es posible definir arbitrariamente la constante "*c*" como:

$$c = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2}$$

con lo cual:

$$\sum_{i=1}^{n}\left(y - \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} x\right)^2 \geq 0$$

Manipulando esta expresión tenemos:

$$\sum_{i=1}^{n}\left(y^2 + x^2 \frac{(\sum_{i=1}^{n} xy)^2}{(\sum_{i=1}^{n} x^2)^2} - 2yx \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2}\right) \geq 0$$

$$\sum_{i=1}^{n} y^2 + \sum_{i=1}^{n} x^2 \frac{(\sum_{i=1}^{n} xy)^2}{(\sum_{i=1}^{n} x^2)^2} - 2\sum_{i=1}^{n} xy \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} \geq 0$$

$$\sum_{i=1}^{n} y^2 + \frac{(\sum_{i=1}^{n} xy)^2}{(\sum_{i=1}^{n} x^2)^2} \sum_{i=1}^{n} x^2 - 2\sum_{i=1}^{n} xy \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} \geq 0$$

$$\sum_{i=1}^{n} y^2 + \frac{(\sum_{i=1}^{n} xy)^2}{\sum_{i=1}^{n} x^2} - 2\frac{(\sum_{i=1}^{n} xy)^2}{\sum_{i=1}^{n} x^2} \geq 0$$

$$\sum_{i=1}^{n} y^2 - \frac{(\sum_{i=1}^{n} xy)^2}{\sum_{i=1}^{n} x^2} \geq 0$$

$$\sum_{i=1}^{n} y^2 \geq \frac{(\sum_{i=1}^{n} xy)^2}{\sum_{i=1}^{n} x^2}$$

$$(\sum_{i=1}^{n} x^2)(\sum_{i=1}^{n} y^2) \geq (\sum_{i=1}^{n} xy)^2$$

Esta última expresión es una forma de escribir la desigualdad de ***Cauchy-Schwartz***. A partir de esta expresión tenemos[14]:

$$1 \geq \left(\frac{(\sum_{i=1}^{n} xy)^2}{(\sum_{i=1}^{n} x^2)(\sum_{i=1}^{n} y^2)} \right) = r^2$$

[14] En realidad, la demostración puede empezar a partir de esta forma de expresar la desigualdad Cauchy-Schawrz.

$$1 \geq \left| \frac{\sum_{i=1}^{n} xy}{\sqrt{\sum_{i=1}^{n} x^2} \sqrt{\sum_{i=1}^{n} y^2}} \right| = | r |$$

Entonces:

$$1 \geq |r|$$

$$-1 \leq r \leq 1$$

Con lo cual queda demostrada la afirmación.