Complementos de Econometría[1]

1. Introducción

En este capítulo vamos a ilustrar mediante ejemplos los aspectos básicos del análisis de regresión múltiple en cuestiones legales. Desplegar visualmente los datos ayuda a menudo a describir las variables utilizadas en semejante análisis. La Figura 1 es un diagrama de dispersión que vincula mediciones de un test de aptitud en el trabajo (*Job Aptitude Test Score*) en el eje de las *x*, con una evaluación del rendimiento en el trabajo (*Job Performance Rating*) en el eje de las *y*. Cada punto indica dónde está situado un individuo medido con el test de aptitud y cómo ha sido su rendimiento laboral. Por ejemplo, el individuo representado en el punto A de la figura registró 49 en el test de aptitud y su rendimiento laboral fue evaluado en 62.

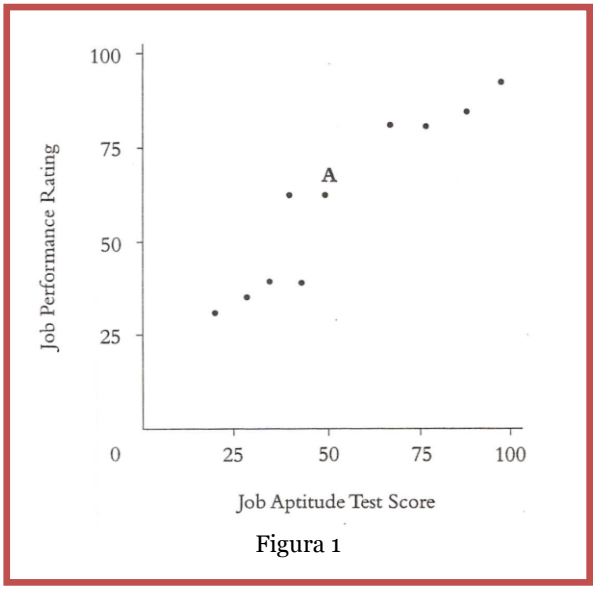


Figura 1

La relación entre dos variables puede resumirse mediante el coeficiente de correlación, que oscila entre -1 (una relación lineal perfecta negativa) y +1 (una relación lineal perfecta positiva). La Figura 2 indica tres relaciones posibles entre la variable de aptitud y la variable de rendimiento laboral. En 3(a) existe correlación positiva: en general, mejores evaluaciones de aptitud laboral van acompañadas por mediciones más altas del rendimiento laboral, y peores evaluaciones de aptitud van acompañadas por mediciones más bajas del rendimiento laboral. En 3(b) se presenta una correlación negativa, ya que mejores evaluaciones de aptitud laboral están asociadas con peores mediciones del rendimiento, y viceversa. Si la relación es suficientemente débil, no existe correlación alguna, como lo ilustra la Figura 3 3(c).

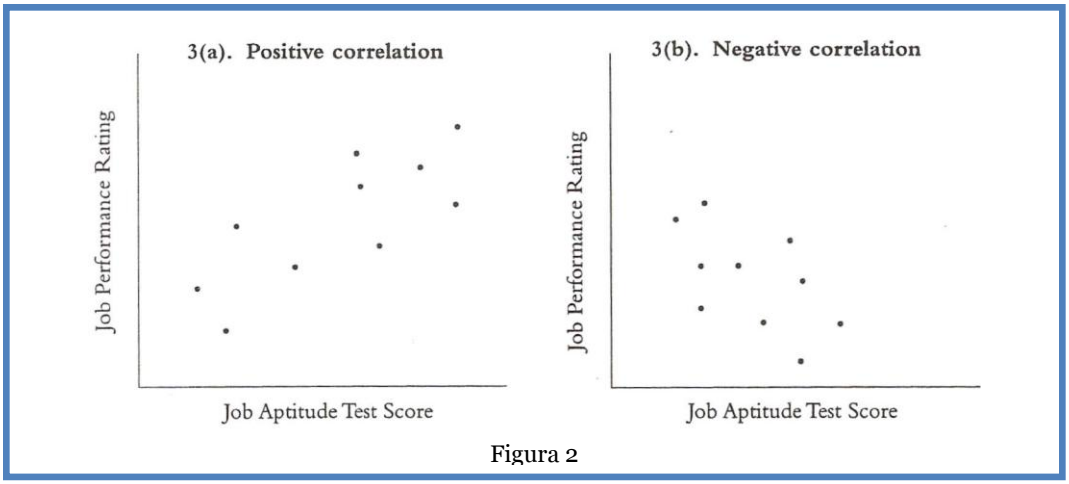
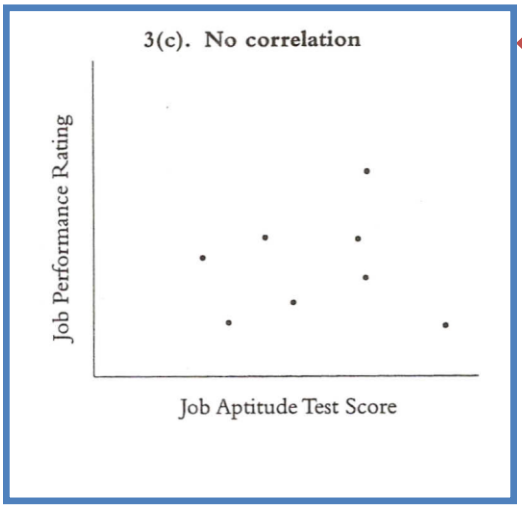


Figura 2

[1] Ver Daniel L. Rubinfeld, Reference Guide on Multiple Regression, *Reference Manual on Scientific Evidence*, 3rd ed., Federal Judicial Center (2011). Appendix: The Basics of Multiple Regression.

El análisis de regresión múltiple va más allá de calcular correlaciones; es un método con el cual se usa una línea de regresión a fin de vincular la media de una variable – la variable dependiente – con los valores de otras variables explicativas. De ello resulta que el análisis de regresión puede ser usado para predecir los valores de una variable usando valores de las otras. Por ejemplo, si la evaluación del rendimiento laboral medio depende del puntaje de las pruebas de aptitud, éstas pueden ser usadas para predecir el rendimiento.



Una línea de regresión es la línea de *mejor ajuste* a un conjunto de puntos de un diagrama de dispersión. Si sólo hay una variable explicativa, la ecuación de la línea recta viene definida por:

[1] $Y = a + b X$

En esta ecuación, "a" es la *ordenada al origen* de la recta (o intersección con el eje de las "y" cuando X es igual a 0), y "b" es la *pendiente* – el cambio de la variable dependiente asociado con el cambio de 1 unidad de la variable explicativa. En la Figura 3, p. ej., cuando la prueba de aptitud es 0, la ordenada al origen (predicha) es 18,4. Asimismo, por cada punto adicional en que se incrementa la prueba de aptitud, el rendimiento en el trabajo crece en 0,73 unidades, que viene dado por la pendiente 0,73. Por consiguiente, la línea de regresión estimada es:

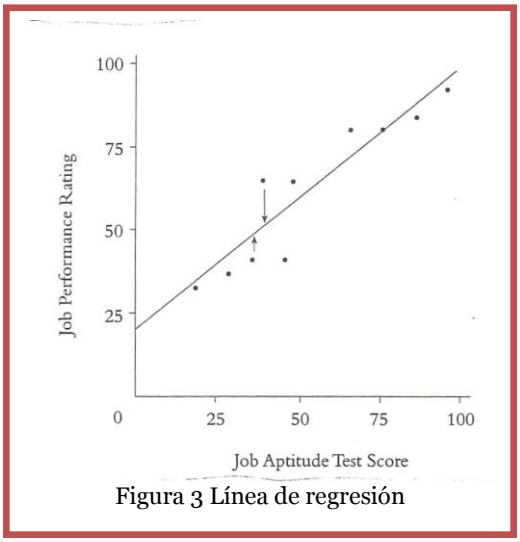


Figura 3 Línea de regresión

[2] $Y = 18,4 + 0,73 X.$

Es típico que la línea de regresión sea estimada usando el método estándar de mínimos cuadrados ordinarios (MCO), donde los valores de "a" y "b" se calculan minimizando la suma de los desvíos al cuadrado de los puntos respecto a la línea de regresión. Así, los desvíos positivos y negativos de igual tamaño son computados de manera similar, mientras que los desvíos amplios cuentan más que los pequeños. En la Figura 3 las líneas de desvío son *verticales porque la ecuación predice la evaluación del rendimiento laboral a partir de los puntajes de las pruebas de aptitud, no los puntajes de las pruebas de aptitud a partir de las evaluaciones del rendimiento laboral.*

Variables importantes que podrían influir sobre la variable dependiente en forma sistemática, cuyos datos puedan ser obtenidos, deberían ser incluidas explícitamente en un modelo estadístico. Las influencias restantes, que pueden ser pequeñas tomadas individualmente, pero que podrían

resultar sustanciales en el agregado, se incluyen dentro de un término adicional de error.[2] La regresión múltiple es un procedimiento que permite separar los efectos sistemáticos (asociados con las variables explicativas) de los efectos aleatorios (asociados con el término de error) y también ofrece un método para evaluar el éxito del proceso llevado a cabo.

2. El Modelo de Regresión Lineal

Con un número arbitrario de variables explicativas, el modelo de regresión lineal adopta la forma:

[3] $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$

En esta forma, Y representa a la variable dependiente (p.ej. el salario de un empleado), y $X_1 \ldots X_k$ a las variables explicativas (p. ej., la experiencia de cada empleado y su sexo, codificado como 1 o 0, respectivamente). El término de error, ε, representa la influencia colectiva *no observable* de todas las variables omitidas. En una representación lineal, cada uno de los términos adicionados implica parámetros desconocidos, $\beta_0, \beta_1, \ldots, \beta_k$, que son estimados "ajustando" la ecuación a los datos usando mínimos cuadrados.

Las mismas variables pueden aparecer en múltiples formas. Por ejemplo, Y podría representar el logaritmo del salario de un empleado, y X_1 representar el número de años de experiencia del empleado. La representación logarítmica es apropiada cuando Y crece en forma exponencial con los incrementos de X – para cada unidad de X, el aumento de Y se va haciendo cada vez más grande. Por ejemplo, si el experto quisiera graficar el crecimiento de la población mundial (Y) a lo largo del tiempo (t), una ecuación con la forma siguiente podría resultar apropiada: $\log(Y) = \beta_0 + \beta_1 \log(t)$.

La mayoría de los estadísticos utilizan la técnica de mínimos cuadrados por su sencillez y sus propiedades deseables. Como resultante, también es utilizada en cuestiones legales. En este sitio de internet hay una presentación didáctica de esta técnica.

Ejemplo Un experto desea analizar los salarios de hombres y mujeres en una gran editorial a fin de descubrir si las diferencias de salarios entre los empleados con experiencia similar son evidencia de discriminación. Los resultados de regresión del ejemplo están basados en datos de 1,715 hombres y mujeres, que fueron usados por el defensor en un caso de discriminación sexual en contra del *New York Times* resuelto en 1978.

Para empezar con el caso más simple, Y, el salario medido en dólares anuales, es la variable dependiente que debe ser explicada, y X_1 es la variable explicativa – el número de años de experiencia del empleado. El modelo de regresión sería escrito así:

[4] $Y = \beta_0 + \beta_1 X_1 + \varepsilon.$

En esta ecuación, β_0 y β_1 son los parámetros que deben ser estimados con los datos, y ε es el término del error aleatorio. ¿Cuál es la interpretación de β_0? Es el salario medio de todos los empleados

[2] Resulta claramente ventajoso que los componentes aleatorios de la relación de regresión sean pequeños con relación a la variación de la variable dependiente.

que carecen de experiencia. ¿Y la interpretación de β_1? Mide el efecto promedio que un año de experiencia adicional tiene sobre el salario medio de los empleados.

Una vez que los parámetros de una ecuación de regresión, como la [3], han sido estimados, pueden calcularse los valores "ajustados". Si en la ecuación [3] denotamos los parámetros de regresión estimados, o *coeficientes de regresión*, como b_o, b_1,... b_k, los valores ajustados de Y, que denotaremos como <Y>, vendrán dados por la siguiente ecuación:

[5] $\quad \langle Y \rangle = b_o + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$

La Figura 4 ilustra esto con un ejemplo que involucra una sola variable explicativa. Los datos aparecen como en un diagrama de dispersión; el salario está en el eje vertical, y los años de experiencia en el eje horizontal. La línea de regresión estimada está dibujada a través del conjunto de puntos. Viene dada por:

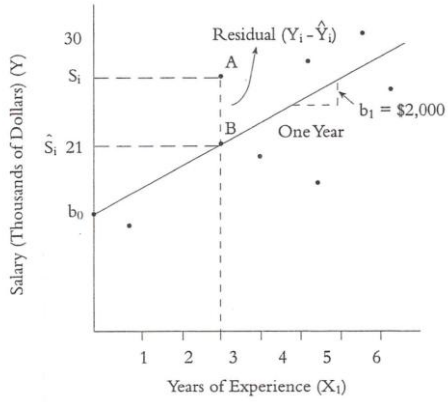


Figura 4 Bondad del ajuste

[6] $\langle Y \rangle = \$15{,}000 + \$2{,}000\ X_1$.

Luego el valor ajustado del salario asociado con X_1 años de experiencia de un individuo está dado por:

[7] $\langle Y_i \rangle = b_o + b_{1i} X_{1i}$ (punto B).

La *ordenada al origen* de la línea recta es el valor medio de la variable dependiente cuando la o las variables independientes son iguales a 0; esta ordenada al origen b_o aparece en el eje vertical de la Figura 4. En forma similar, la *pendiente de la línea* mide el cambio (promedio) de la variable dependiente asociado al incremento de 1 unidad de la variable explicativa. También está representada la pendiente b_1. Según la ecuación [6], la ordenada al origen igual a $15,000 indica que los empleados inexpertos ganan $15,000 por año. El parámetro de la pendiente implica que cada año de experiencia añade $2,000 al salario de un empleado "promedio".

Ahora supongan que la variable salario está relacionada con el sexo del empleado. La variable relevante indicativa, que a menudo es llamada una variable *dummy*, es ahora X_2, igual a 1 si el empleado es del sexo masculino y 0 si es del sexo femenino. Supongan que la regresión del salario con respecto a X_2 produce el siguiente resultado: $\langle Y \rangle = \$30{,}449 + \$10{,}979\ X_2$. El coeficiente $10,979 mide la diferencia entre el salario medio de los hombres y el salario medio de las mujeres.[3]

3. Residuos de Regresión

Para todo conjunto de datos puntuales, el residuo de regresión es la diferencia entre el valor observado y el valor ajustado de la variable dependiente. Supongan, por ejemplo, que estudiamos el caso

[3] Para apreciar por qué sucede así, observar que si X_2 es igual a 0, el salario medio de las mujeres es $30,449. Para los hombres (cuando X_2=1) el salario medio es $30,449 + $10,979 x 1 = $41,428. Luego, la diferencia es igual a $41,428 - $30,449 = $10,979.

de un individuo con 3 años de experiencia y un salario de $27,000. Según la línea de regresión de la Figura 4, el salario medio de un individuo con esa experiencia se ubica en $21,000. Luego, estamos en presencia de un residuo positivo, igual a $6,000. En términos generales, el residuo *e* asociado con un dato puntual como el punto A de la Figura 4, viene dado por $e_i = Y_i - \langle Y_i \rangle$. Cada punto de la figura tiene un residuo, que es el error cometido por el método de regresión mínimo-cuadrático con ese individuo.

4. No linealidades

Los modelos no lineales toman en cuenta la posibilidad de que la magnitud del efecto de una variable explicativa sobre la variable dependiente cambie a medida que cambia el nivel de la variable explicativa. Hay un modelo útil que produce este efecto, el modelo de interacción entre las variables. Por ejemplo, supongan que:

[8] $S = \beta_1 + \beta_2 \, SEXO + \beta_3 \, EXP + \beta_4 \, (EXP)\,(SEXO) + \varepsilon.$

En esta ecuación, S es el salario anual, SEXO es igual a 1 para las mujeres y a 0 para los hombres, EXP representa los años de experiencia laboral, y ε es un término de error. El coeficiente β_2 mide la diferencia del salario medio (para todos los niveles de experiencia) entre hombres y mujeres que no tienen experiencia. El coeficiente β_3 mide el efecto de la experiencia sobre el salario de los hombres (cuando SEXO=0), y el coeficiente β_4 mide la diferencia en el efecto de la experiencia sobre el salario de hombres y mujeres. Se desprende, por ejemplo, que el efecto de un año de experiencia sobre el salario de los hombres es β_3, mientras que el efecto comparativo para las mujeres es $\beta_3 + \beta_4$. *Nota:* Estimar una ecuación en la cual hay términos de interacción para todas las variables explicativas, como en la [8], es esencialmente lo mismo que estimar dos regresiones por separado, una para los hombres y otra para las mujeres.

5. Interpretación de los Resultados de una Regresión

Para ver cómo se interpretan los resultados de regresión, ampliamos el ejemplo anterior de la Figura 4 a fin de considerar la posibilidad de una variable explicativa adicional – el número de años de experiencia, X_3, elevado al cuadrado. Esta variable está pensada para captar el hecho de que para la mayoría de los individuos, los salarios se incrementan con la experiencia, pero eventualmente tienden a nivelarse. La línea de regresión estimada utilizando la tercera variable explicativa, así como la variable representativa de los años de experiencia (X_1) y la variable dummy de sexo (X_2), es la siguiente:

[9] $\langle Y \rangle = \$14{,}085 + \$2{,}323 \, X_1 + \$1{,}675 \, X_2 - \$36 \, X_3.$

El cambio de los coeficientes de regresión luego de incluir X_3 y X_1 ilustra la importancia de incluir variables explicativas relevantes. El coeficiente de regresión de X_2 mide la diferencia de salarios entre hombres y mujeres manteniendo constante el efecto de la experiencia. Este diferencial es sustancialmente más bajo que el medido previamente ($10,979). Si fallamos en controlar este efecto de la experiencia se sobre-estimará la diferencia de salarios entre hombres y mujeres.

Consideren ahora la interpretación de las dos variables de experiencia, X_1 y X_3. El signo positivo de X_1 muestra que el salario crece con la experiencia. El signo negativo de X_3 indica que la tasa de crecimiento del salario disminuye con la experiencia. Para calcular el efecto combinado de X_1 y X_3 podemos hacer algunos cálculos: por ejemplo, veamos cómo cambia el salario medio de las mujeres (X_2=0) a medida que cambia el nivel de experiencia. A medida que la experiencia crece desde 0 a 1 año, el salario medio crece en $2,251, desde $14,085 a $16,336. Sin embargo, las mujeres con 2 años de experiencia ganan sólo $2,179 más que las mujeres con 1 año de experiencia, y las mujeres con 3 años de experiencia sólo ganan $2,127 más que las que tienen 2 años de experiencia. Además, las que tienen 7 años de experiencia ganan $28,582 por año, que sólo representa $1,855 más que los $ 26,727 ganados por las mujeres con 6 o más años de experiencia.[4] La Figura 5 ilustra estos resultados; la línea de regresión representada corresponde a los salarios de las mujeres; la línea correspondiente a los hombres sería paralela y más elevada en $1,675.

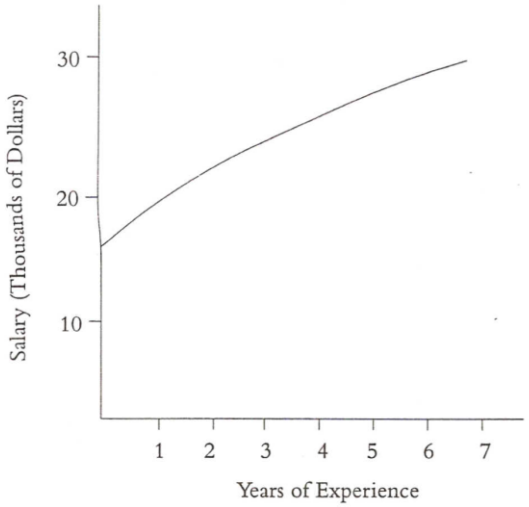


Figura 5 Pendiente de la regresión

6. Resultados de Regresión por MCO

La regresión mínimo-cuadrática proporciona no sólo estimadores de los parámetros que indican la dirección y magnitud del efecto de un cambio de la variable explicativa sobre la variable dependiente, sino además un estimador de la confiabilidad del estimador del parámetro y una medida global de bondad del ajuste del modelo de regresión.

Los estimadores de los parámetros verdaderos (pero desconocidos) de un modelo de regresión, son números que dependen de qué muestra de observaciones se utilizó para el estudio. Esto es, si se hubiera utilizado otra muestra, se hubiera calculado un estimador distinto (ya que la fórmula que genera los coeficientes es denominado el estimador mínimo-cuadrático, cuyo valor cambia según la muestra). Si el experto continúa recogiendo más y más muestras generando estimadores adicionales, como podría suceder si hubiera nuevos datos disponibles a lo largo del tiempo, los estimadores de cada parámetro tendrían una distribución de probabilidad (es decir, el experto podría determinar el porcentaje o frecuencia de tiempo que sucede un estimador). Esta distribución de probabilidad podría ser resumida por una media y una medida de dispersión en torno a la media, un desvío estándar, que habitualmente es denominado el error estándar del coeficiente o error estándar (SE). Por ejemplo, supongan que el experto está interesado en estimar el precio medio pagado por litro de nafta sin plomo por los consumidores de cierta zona de Argentina en un momento determinado del tiempo. El precio medio de una muestra de estaciones de nafta pudo haber sido $4,04, en otra muestra de $3,872 y en otra tercera de $4,264. Sobre esta base, el experto calcula el precio medio de la nafta sin plomo en surtidores de Argentina en $4,04 y un desvío estándar de $0,197.

[4] Estos guarismos surgen de sustituir distintos valores en la ecuación [9] para X_1 y X_3.

La regresión por mínimos cuadrados generaliza este resultado, calculando medias cuyos valores dependen de una o más variables explicativas. El error estándar de un coeficiente de regresión le dice al experto en cuánto es probable que los estimadores de los parámetros varíen de muestra en muestra. A mayor variación de los estimadores de los parámetros entre muestra y muestra, más elevado será el error estándar y, en consecuencia, menos confiable será el resultado de la regresión. Pequeños errores estándar implican resultados que son probablemente similares entre distintas muestras, mientras que grandes errores estándar son evidencia de gran variabilidad.

Bajo supuestos adecuados, los estimadores mínimo-cuadráticos son las "mejores" determinaciones de los verdaderos parámetros subyacentes.[5] De hecho, *los mínimos cuadrados tienen diversas propiedades deseables*. Primero, los estimadores mínimo-cuadráticos son *insesgados*. Esto significa, intuitivamente, que si la regresión fuera calculada una y otra vez con muestras diferentes, la media de los diversos estimadores de cada coeficiente obtenidos sería el verdadero parámetro. Segundo, los estimadores mínimo-cuadráticos son *consistentes*; si la muestra fuera muy grande, los estimadores estarían próximos a los verdaderos parámetros. Tercero, los estimadores mínimo-cuadráticos son *eficientes*, en el sentido de que los estimadores tienen la menor varianza de todos los posibles estimadores (lineales) insesgados.

Si además hacemos un supuesto sobre la distribución de probabilidad de cada uno de los términos de error, es posible enunciar algo acerca de la precisión de los coeficientes estimados. En muestras relativamente grandes (a menudo, unos 30 o 40 puntos serán suficientes para regresiones con un pequeño número de variables explicativas), la probabilidad de que el estimador de un parámetro esté dentro del intervalo de *2 errores estándar* del verdadero parámetro será aproximadamente 0,95, o sea 95%. Hay un supuesto que se hace a veces sobre el término de error, que no siempre es apropiado, consistente en que los parámetros siguen una distribución normal. Esta distribución tiene la propiedad de que el área comprendida entre 1.96 errores estándar de la media es igual al 95% del área total. Fíjense que *no es necesario hacer el supuesto de normalidad para aplicar mínimos cuadrados*, ya que la mayoría de las propiedades de los mínimos cuadrados surgen en forma independiente de la hipótesis de normalidad.

En general, para cualquier estimador paramétrico b, el experto puede construir un intervalo en torno a b en el cual hay una "masa" de probabilidad de 95% tal que el intervalo abarca al parámetro verdadero. El intervalo de confianza al 95% está dado por:

[10] $b \pm 1.96.$ (SE de b).

Ya sabemos que los intervalos de confianza son comúnmente utilizados en los análisis estadísticos, porque el experto nunca puede estar seguro de que el estimador del parámetro sea igual al verdadero parámetro de la población. El experto puede contrastar la hipótesis de que el parámetro en realidad es cero (llamada la hipótesis nula) mirando el estadístico-t, definido como:

---

[5] Los *supuestos del modelo de regresión* incluyen que: (a) el modelo esté correctamente especificado; (b) que los errores asociados con cada observación sean extracciones aleatorias de la misma distribución de probabilidad y que sean independientes unos de otros; (c) que los errores asociados con cada observación sean independientes de las observaciones correspondientes de cada una de las variables explicativas del modelo; y (d) que no haya ninguna variable explicativa perfectamente correlacionada con una combinación de otras variables.

[11] $t = b / \text{SE}(b)$

Si este estadístico-*t* resulta de magnitud inferior a 1.96, el intervalo de confianza al 95% alrededor de *b* debe incluir 0. Este estadístico-t es aplicable a muestras de cualquier tamaño. A medida que la muestra se hace más grande, la distribución subyacente, que es la fuente del estadístico-t (la distribución t de Student) se va aproximando a la distribución normal. Como esto significa que el experto no puede rechazar la hipótesis de que β=0, el estimador – cualquiera resulte su valor – se dice que no es estadísticamente significativo. Recíprocamente, si el estadístico-*t* es mayor que 1.96 en valor absoluto, el experto concluye que es improbable que el verdadero valor de β sea 0 y dice que este estimador es estadísticamente significativo (intuitivamente, *b* está "demasiado lejos" de 0 como para que sea consistente con β=0). En tal caso, el experto rechaza la hipótesis de que β=0 sea verdadera y dice que el estimador es estadísticamente significativo. Si la hipótesis nula β=0 es verdadera, usar un intervalo de confianza al 95% implicará que el experto rechace erróneamente la hipótesis nula 5% de las veces. Por consiguiente, decimos que los resultados son significativos al 5%. (Un estadístico-t de magnitud igual o mayor que 2.57 está asociado a un nivel de confianza del 99%, o un nivel de significación del 1%, que incluye una banda igual a 2.57 desvíos estándar a ambos lados de los coeficientes estimados.)

Como ejemplo, veamos un conjunto más completo de resultados de regresión asociados a la regresión del salario descripta en [9]:

[12] $\langle Y \rangle = \$14{,}085 + \$2{,}323\, X_1 + \$1{,}675\, X_2 - \$36\, X_3$

	(1,577)	(140)	(1,435)	(3,4)
t =	8,9	16,5	1,2	-10,8

El error estándar de cada parámetro estimado viene puesto entre paréntesis directamente debajo de cada coeficiente, mientras que el estadístico-*t* aparece debajo de cada error estándar.

Tomemos el coeficiente de la variable *dummy* X_2. Está indicando que $1,675 es la mejor estimación de la diferencia salarial promedio entre hombres y mujeres. Empero, su error estándar es amplio ($1,435 con respecto al parámetro $1,675). Como el error estándar es relativamente amplio, el rango de valores posibles para medir la verdadera diferencia salarial (el verdadero parámetro) es grande. De hecho, un intervalo de confianza al 95% viene dado por:

[13] $\$1{,}675 \pm \$1{,}435 \times 1.96 = \$1{,}675 \pm \$2{,}813$.

En otras palabras, el experto puede tener 95% de confianza de que el valor verdadero del coeficiente esté comprendido entre - $1,138 y $4,488. Como este intervalo incluye al 0, el efecto del sexo sobre el salario se dice que *no es estadísticamente significativo al 5% de significación*.

Observen que la experiencia es una variable de alta significación en el salario, ya que X_1 y X_3 tienen variables *t* de magnitud sustancialmente mayor que 1.96. Mayor experiencia tiene un efecto significativo sobre el salario, pero el tamaño de este efecto tiende a disminuir significativamente con la experiencia.

La información proporcionada por los resultados de regresión contiene no sólo los estimadores puntuales de los parámetros y sus errores estándar o estadísticos-*t*, sino además otra información que nos dice cuán buena es la aproximación de la línea de regresión a los datos. Hay un estadístico, llamado el error estándar de regresión (SER) que es un estimador del tamaño promedio de los residuos de regresión.[6] Un SER=0 significaría que *todos los puntos de los datos yacen exactamente sobre la línea de regresión* – lo cual es algo prácticamente imposible. A otras cosas iguales, a mayor SER, peor será el ajuste de los datos del modelo.

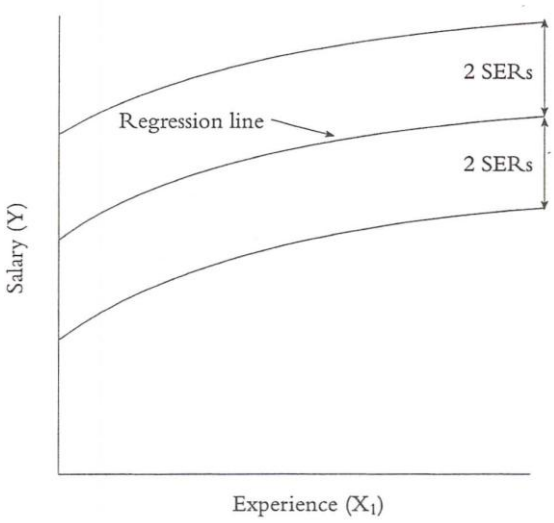


Figura 7 Error estándar de regresión

Si el término del error está distribuido en forma normal, el experto podría esperar que aproximadamente 95% de los puntos de los datos estén ubicados a una distancia de 2 SERs de la línea de regresión, como se muestra en la Figura 7 (en esta figura, SER ≈ $5,000).

El estadístico R^2 mide el porcentaje de variación de la variable dependiente que es tenido en cuenta por todas las variables explicativas. Luego, R^2 facilita una medida global de bondad del ajuste. Su valor oscila entre 0 y 1. Un $R^2=0$ significa que las variables explicativas no explican nada de la variación de la variable dependiente.; si $R^2=1$, las variables explicativas explican toda la variación. El R^2 de la ecuación [12] es .56, lo cual implica que las tres variables explicativas dan cuenta del 56% de la variación de los salarios. Hay que tener en cuenta que R^2 y SER dan aproximadamente la misma información, ya que R^2 es más o menos igual a 1 – SER^2/Varianza de Y. La variación es computada como el cuadrado de la diferencia entre cada Y y el Y medio, sumado a lo largo de todas las observaciones.

¿Hay alguna forma de saber el R^2 que indica que el modelo es satisfactorio? Lamentablemente no existe una respuesta clara a esta pregunta, dado que la magnitud de R^2 depende de los datos usados y, en particular, de si los datos cambian a través del tiempo o entre los individuos. Es típico un R^2 bajo en estudios de sección cruzada en los que se busca explicar estas diferencias. Es muy probable que las diferencias individuales sean causadas por varios factores que no pueden medirse. De resultas, el experto no tiene que esperar poder explicar gran parte de la variación. En contraste, en los estudios de series de tiempo, el experto se encuentra explicando movimientos de agregados a lo largo del tiempo. Como la mayoría de las series tienen un crecimiento sustancial, o tendencia, común a todas ellas, no resulta difícil "explicar" una serie temporal utilizando otra serie temporal, simplemente porque se mueven en forma conjunta. Se desprende en calidad de corolario que *un elevado R^2 de por sí no significa que las variables incluidas en el modelo sean las adecuadas*.

[6] Específicamente, es una medida del desvío estándar del error de regresión, *e*. A veces se lo llama error cuadrático medio de la línea de regresión.

Por regla general, los tribunales deberían evitar basarse exclusivamente en un estadístico como el R^2 para elegir un modelo en lugar de otro. El experto debería hurgar especialmente en el comportamiento de los residuos (estadístico de Durbin-Watson) y otras propiedades como los *F*-test.

La línea de regresión mínimo-cuadrática puede ser sensible a los puntos extremos. Esto se puede apreciar en la Figura 8. Supóngase que inicialmente hay tres puntos (A, B, y C), que vinculan la información de la variable X_1 con la variable Y. La línea 1 representa la mejor regresión entre estos puntos. El punto D es un valor atípico porque se encuentra muy alejado de la línea de regresión que ajusta a los puntos restantes. Si se re-estima la línea de regresión mínimo-cuadrática incluyendo ahora el punto D, se obtiene la Línea 2. Esta figura muestra que D es un dato influyente, ya que tiene un efecto dominante tanto sobre la pendiente como sobre la ordenada al origen de la línea de mínimos cuadrados. Como en mínimos cuadrados se trata de minimizar la suma de los desvíos al cuadrado, la sensibilidad de la línea a estos puntos individuales puede ser a veces sustancial. No siempre es indeseable la insensibilidad, por ejemplo cuando es mucho más importante predecir un punto como D cuando se produce un gran cambio, en comparación con medir cambios pequeños.

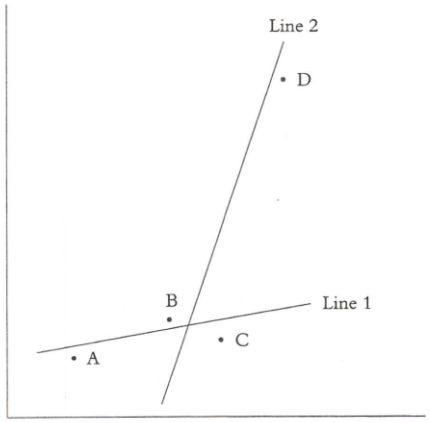


Figura 8 Línea de regresión MC

Lo que torna más difícil el problema de los datos influyentes es que el efecto de un valor atípico puede no ser apreciado cómodamente si el desvío es medido a partir de la línea de regresión final. El motivo es que la influencia del punto D sobre la línea 2 es tan sustancial que *su desvío de la línea de regresión no es necesariamente mayor que el desvío de los puntos restantes de la línea de regresión*. La importancia de un valor atípico también depende de su ubicación en el espacio de datos. Es probable que los valores atípicos asociados con valores relativamente extremos de las variables explicativas sean muy influyentes. Ver, p.ej., Fisher v. Vassar College, 70 F.3d 1420, 1436 (2d Cir. 1995) (el tribunal requirió evaluar el "servicio en la comunidad académica", porque el concepto era demasiado amorfo y no constituía un factor significativo para revisar la continuidad de la cátedra). Aunque no tan populares como los mínimos cuadrados, hay otras técnicas de estimación alternativas menos sensibles a los valores atípicos, como la regresión robusta.

7. Lectura del Output de Computadora de Regresión Múltiple

Como aplicación, vamos a analizar los resultados de Ernesto Gaba y Lucio Reca de estimación de la demanda de carne vacuna en Argentina entre 1950 y 1972,[7] que construyeron un índice de precios de un conjunto de alimentos que componen una canasta suficientemente amplia de posibles sustitutos de la carne vacuna e incorporaron el posible efecto de la veda sobre el consumo mediante una

[7] Lucio G. Reca y Ernesto Gaba, Poder adquisitivo, veda y sustitutos: un reexamen de la demanda interna de carne vacuna en la Argentina, 1950-1972, Desarrollo Económico 50, vol. 13, julio-sept. 1973. Reproducido en J. C. de Pablo y F. V. Tow (eds.), Lecturas de Microeconomía por Economistas Argentinos, Buenos Aires, 1975. Ver Enrique A. Bour, Tratado de Microeconomía, 2009, pág. 140-144.

variable binaria que adopta el valor 1 en los años en que rigió algún tipo de veda y 10 en los restantes. Para otra variable explicativa – el ingreso por habitante – usaron el ingreso medio de la población a partir de las series de producto bruto nacional y de población. Finalmente, el precio de la carne vacuna fue introducido en el modelo de dos formas alternativas: 1) deflactado (dividido) por el índice de precios implícitos en el producto; 2) como precio relativo a los sustitutos. Reca y Gaba optaron por usar variables expresadas en logaritmos (lo que confiere a las derivadas parciales de la función demanda de carne el sentido de elasticidades), donde:

[14] $\varepsilon_{cv,\,pv} = -\,(\partial \log cv / \partial \log pv)$

representa la elasticidad de la demanda de carne vacuna (cv) al precio minorista de la misma (pv, deflactado por el índice de precios implícitos del Banco Central o, según se verá luego, por un índice de precios de los sustitutos). Definiciones similares permiten definir la elasticidad de la demanda de carne vacuna ante el precio de los sustitutos ps (usando un deflactor similar) $\varepsilon_{cv,ps}$ (con un signo *a priori* positivo); la elasticidad-ingreso *per capita* de la demanda de carne vacuna $\varepsilon_{cv,y}$. El Cuadro 2 la página 141 de mi *Tratado* resume las principales ecuaciones estimadas. Cabe tener en cuenta que: 1º) Se utilizaron especificaciones logarítmicas porque ello permite un cálculo simple y directo de las elasticidades; 2º) Los autores incorporaron el posible efecto de la veda sobre el consumo mediante una variable binaria = 1 en los años que rigió algún tipo de veda e = 10 en los restantes; 3º) Para otra variable explicativa – el ingreso por habitante – usaron el ingreso medio de la población medido por las series de producto bruto nacional y de población; 4º) Finalmente, el precio de la carne vacuna fue introducido en el modelo de dos formas alternativas: primero, deflactado por el índice de precios implícitos en el producto; segundo, relativo a los sustitutos.

Las variables presentadas son habituales en los modelos econométricos, p.ej. los *t*-Student de los distintos coeficientes; el coeficiente R^2_c es el porcentaje de variación de la variable dependiente del que dan cuenta las variables independientes incluidas (idéntico al R^2 ajustado por el número de grados de libertad); el coeficiente SE es el error típico de estimación; el coeficiente DW es el estadístico de Durbin y Watson utilizado para probar la presencia de auto-correlación de los residuos. No se incluyen, aunque es habitual hacerlo, los coeficientes de la prueba *F*; el *p*-valor de la regresión conjunta y de cada coeficiente por separado.

La veda, representada en el modelo por la variable binaria, muestra una elevada significación estadística y una considerable estabilidad en todas las regresiones. La primera veda al consumo interno fue en 1952, y si bien fue suspendida 3 años después, de hecho cesó de funcionar aproximadamente al año y medio de ser aplicada. La segunda veda (1964 y 1965) introdujo la restricción en la veda de carne vacuna 2 días por semana, y la última, a partir de marzo de 1971, si bien con interrupciones y contramarchas, se caracterizó por fijar la duración del período de veda en una semana, hasta su derogación en abril de 1973. Los estímulos para la faena clandestina de ganado tal vez incidan en una sobreestimación del coeficiente de la variable binaria, sin embargo. Es decir, que las cifras tomadas para confeccionar el análisis de regresión podrían ser inferiores a las reales y, por consiguiente, la variable binaria recoger además del efecto propio de la veda la subestimación proveniente del consumo no registrado. En cuanto a la magnitud del coeficiente, el promedio simple de la diferencia de consumo con veda y sin veda en los años 1952, 1964 y 1971 arroja una reducción de consumo por habitante y año de 5,7 kg de carne. En otras ecuaciones, la reducción del consumo imputable a la veda arroja 9,5 kg por habitante y por año. Lo cual implica una caída del consumo

ligeramente superior a 10%. Para que se logre un efecto similar dejando operar al sistema de precios sin interferencias, dada la elasticidad propia de la demanda de carnes, hubiera sido necesario un aumento del precio de la carne de 35%.

Una de las mejores ecuaciones es la C2, que se escribe de la forma siguiente:

[15] $$\log(cv) = \underset{(1,3)}{1{,}245} + \underset{(6,4)\,**}{0{,}159}\ \text{veda} - \underset{(8,9)\,**}{0{,}369}\log(pv) + \underset{(0,07)}{0{,}048}\log(ps) + \underset{(3,8)\,**}{0{,}38}\log(y)$$

$R^2_{aj} = 0{,}93$
SER = 0,033
DW = 2,03

La elasticidad-precio propia de la carne es, en esta ecuación, -0,369. La elasticidad-ingreso es igual a 0,38 (es usual una elasticidad-ingreso inferior a la unidad en bienes que satisfacen necesidades primarias). El doble asterisco (**) indica que los dos coeficientes difieren en forma significativa de cero al 99% de probabilidad. En cambio, el precio de los sustitutos no es estadísticamente significativo, aunque tenga el signo esperado *a priori*. El valor del SER, como la variable dependiente viene expresada en logaritmos, puede considerarse como aproximando el error de la ecuación en tanto por uno (es decir, se tendría un error del 3.3%).

El documento de Gaba y Reca llega hasta 1972, e ignoro si hay nuevas estimaciones. El sector de la carne vacuna se ha caracterizado por su alta complejidad a lo largo de los principales eslabones que integran la cadena: producción, industria, distribución y consumidor. La Argentina ocupaba hasta el año 2000 (2001 resultó totalmente atípico con el cierre de casi el 98% de los mercados de carnes frescas) el *quinto* lugar como productor y el *sexto* lugar como exportador, *siendo el país con mayor consumo de carnes por habitante*; por tal razón, el arraigo de este producto en el país hace que 85% de la producción total sea consumida localmente y el resto se exporte con una participación promedio, en la última década, de 700 millones de dólares por año. La ganadería vacuna participa en 18% del PBI agropecuario y en 3% del PBI total, la carne de vaca representa aproximadamente 68% del consumo total de carnes en nuestro país, y 7,1% del gasto total en alimentos por habitante. El negocio de la carne vacuna, a la salida del frigorífico, tiene una facturación aproximada de 6.500 millones de pesos anuales. Considerando las exportaciones totales del 2000 en 25.000 millones de dólares, su participación era 2,5%.

Un punto final que cabe señalar es que la participación de los supermercados introdujo tanto una nueva modalidad de compra para el consumidor como también cambios en las tradicionales reglas de juego de la comercialización minorista: la concentración de mercadería de los supermercados conllevó nuevos plazos de pago y condiciones de compra que se extendieron a toda la cadena de comercialización, llegando inclusive al productor que vio reducida su rentabilidad. En la actualidad se estima que entre 35 y 40% de la venta de carnes se concentra en este canal y sólo en Capital Federal representa 60% de la comercialización de carnes vacunas. Hay supermercados que tienen sus propias plantas de faena, lo que les brinda un mayor control del negocio hacia adelante y hacia atrás en la cadena y también elementos clave como la seguridad en la cadena de frío, calidad continua (contratos con los productores en forma directa), mejor distribución de los cortes por zona de consumo, venta de cortes con marca propia y diferente presentación, etc.

La comercialización tradicional de carne en el mercado interno se basaba en el traslado de la media res salida de la planta de faena y transportada hacia cada boca de expendio en camiones refrigerados. Esta modalidad a principios de los 1990s representaba 95% y el resto se basaba en la distribución de cortes, pero esta modalidad de venta mostró un fuerte cambio ya que la relación pasó de 95 % a 75 %. Ambos mercados constituyen mundos muy diferentes, en el mercado de la media res las categorías más utilizadas son vaquillona, ternero y novillito colocadas directamente en el local de expendio para su desposte, siendo la modalidad utilizada por los frigoríficos consumeros. De esta forma se incrementaron los beneficios obtenidos en los mercados de alto valor. Estos frigoríficos proveen de carne a supermercados e hipermercados, algunos a través del ingreso a una nómina de proveedores continuos con especificaciones en cuanto a calibre de los cortes; los plazos más comunes oscilan entre 30 y 45 días. Los operadores del mercado interno son los matarifes carniceros.

8. Proyecciones

Volvamos a nuestro ejemplo salarial. Una proyección es una predicción sobre los valores que alcanzará una variable dependiente usando información sobre las variables independientes. Frecuentemente lo que se hace es una proyección *ex ante*: en tal caso, los valores de la variable dependiente son predichos más allá de la muestra (es decir, más allá del período en que el modelo fue estimado). Empero, las proyecciones *ex post* son utilizadas a menudo en el análisis de daños. En los casos que implican daños, surge la pregunta frecuente de cómo hubiera sido el mundo si cierto evento no hubiera sucedido. Por ejemplo, en la fijación de precios anti-monopolística, el experto puede preguntarse cuál hubiera sido el precio de un producto si cierto evento asociado con la fijación de precios no hubiera ocurrido. Si los precios hubieran sido más bajos, la evidencia sugiere un impacto. Si el experto puede predecir cuán bajos hubieran podido llegar a ser, los datos pueden ayudarlo a desarrollar un estimador numérico de los daños. Una proyección *ex post* se caracteriza porque en el horizonte de proyección todas las variables dependientes y explicativas son conocidas; las proyecciones ex post pueden ser chequeadas comparando con los datos existentes y facilitar un método directo de evaluación.

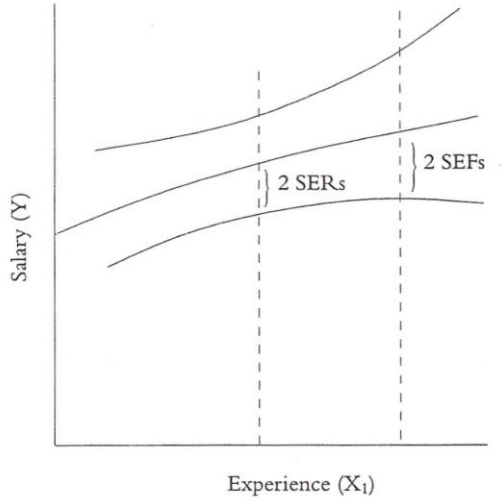


Figura 9 Error Estándar de Proyección

Por ejemplo, a fin de calcular una proyección de la regresión salarial de más arriba, el experto utiliza la ecuación de salarios estimada siguiente:

[16] $\langle Y \rangle = \$14{,}085 + \$2{,}323\ X_1 + \$1{,}675\ X_2 - \$36\ X_3$

Para predecir el salario de un hombre con 2 años de experiencia, el experto calcula:

[17] $\langle Y\ (2) \rangle = \$14{,}085 + (\$2{,}323 \times 2) + (\$1{,}675) - (\$36 \times 2) = \$20{,}262$

El grado de precisión de las proyecciones ex ante y ex post puede calcularse si la especificación del modelo es correcta y los errores están distribuidos normalmente y son independientes. Al estadístico se lo conoce como error estándar de proyección (SEF, por *standard error of forecast*). El SEF mide el desvío estándar del error de proyección cometido dentro de una muestra de la que se conocen con certeza las variables explicativas.[8] El SEF puede usarse para determinar cuán precisa es una proyección. En [17] el SEF asociado a la proyección es aproximadamente $5,000. Si se utiliza una muestra amplia, la probabilidad es, aproximadamente, de 95% de que el salario predicho esté comprendido entre 1.96 errores estándar del valor proyectado. En tal caso, el intervalo apropiado al 95% va de $10,822 a $30,422. Como el modelo estimado no explica efectivamente los salarios, el SEF es amplio, como lo es el intervalo al 95%. Un modelo más completo con variables explicativas adicionales resultará en un SEF más reducido y un intervalo al 95% más pequeño de predicción.

Hay un peligro en usar el SEF, que también se aplica a los errores estándar de los coeficientes estimados. El SEF se calcula basándose en el supuesto de que el modelo incluye el conjunto correcto de variables explicativas y usando la forma funcional correcta. Si hemos cometido un error al elegir las variables o la forma funcional, el error de proyección estimado es engañoso. Hay ocasiones en que puede ser menor, quizá sustancialmente menor, que el verdadero SEF; en otras, puede ser más grande, por ejemplo si las variables incorrectas terminan capturando los efectos de las variables correctas.

En la Figura 9 se aprecia la diferencia entre el SEF y el SER. El SER mide desvíos dentro de la muestra. El SEF es más general, dado que calcula desvíos dentro o fuera del período muestral. En general, la diferencia entre el SEF y el SER aumenta a medida que aumenta la distancia de las variables explicativas de sus valores medios. La Figura 9 muestra el intervalo de predicción al 95% creado midiendo 2 SEF en torno a la línea de regresión.

[8] En realidad hay dos fuentes de error implícitas en el SEF. Una surge porque los parámetros estimados pueden no ser exactamente los verdaderos parámetros. Otra es el propio término de error; cuando se proyecta, típicamente el experto hace que este error sea igual a 0 aunque un trastorno de eventos no tomados en cuenta en el modelo de regresión podrían requerir que el error fuera positivo o negativo. En la metodología aplicada, los que efectúan proyecciones usualmente manipulan el término aleatorio uno o dos períodos fuera de la muestra para tomar en cuenta ciertos factores aleatorios, que al entrar aditivamente suelen ser conocidos como *add factors*.

Apéndice 1 - Tablas

Para comenzar el análisis de este apéndice, voy a sugerir un repaso de conceptos introductorios que hemos venido brindando en capítulos previos. He utilizado material producido por StatSoft Electronic Statistics Textbook, StatSoft, Inc. (2013). Tulsa, OK, que puede ser bajado de internet.

1. La Función Normal Estándar (Z)

La distribución Normal Estándar es utilizada en tests de hipótesis que incluyen contrastes sobre medias aisladas, diferencia entre dos medias, y contrastes sobre proporciones. La Normal Estándar tiene 0 como media y desvío típico de 1. Ustedes pueden ver una animación que muestra varias colas (a la izquierda) de esta distribución. Más información sobre la Distribución Normal tal como es usada en el contraste estadístico puede ser hallada repasando algunos conceptos elementales. También hallarán un desarrollo matemático algo más avanzado.

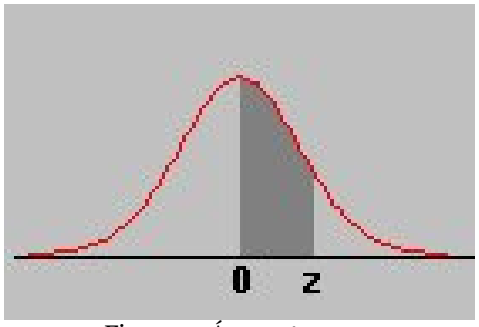


Figura 10 Área entre 0 y z

Como se muestra en la Figura 10, los valores de la tabla representan el área por debajo de la curva normal estándar entre 0 y el número relativo z. Por ejemplo, a fin de calcular el área entre 0 y 2.36, hay que buscar en la tabla de la Normal la celda de intersección de la fila indicada 2.30 y la columna indicada 0.06. El área resultante es 0.4909. Para calcular el área entre 0 y un valor negativo, observar que la simetría de la distribución normal requiere usar el valor positivo correspondiente. Por ejemplo, el área de la curva entre -1.3 y 0 es igual al área de la curva entre 1.3 y 0, de modo que hay que buscar la celda correspondiente a la fila 1.3 y la columna 0.00 (el área es 0.4032).

2. La Distribución *t* de Student

La forma de la distribución t de Student queda determinada por el número de grados de libertad. En esta animación, su forma cambia al aumentar el número de grados de libertad. Para ver cómo esta distribución es usada en el contraste de hipótesis, véase el test *t* de muestras independientes y el test *t* de muestras dependientes, en Estadísticos y Tablas Básicas. Como indica la Figura 11, el área en la parte de arriba de la tabla es el área de la cola derecha del valor-*t* que está en tabla. Para calcular el valor crítico al 0.05 de la distribución *t* con 6 grados de libertad, por ejemplo, hay que fijarse en la columna 0.05 y en la fila 6: $t_{(.05,\ 6)} = 1.94$.

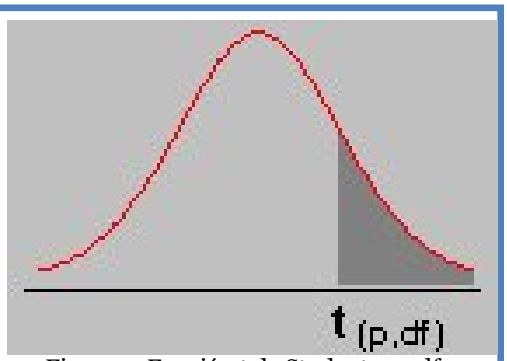


Figura 11 Función t de Student con df grados de libertad

3. La distribución χ^2 (chi cuadrado)

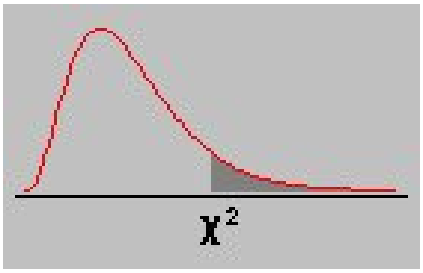


Figura 12 Función chi cuadrado

Al igual que la distribución *t* de Student, la forma de la distribución chi cuadrado queda determinada por sus grados de libertad. La animación muestra cómo se transforma la curva de la distribución chi cuadrado a medida que los grados de libertad aumentan (1, 2, 5, 10, 25 y 50). Como se indica en la Figura 12, los valores de esta tabla son valores críticos de la distribución chi cuadrado con los grados de libertad correspondientes. Para calcular el valor de una distribución chi cuadrado (con determinados grados de libertad) que contiene cierta área, hay que ir a la columna del área dada y a la fila de los grados deseados de libertad. Por ejemplo, el valor crítico al 25% de una chi cuadrado con 4 grados de libertad es 5.38527. Lo cual significa que el área a la derecha de 5.38527 en una distribución chi cuadrado con 4 grados de libertad es 25%.

4. Grados de Libertad

Hemos usado el concepto de "grados de libertad" como el de un estimador del número de categorías independientes en un test particular o experimento estadístico. Se encuentran mediante la fórmula n−r, donde n=número de sujetos en la muestra y r es el número de sujetos o grupos estadísticamente dependientes.

Cuando uno ajusta un modelo estadístico a un conjunto de datos, los residuos -expresados en forma de vector- se encuentran habitualmente en un espacio de menor dimensión que aquel en el que se encontraban los datos originales. Los grados de libertad del error los determina, justamente, el valor de esta menor dimensión. Por ejemplo, supongan que hay *n* variables aleatorias Y_1, Y_2,... Y_n, que constituyen una muestra de datos con media muestral igual a ‹Y›. La media muestral viene expresada como $\langle Y \rangle = (\sum_{i=1}^{n} Y_i)/n$. Entonces las cantidades Y_i - ‹Y› son los residuos, que pueden ser considerados estimadores de los errores $Y_i - \mu$ (donde μ es la media de la población). La suma de los residuos (a diferencia de la suma de los errores, que no es conocida) es necesariamente cero, pues $\sum_{i=1}^{n} (Y_i - \langle Y \rangle) = \sum_{i=1}^{n} Y_i - n \langle Y \rangle = \langle Y \rangle - \langle Y \rangle = 0$, ya que existen variables con valores superiores e inferiores a la media muestral. Esto también significa que los residuos están restringidos a encontrarse en un espacio de dimensión n − 1 (en este ejemplo, en el caso general a n − r) ya que, si se conoce el valor de n − 1 de estos residuos, la determinación del valor del residuo restante es inmediata. Luego, decimos que "el error tiene n − 1 grados de libertad" (o que el error tiene n − r grados de libertad en el caso general).

5. La Distribución F

La distribución F (distribución F de Snedecor o distribución F de Fisher-Snedecor) es una distribución de probabilidad continua. Esta distribución está sesgada hacia la derecha y es utilizada frecuentemente en lo que se conoce como Análisis de la Varianza (ver ANOVA/MANOVA). Es un cociente de dos distribuciones chi cuadrado, y una distribución F específica se denota indicando los grados de libertad de la chi cuadrado del numerador y los grados de libertad de la chi cuadrado del denominador. Este es un ejemplo de una distribución $F_{(10, 10)}$. En las cuatro tablas de la función F, las filas representan los grados de libertad del denominador y las columnas los del numerador. El área de la cola a la derecha le da el nombre a la tabla (p.ej., .025). Ejemplo: para determinar el

valor crítico al 5% de una distribución F con 7 y 12 grados de libertad, hay que fijarse en la columna 7 (numerador) y la fila 12 (denominador) para un alfa=0.05. $F_{(0.05,\ 7,\ 12)} = 2.9134$. Las cuatro tablas corresponden a los valores críticos .10, .05, .025 y .01. La Figura 13 representa una función F al 10% con *df1* grados de libertad en el numerador y *df2* grados de libertad en el denominador. El test F es utilizado para calcular la probabilidad unilateral de la eventualidad de que dos varianzas sean distintas.

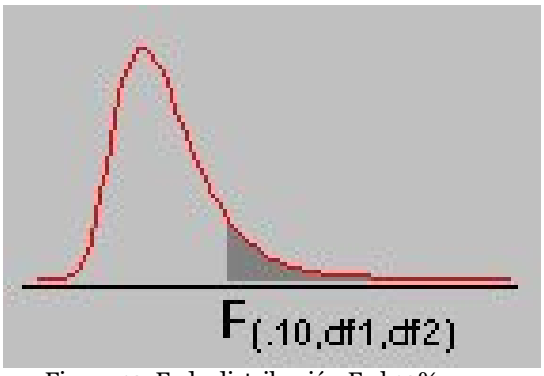


Figura 13. F. de distribución F al 10% y con df1 y df2 grados de libertad

Apéndice 2 – Causalidad en econometría

El **test de causalidad de Granger** es un test de hipótesis estadístico para determinar si una serie de tiempo es útil al pronosticar otra, propuesto por primera vez en 1969.[9] Por lo general, las regresiones reflejan "simples" correlaciones, pero Clive Granger argumentó que la causalidad en economía podría ponerse a prueba midiendo la capacidad de predecir los valores futuros de una serie de tiempo utilizando los valores anteriores de otra serie de tiempo. Dado que la cuestión de la "verdadera causalidad" es de raíz filosófica, y debido a la falacia *post hoc ergo propter hoc* de asumir que una cosa precedente a otra puede ser utilizada como una prueba de causalidad, los económetras afirman que la prueba de Granger sólo halla la "causalidad predictiva".[10] Una serie de tiempo X se dice que **causa en sentido de Granger** a Y si se puede demostrar, en general a través de una serie de tests t y F aplicados sobre los valores retardados de las X (y con valores retardados de Y también incluidos), que esos valores X proporcionan información estadísticamente significativa sobre valores futuros de Y.


Sir Clive William John Granger (1934-2009)
Nobel 2009 Nobelprize.org

La primera aplicación importante de causalidad de Granger a la economía aparece en un artículo de 1972 de Christopher Sims en el que demostró que el dinero *causa en sentido de Granger* al PBN nominal, al parecer, reforzando la idea monetarista de que las fluctuaciones monetarias son la causa principal de los ciclos económicos. En el debate que siguió, se aclararon los límites de la causalidad de Granger: el concepto se refiere a la predictibilidad y no al control, por lo que una constatación de que el dinero causa en sentido de Granger al PBN *no implica que la Reserva Federal tenga un instrumento eficaz para dirigir la economía*. Mientras que el propio Granger se había referido simplemente a "causalidad", el adjetivo "en sentido de Granger" hoy va unido a su idea para distinguirla de la causalidad basada en el control.

El propio Granger comenta:

El tema de cómo definir la causalidad ha mantenido ocupados a los filósofos por más de dos mil años y aún no se ha resuelto. Es una pregunta profunda enrevesada con muchas respuestas posibles que no satisfacen a todos, y sin embargo sigue siendo de cierta importancia. Los investigadores desearían pensar que han hallado una "causa", es decir una relación profunda, fundamental y, posiblemente, potencialmente útil. A principios de la década de los 1960s yo estaba considerando un par de procesos estocásticos relacionados que estaban claramente relacionados entre sí y quise saber si esta relación podría ser dividida en un par de relaciones en una sola dirección. Me sugirieron mirar una definición de causalidad propuesta por un famoso matemático, Norbert

[9] C. W. J. Granger, *Investigating Causal Relations by Econometric Models and Cross-spectral Methods* (Econometrica, 1969).
[10] Francis X. Diebold, Elements of Forecasting, 1998.

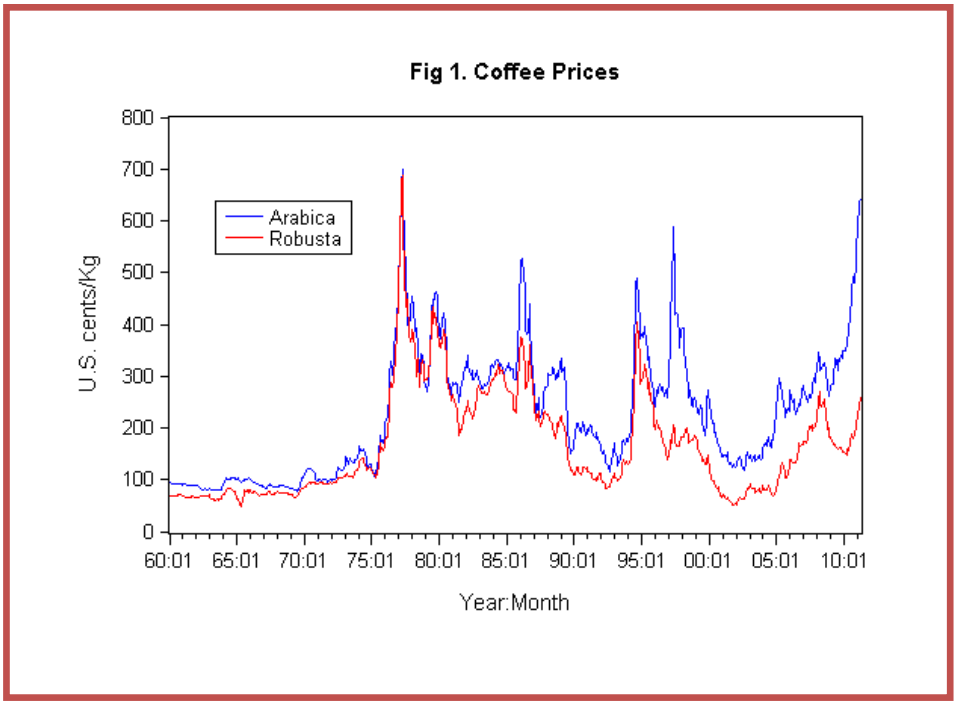
Wiener, así que adapté esta definición de forma práctica y la discutí. Los economistas aplicados hallaron la definición comprensible y utilizable y comenzaron a aparecer aplicaciones de la misma. Sin embargo, varios autores afirmaron que "por supuesto, esto no es causalidad real, es solamente causalidad de Granger." Por lo tanto, desde el principio, las aplicaciones utilizan este término para distinguirla de otras posibles definiciones.

La definición básica de "causalidad de Granger" es bastante simple. Supongamos que tenemos tres términos, X_t, Y_t, y W_t, y que se intenta primero pronosticar X_{t+1} utilizando términos anteriores de X_t y W_t. Luego tratamos de predecir X_{t+1} utilizando los términos pasados de X_t, Y_t, y W_t. Si se determina que el segundo pronóstico es más exitoso, de acuerdo con funciones de costo estándar, entonces el pasado de Y parece contener información para ayudar en el pronóstico X_{t+1} que no está en el pasado X_t o W_t. En particular, W_t podría ser un vector de posibles variables explicativas. Por lo tanto, Y_t sería "causa en sentido de Granger" de X_{t+1} si (a) Y_t se produce antes de X_{t+1}; y (b) contiene información útil para pronosticar X_{t+1} que no se encuentra en un grupo de otras variables apropiadas.

Naturalmente, cuanto más amplio sea W_t, y más cuidadosamente su contenido sea seleccionado, más estricto será el criterio que deberá pasar Y_t. Eventualmente, Y_t podría parecer que contenga información única sobre X_{t+1} que no se encuentre en otras variables por lo que la etiqueta de "causalidad" tal vez sea apropiada. La definición se apoya fuertemente en la idea de que la causa aparece antes que el efecto, que es la base de la mayoría, pero no todas, las definiciones de causalidad. Algunas implicancias son que es posible que Y_t sea causa de X_{t+1} y que X_t sea causa de Y_{t+1}, un sistema de retroalimentación estocástico. Sin embargo, no es posible que un proceso determinado, como una tendencia exponencial, sea una causa o sea causado por otra variable. Es posible formular pruebas estadísticas para lo que ahora se denomina G-causalidad, y muchas están disponibles y se describen en algunos libros de texto de econometría. La definición ha sido ampliamente citada y aplicada, ya que es pragmática, fácil de entender y de aplicar. En general se acepta que no refleja todos los aspectos de la causalidad, sino los suficientes como para merecer la pena de que sea considerada en una prueba empírica. (Granger, en scholarpedia).

Resumiendo, la causalidad de Granger es un concepto estadístico de causalidad que se basa en la predicción. De acuerdo con la causalidad de Granger, si una señal X_1 es "Granger-causa de" (o "G-causa") una señal X_2, entonces los valores pasados de X_1 deben contener información que ayude a predecir X_2 por encima y más allá de la información contenida en los valores pasados de X_2. Su formulación matemática se basa en un modelo de regresión lineal de procesos estocásticos. Existen extensiones más complejas a casos no lineales; sin embargo, estas extensiones son a menudo más difíciles de aplicar en la práctica.

Ejemplo 1 Como un primer ejemplo, véase el blog econométrico de Dave Giles: *Econometrics Beat: Dave Giles' Blog*. “Vamos a echar un vistazo a los precios mundiales de los cafés Arábica y Robusta. He aquí una representación gráfica de los datos mensuales desde enero 1960 a marzo 2011 - una bonita serie de largo plazo con una gran cantidad de observaciones:



Luego de diversos análisis usando el paquete de econometría *EViews*, el autor concluye que "se tiene evidencia razonable de causalidad de Granger *desde* el precio del café Arábica *al* precio del café Robusta, pero no viceversa." Estos análisis presuponen un conocimiento de técnicas de cointegración, por cuyo motivo no se expone el razonamiento aquí.

Ejemplo 2 En el campo de la neurociencia ha habido aplicaciones importantes. Diferentes métodos de obtención de alguna medida del flujo de información de las actividades de disparo de una neurona y su conjunto circundante han sido explorados en el pasado, pero son limitados por los tipos de conclusiones que se pueden extraer y proporcionan poca información acerca del flujo direccional de la información, el tamaño del efecto, y cómo puede cambiar con el tiempo. Recientemente la causalidad de Granger se ha aplicado a abordar algunas de estas cuestiones con gran éxito. Dicho en forma simple, se examina cómo predecir mejor el futuro de una neurona: utilizando todo el conjunto o el conjunto entero, excepto cierta neurona diana. Si la predicción se ve agravada por la exclusión de la neurona diana, entonces se dice que tiene una relación "g-causal" con la neurona actual (ver Figura 14). En el campo de la resonancia magnética funcional, se ha aplicado la G-causalidad a

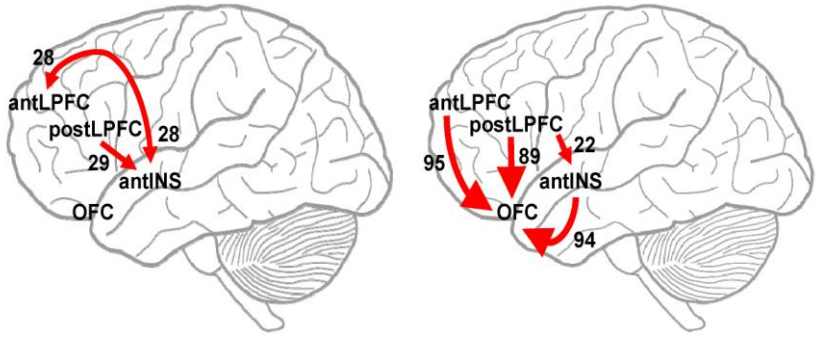
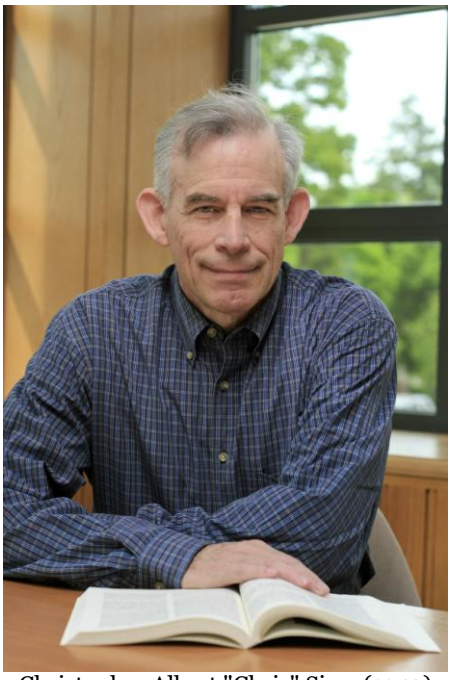


Figura 14 Un ejemplo de G-causalidad estimada a partir de datos de fMRI

los datos adquiridos durante una tarea compleja visuomotora, y se utilizó una variación de ondículas de causalidad-G para identificar influencias causales variables en el tiempo. La G-causalidad también se ha aplicado a sistemas neuronales simulados con el fin de investigar la relación entre la neuroanatomía, la dinámica de la red, y el comportamiento. Un desafío clave en la neurociencia y, en particular, la neuroimagen, es ir más allá de la identificación de activaciones regionales hacia la caracterización de circuitos funcionales que sustentan la percepción, la cognición, la conducta y la conciencia. El análisis de causalidad de Granger (G-causalidad) proporciona un poderoso método para lograrlo, mediante la identificación de interacciones dirigidas funcionales ("causales") a partir de datos de series de tiempo.

Un artículo reciente explica el procedimiento usando la técnica VAR:[11] En la práctica, la estrategia computacional para el análisis de la implementación de causalidad-G (GCA) se basa en la estimación y comparación de dos modelos VAR, dados un conjunto de datos de series de tiempo. Asumamos que tenemos 3 variables: X, Y, y Z, y estamos interesados en medir el flujo de información de X a Y. En primer lugar, un modelo VAR "pleno" se estima en forma conjunta para todas las variables. Esto conduce a un error de predicción / estimación particular, para cada variable dentro del conjunto. Un segundo modelo VAR "reducido" se estima entonces, que omite la causa potencial (X, en el ejemplo anterior). Esto conduce a un segundo conjunto de errores de predicción para cada variable restante. Si el error de predicción para Y es significativamente menor para la regresión completa (incluyendo X), en comparación con la regresión reducida (excluyendo X), entonces se dice que X G-causa Y, condicionado a Z. (Obsérvese que también tenemos, a partir de los mismos modelos, la G-causalidad de X a Z condicionado a Y). Técnicamente, la magnitud de la G-causalidad está dada por la relación de la varianza de los términos de errores de predicción para las regresiones reducidas y completas.


Christopher Albert "Chris" Sims (1942)
Chris page

La autorregresión vectorial (VAR) es un modelo econométrico utilizado para captar interdependencias lineales entre múltiples series de tiempo. Los modelos VAR generalizan el modelo autorregresivo univariado (modelo AR) al permitir más de una variable en evolución. Todas las variables en un VAR entran en el modelo de la misma manera: cada variable tiene una ecuación que explica su evolución en función de sus propios rezagos y los retrasos de las demás variables del modelo. La modelización VAR no requiere tanto conocimiento acerca de las fuerzas que influyen en una variable al igual que los modelos estructurales con ecuaciones simultáneas: el único conocimiento previo requerido es una lista de variables que puede ser considerado que se afectan entre sí intertemporalmente.

Christopher Sims abogó por el uso de modelos VAR, criticando las pretensiones y el rendimiento de modelos más

[11] Anil K. Seth, Adam B. Barrett, and Lionel Barnett, Granger Causality Analysis in Neuroscience and Neuroimaging, 2015.

tempranos en econometría macroeconómica. Recomendó los modelos VAR, que habían aparecido previamente en las estadísticas de series de tiempo y en la identificación de sistemas, una especialidad estadística en la teoría del control. Sims abogó porque los modelos VAR proporcionan un método "libre de teoría" para estimar relaciones económicas, siendo así una alternativa a las restricciones de identificación "increíbles" de los modelos estructurales.

Ejemplo 3 La integración del crecimiento y de una política pro-mercado fue una de las hipótesis centrales que este autor analizó formando parte del equipo de FIEL que, en 1999, diagnosticó el *síndrome de crecimiento* asociado a la política de convertibilidad entre 1992 y aquel año, con la colaboración de Arnold C. Harberger.[12] Uno de los instrumentos usados fue el test de Causalidad de Granger. Se acompañó el resultado del test de Granger entre variables que potencialmente podrían estar vinculadas: el índice de PTF (productividad total de los factores), el producto de la economía de negocios Y_r (equivalente al producto total excluidos el sector público no empresario, los servicios de la vivienda, y el sector agropecuario –excepto la pesca) y la inversión bruta fija IBF. Un primer test indicó que, al 95% de confianza, solamente se detectó causalidad unidireccional desde PTF hacia la inversión. En un segundo test, la PTF es causa del producto, en sentido de Granger, al 99% de confianza, no descartándose la causalidad bidireccional. Luego, puede afirmarse que entre 1992 y 1999 hubo una clara incidencia de la productividad factorial (la clásica TFP) sobre el producto de la economía, detectándose indicios de causalidad bidireccional (es decir, TFP $\rightleftarrows$ Y_r). Este resultado pondría en evidencia los beneficios derivados de la estabilidad macroeconómica, que permite que un sistema pueda funcionar de modo eficiente en un escenario descentralizado de política económica.

Como indica este documento de FIEL, es natural asociar el aumento promedio de la productividad de las firmas al mejor clima de negocios que prevaleció en la Argentina en buena parte de los 1990s. Varios son los caminos por los que las reformas de principios de esa década pueden haber alentado la productividad. En principio, *la estabilidad macroeconómica tiene un papel central*. Con incertidumbre macroeconómica es difícil que los empresarios se embarquen en proyectos de cambios tecnológicos u organizacionales, y si lo hacen, la probabilidad de éxito resulta menor en un contexto donde las genuinas señales económicas se encuentran borrosas. Más allá de la estabilidad macro, tímidamente hacia fines de los ochenta y con más fuerza en los 1990s se instrumentaron varias *reformas pro mercado: privatizaciones, apertura comercial y financiera, y desregulaciones*. En ese escenario con menor incertidumbre, señales económicas más claras, menor riesgo de expropiación por decisiones estatales y nuevos mercados para el negocio privado, es natural que los empresarios hayan evidenciado un mayor dinamismo y una búsqueda más intensa de fuentes de ahorro real de costos. Adicionalmente, *la mayor competencia generada en muchos mercados por la apertura y las desregulaciones* implicó la necesidad de obtener ganancias de productividad para sobrevivir en un contexto competitivo.

Este trabajo de FIEL fue seguido en 2011 por otro documento que actualizó este diagnóstico incluyendo la mayor parte del período kirchnerista, *Terms of Trade and Economic Growth in Argentina*, en el cual participé con Daniel Artana, Juan Luis Bour, Cynthia Moskovits y Nuria Susmel. “En este documento tenemos la intención de presentar una teoría de la evolución de la productividad total de los factores y condiciones de comercio a lo largo de la última década en Argentina, y some-

[12] FIEL, 2002, Productividad, Competitividad y Empresas – Los engranajes del crecimiento, capítulo 2.

terla a una prueba econométrica. Nuestra creencia a priori es que una parte significativa del crecimiento económico reciente se puede atribuir a un factor exógeno, es decir, mejores términos de intercambio, más favorables enfrentados por Argentina desde 2003." En este último proyecto se introdujeron algunas modificaciones con relación al trabajo anterior. El enfoque económico de los índices de precios se basa en el supuesto de que los agentes económicos (consumidores o productores) optimizan en condiciones competitivas. Vamos a incluir la totalidad de la economía - hay que subrayar que, en FIEL (2002), se consideró sólo al "sector empresarial", manteniendo cuentas separadas para el sector agrícola. Hubiera sido mejor centrarse en el sector empresarial, pero los datos disponibles no permitieron hacerlo. Por ejemplo, para la vivienda ocupada por el propietario, el producto es igual a los insumos y por lo tanto no hay mejoras de productividad que puedan ser generadas en este sector de acuerdo con las normas del Sistema de Cuentas Nacionales. Existen problemas similares para medir la productividad del gobierno.

"El crecimiento de los ingresos reales en el tiempo puede ser descompuesto en tres factores principales: el progreso técnico o *productividad total de los factores*, el *crecimiento de los precios de producción reales* y el *crecimiento de los insumos primarios (capital y trabajo)*. En este documento nos concentraremos en los primeros y últimos conductores, por la siguiente razón: es bien sabido que el Crecimiento de la Eficiencia y de la Tecnología son considerados como dos de las mayores sub-secciones de la Productividad Total de los Factores, teniendo el primero características inherentes "especiales" tales como las externalidades positivas y la no rivalidad, que mejoran su posición como motor del crecimiento económico. La Productividad Total de los Factores es a menudo vista como el verdadero motor de crecimiento dentro de la economía y los estudios revelan que, si bien el trabajo y la inversión son importantes contribuyentes, la Productividad Total de los Factores puede ser responsable de hasta el 60% del crecimiento en las economías. Durante el período de convertibilidad en la Argentina, la PTF creció 58% desde 1992 hasta 1998 y 113% acumulado en comparación con 1990, el año de productividad más baja de la década. Esto implicó ocho años con un crecimiento acumulado de PTF al 9,9% anual.

El cuadro básico

"Como puede verse en el gráfico adjunto Nº [15], el comportamiento del PIB a precios constantes experimentó desde 1980 fuertes fluctuaciones. Una simple regresión del logaritmo del PIB contra el tiempo, a partir de datos oficiales, permite obtener una tasa de crecimiento anual de alrededor del 2,2% en todo el período, pero es útil distinguir varios sub períodos:

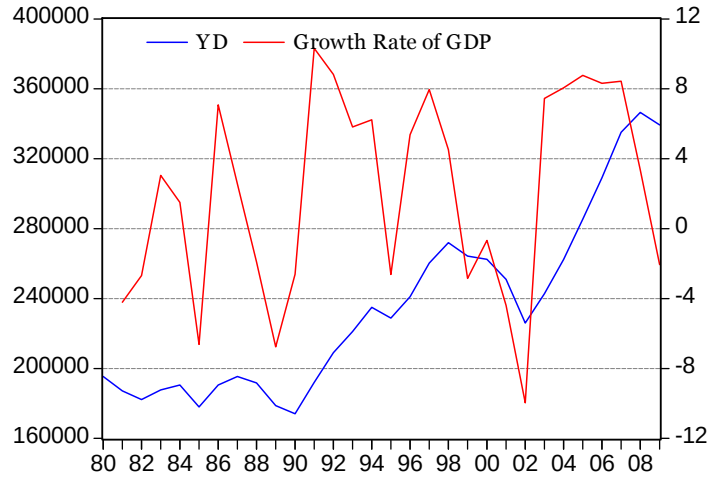


Figura 15 Evolución del PBI de Argentina

1) En el período 1980-1993, la economía creció a una tasa promedio de 0,6%; 2) En 1994-1998, el crecimiento fue de un 2,3% en un año; 3) En 1999-2002 hubo un retroceso a una tasa anual del -5,1%; 4) Entre 2003 y 2007 la tasa de crecimiento alcanzó 8,1% al año. FIEL obtuvo una nueva estimación

del PIB para 2008 y 2009, lo que implica un PIB inferior al oficial en esos años, por una cantidad relativa de -2,8% (2008) y -5,7% (2009). Estos datos se representan en el gráfico Nº 15, torciendo la expansión de la economía global a un nivel más bajo que los datos oficiales."

Usando un programa de estimación econométrica (EViews) se obtuvo la mejor aproximación teórica de este movimiento histórico (ecuación [11] del documento):

y'=0.45*(k(t-1)*u)'+0.42*(hrs*nt*(1-inf))' +0.09* (hrs*nt*inf)'+0.019-0.16*log(ri) -0.00002*crisk
(0.08) (0.04) (0.03) (0.004) (0.03) (0.000004)

(Un ' indica una operación Δlog)

R2=0.95; SE= 0.004; DW=1.84; MA (1)=1.00

Analicemos un poco esta ecuación.

En primer término, ¿cuáles son las variables dependientes e independientes?

La variable dependiente es y', a saber el *cambio logarítmico del PBI de un año al siguiente*. Es decir, podemos escribir $y = log(PBI_t) - log(PBI_{t-1})$. Por una propiedad del logaritmo,

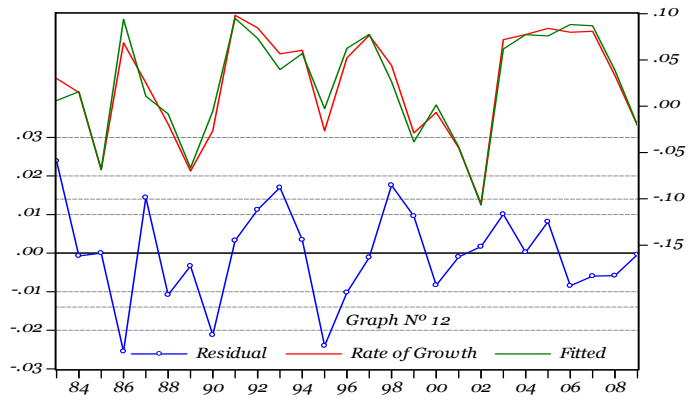


Figura 16 Residuos de la ecuación estimada

$log(PBI_t) - log(PBI_{t-1}) = log(PBI_t/PBI_{t-1})$.

Esto significa que y' está representando la operación de extraer el logaritmo al factor de expansión del PBI entre un año y el siguiente. Dado el valor que sea predicho por la ecuación, para recuperar este factor hay que sacar el *antilogaritmo*. Lo mismo hay que hacer con toda otra variable que aparece con una tilde en esta ecuación elegida.

Pasemos al lado derecho de la ecuación. Para analizarlo, primero es conveniente considerar que el PBI lo estaremos explicando por la combinación de dos elementos: 1) los factores productivos aplicados (capital y trabajo, formal e informal); y 2) la productividad media de ese complejo de factores. La cantidad de capital y de trabajadores está medida según estimaciones explicadas en el documento. La productividad media es explicada, a su vez, por tres variables: una tendencia temporal, la relación real de intercambio de Argentina y un coeficiente de riesgo-país.

El complejo de factores de capital y trabajadores viene medido por una función de producción Cobb-Douglas. La forma de esta función de producción es la siguiente:

$$PBI_t = A\ K_{t-1}\ L_t$$

Donde K_{t-1} es el acervo de capital productivo dejado a fines del año precedente (t-1), y L_t son los trabajadores empleados en el año t. El *capital* de cada año lo obtenemos sumando al capital dejado el año anterior la totalidad de las inversiones realizadas en el año corriente – a saber, construcciones, nuevas máquinas, pozos perforados, rutas construidas, y así sucesivamente – y restando el capital *amortizado* cada año en uso (este componente - no incluido en las estadísticas oficiales - es obtenido aplicando una tasa de amortización media al stock de cada período). En consecuencia, el stock de capital sólo puede aumentar si se registra *inversión neta positiva* (es decir, si la inversión bruta supera el monto de capital amortizado y/o depreciado). La evolución histórica del capital en Argentina puede apreciarse en el gráfico 2 del documento: allí está graficado el capital *efectivo* disponible, que se obtiene corrigiendo el stock de capital por un índice de uso de la capacidad (este índice es preparado por FIEL, pero también hay un índice oficial). Este índice viene representado con la letra *u*.

Pasemos ahora al factor *trabajo*. A éste lo hemos dividido en un segmento formal y otro informal. Se mide el número total de personas empleadas en ambos segmentos. A tal efecto fueron procesados en FIEL datos de la Encuesta de Hogares (INDEC) con los cuales se obtuvieron datos sobre informalidad. El gráfico [11] del documento muestra la elevada correlación que resulta entre el agregado de trabajadores formales así obtenido y el nivel del PIB. Para obtener el insumo total de trabajo fue utilizado el índice de horas trabajadas (un dato oficial complementado con estimaciones propias de FIEL). El dato de empleo es así uno de horas trabajadas (hrs) en la economía, diferenciado entre trabajadores formales e informales (para ambas categorías fue utilizada la misma cantidad de horas por empleado, ya que se carece de información para su apertura).

Comentarios sobre la ecuación estimada

En base a los análisis incluidos en el documento, se tomaron distintas decisiones de especificación del componente de productividad. Algunas tienen que ver con la forma de la función de producción y otras sobre las variables que inciden sobre la productividad:

1.- En primer lugar, el PIB total parece obedecer a una función de producción de rendimientos constantes a escala, como se subraya en FIEL (2002). En una especificación previa (ecuación [4]) la suma de las elasticidades de la producción de capital (0,37) y el trabajo (0.64) no resultó igual a uno, pero no fue significativamente diferente de la unidad. Un test sobre la restricción de que la suma = 1 es un estadístico F con 1 y 22 grados de libertad, con una probabilidad del 26%. Por lo tanto, debemos rechazar la diferencia de la unidad como no significativa. Las elasticidades capital y trabajo están dentro del rango de práctica internacional (capital 0.45, trabajo formal 0.42, informal 0.09). A partir de estas elasticidades podemos computar las productividades marginales. En efecto, recordando que la elasticidad-capital, por ejemplo, se puede escribir como

$$EY/EK = \partial Y/\partial K * K/Y,$$

observando que la elasticidad capital estimada es 0.45, podemos calcular el producto marginal de ambos factores en el punto de la media muestral, de lo cual surge que *el producto marginal de los trabajadores formales supera al de los informales por un factor del orden del 80%*.

2.- Los coeficientes hallados para la función de producción *han permanecido sin variación* con respecto a los hallados en el documento de 2002. Esto es significativo, dados los cambios que experimentó la Argentina a principios del siglo 21.

3.- Recuérdese que en el caso de los coeficientes de regresión es utilizado un test de Student. Todos los coeficientes – a excepción del coeficiente de los trabajadores informales – son significativos al 0.1%. El coeficiente de los trabajadores informales es significativo al 1.3%.

4.- La productividad tiende a crecer en forma autónoma a una tasa del 1.9% anual. *Ésta es una tasa inferior a la del período de convertibilidad (2.1%)* que es aún más reducida si a los trabajadores se los agrupa en una categoría común (1.6%). El autor estima que de haber sido incluidos datos de años posteriores, la diferencia habría sido aún mayor.

5.- La ecuación estimada involucra un efecto importante de la relación de intercambio. Una mejora de la relación en torno del 10% tiende a mejorar *la tasa de crecimiento* de la economía en 1.6%.

6.- Usando un esquema de contabilidad del crecimiento se logró extraer una serie del residuo del PIB. Hemos probado esta variable utilizando como factores explicativos la relación de los términos de intercambio, una tendencia lineal y el riesgo país. El principal resultado es la confirmación de la influencia de estos factores, así como la conveniencia de *modelar la PTF como un proceso de promedio móvil, con choques aleatorios que se propagan a los valores futuros de la serie temporal*.

7.- *El riesgo país, es decir, el precio que se debe pagar por encima de la tasa del Tesoro de EE.UU. para invertir en Argentina, es muy significativo y su coeficiente ha ido en aumento desde 2007* en adelante (ver Gráfico Nº 5 del documento). Por otra parte, el coeficiente de los términos de intercambio se ha mantenido estable desde 1995. Una prueba de causalidad de Granger con dos retardos sugiere que debemos rechazar el enunciado unilateral "Riesgo-país no es causa (en sentido de Granger) de la tasa de crecimiento del PBI" con una confianza del 99%.

8.- Hicimos un ejercicio *ex post* a fin de determinar el impacto neto de los términos del intercambio sobre el crecimiento de la economía argentina. Para ello, se simuló con el nombre de Cs el *Crecimiento simulado* de la economía argentina si no hubiera existido una modificación favorable de la relación de intercambio, permaneciendo ésta al mismo nivel al que llegó en 2003. La tabla siguiente muestra la tasa de crecimiento simulada y la tasa acumulativa de crecimiento simulada a partir de enero de 2004:

	(Cs) Tasa simulada	Factor acumulativo
2004	1.000244	1.000244
2005	1.035482	1.035735
2006	1.027812	1.064541
2007	1.049425	1.117157
2008	1.064715	1.189454
2009	1.042941	1.240530

Dado que el crecimiento total de la economía argentina en el mismo período fue a una velocidad de 48,2% según las estadísticas oficiales, se puede inferir que *la mitad del crecimiento de Argentina*

devengado en este período se explica en su totalidad por mejores términos de intercambio. Este cálculo probablemente sobre-estima la diferencia en cierta medida, ya que no tiene en cuenta el *efecto incentivo* que mejores términos del intercambio tiene sobre la decisión de inversión.