Ejercicio: Ventas, precio y gastos en publicidad

Alfonso Novales
Departamento de Economía Cuantitativa
Universidad Complutense*

10 de mayo de 2004

1.

*Versión muy preliminar. No citar sin permiso del autor. Las sucesivas versiones de este trabajo irán estando disponibles en http://www.ucm.es/info/ecocuan/anc, o pueden solicitarse en anovales@ccee.ucm.es

Ventas y gastos en publicidad[1]

Este ejemplo utiliza datos tomados de Novales, Estadística y Econometría. Son datos ficticios consistentes en 10 observaciones anuales sobre las ventas, gastos en publicidad y precio del producto de una empresa. El interés del ejemplo es:

- ilustrar el modo de interpretar los valores numéricos estimados para coeficientes individuales en un contexto de colinealidad,
- mostrar la manera de analizar el contenido informativo de las variables explicativas en un contexto de alta colinealidad,
- proponer un modo de tratar la colinealidad entre variables explicativas,

1.1. Algunas características de las variables

El archivo de trabajo contiene información acerca de la cifra de ventas anuales V_t de una empresa, junto con sus gastos en publicidad, Pub_t, ambos en miles de euros, y el precio de venta de su producto, P_t, asimismo en miles de euros por unidad. Son datos artificiales, formados por 10 observaciones de cada variable, pero serán suficiente para ilustrar las cuestiones que nos interesan. Las tres variables muestran, dentro del breve espacio de tiempo cubierto por la muestra, un comportamiento tendencial, que es creciente en el caso de las ventas y los gastos en publicidad, y decreciente en el caso del precio del producto.

Las nubes de puntos representan la relación entre la cifra de ventas anual y cada una de las dos potenciales variables explicativas, precio y gasto en publicidad $(NUBE\_VENT\_PRECIO, NUBE\_VENT\_PUB)$, mostrando claramente una asociación negativa entre V_t y P_t, y positiva entre V_t y Pub_t.

Las covarianzas y coeficientes de correlación entre las variables pueden resumirse en la matriz,

$$\Sigma = \begin{pmatrix} Correlaciones/Covarianzas & Ventas & Publicidad & \mathrm{Pr}\,ecio \\ Ventas & S_V^2 & S_{V,Pub} & S_{V,P} \\ Publicidad & \rho_{V,Pub} & S_{Pub}^2 & S_{Pub,P} \\ \mathrm{Pr}\,ecio & \rho_{V,P} & \rho_{Pub,P} & S_P^2 \end{pmatrix} =$$

[1]Fichero de trabajo: Ventas.wf1. Fichero de Excel: Ventas.xls.

$$= \begin{pmatrix} Corr./Cov. & Ventas & Publicidad & \text{Pr}ecio \\ Ventas & 443,5 & 124,1 & -99,0 \\ Publicidad & 0,950 & 38,5 & -26,8 \\ \text{Pr}ecio & -0,901 & -0,829 & 27,2 \end{pmatrix}$$

que muestra en su diagonal las varianzas de las tres variables; sus coeficientes de correlación consigo mismas son igual a uno, por lo que no es preciso mostrarlos. Debajo de la diagonal aparecen los coeficientes de correlación entre cada par de variables, todos ellos entre -1 y +1, mientras que por encima de la diagonal aparecen las covarianzas. Como puede verse, los tres coeficientes de correlación son muy elevados en valor absoluto.

Las desviaciones típicas muestrales de las variables son,

$$D.T.(V_t) = \sqrt{443,5} = 21,06;\ D.T.(Pub_t) = \sqrt{38,5} = 6,20;\ D.T.(P_t) = \sqrt{27,2} = 5,22.$$

Para obtener una medida comparable entre variables, debe utilizarse el coeficiente de variación, definido como cociente entre la desviación típica y media muestral de una variable.[2]

El modelo de ventas estimado utilizando tanto los gastos en publicidad como el precio el bien como variables explicativas es,

$$V_t = \underset{(67,3)}{247,6} + \underset{(0,545)}{2,204}\ Pub_t - \underset{(0,649)}{1,464}\ P_t, \tag{1.1}$$

con coeficiente de determinación y varianza residual,

$$\begin{aligned} R^2_{V.[Pub,P]} &= 1 - \frac{SR^2_{V.[Pub,P]}}{T.S^2_V} = 0,943 \\ \hat{\sigma}^2_u &= \frac{SR^2_{V.[Pub,P]}}{T-3} = \frac{250,6}{10-3} = 35,8 \ \Rightarrow \hat{\sigma}_u = \sqrt{35,8} = 5,98 \end{aligned}$$

El coeficiente de determinación es elevado, indicando que más de un 94% de las fluctuaciones en las cifras de ventas anuales está explicada por cambios en el precio del producto y en el gasto en publicidad. El ajuste parece bastante bueno.

[2]Que en todo caso, no tiene sentido en el caso de variables tendenciales, como ocurre en este ejemplo, por lo que no calculamos dichos coeficientes, renunciando a comparar entre sí el grado de volatilidad de las variables del modelo.

Otra manera de ver este hecho consiste en comparar la desviación típica de los residuos, 5,98, que indica el tamaño[3] del componente de las ventas no explicado por el modelo, y la desviación típica de las propias ventas, que es de 21,2. El término de error parece pequeño en relación con el tamaño medio de las fluctuaciones anuales en las cifras de ventas, y el indicador de ajuste asociado toma un valor numérico:

$$Ratio = 1 - \frac{\hat{\sigma}_u}{\sqrt{Var(V_t)}} = 1 - \frac{5,98}{21,1} = 0,717$$

indicando que el 72% del tamaño medio de las fluctuaciones anuales en ventas ha quedado explicada por el modelo anterior.[4]

1.2. Interpretación de los coeficientes estimados

Este ejemplo nos proporciona una ilustración de las dificultades que apuntamos en la Sección XX acerca de la interpretación de los coeficientes de regresión en un contexto de colinealidad.

Una lectura del modelo (1.1) sugeriría que las cifras de ventas aumentan en 220,4 euros por cada 100 euros de incremento en gastos de publicidad, suponiendo que el precio del producto no variase. Esta es la interpretación *ceteris paribus*, tan habitualmente utilizada, pero también tan poco consisitente con la mayoría de las situaciones a que se enfrenta un analista de datos económicos, con variables epxlicativas correlacionadas entre sí. De modo análogo, las ventas disminuirían en 146,4 euros por cada 100 euros de incremento en el precio unitario del articulo comercializado por la empresa.

Pero, al existir la simultaneidad mencionada entre los niveles de gastos en publicidad y de precios, la interpretación *ceteris paribus* no es rigurosa, puesto que, como indica el elevado coeficiente de correlación negativo entre ambas variables, de -0,829, indicando que la empresa gasta más en publicidad en períodos en que el precio del producto es bajo, y menos cuando el precio del producto es alto, lo cual podría ser espúrio o, por el contrario, fruto de una estrategia deliberada

[3]Una vez más, interpretamos la desviación típica de una variable aleatoria de esperanza matemática cero como inidicador de tamaño de dicha variable.

[4]Recordemos que $R^2 = 1 - \frac{SR^2}{ST}$, por lo que $SR^2 = (1-R^2)ST$, y $\hat{\sigma}_u^2 = \frac{SR^2}{T-k} = (1-R^2)\frac{ST}{T-k} = (1-R^2)\frac{T}{T-k}Var(y_t)$, por lo que, $Ratio = 1 - \frac{\hat{\sigma}_u}{\sqrt{Var(V_t)}} = 1 - \sqrt{(1-R^2)\frac{T}{T-k}}$. Para valores grandes de T en relación con k, $Ratio = 1 - \sqrt{(1-R^2)}$, siendo siempre $Ratio$ inferior a R^2.

de marketing. Con independencia de las razones que generan dicha correlación, un mayor gasto en publicidad suele venir asociado a una reducción en el precio del producto, siendo el efecto sobre las ventas la conjunción de ambos efectos. En consecuencia, al incrementar el gasto en publicidad en 100 euros, las ventas aumentarían en más de 220,4.

De hecho, las regresiones simples de las ventas anuales sobre cada variable explicativa son,

$$V_t = \underset{(0,375)}{96,0 + 3,224}\ Pub_t,\ R^2_{V.Pub} = 0,902,\ \hat{\sigma}_u = 7,36 \tag{1.2}$$

$$V_t = \underset{(0,619)}{483,6 - 3,637}\ P_t,\ R^2_{V.P} = 0,812,\ \hat{\sigma}_u = 10,22 \tag{1.3}$$

cuyos coeficientes de determinación son, por supuesto, el cuadrado de los coeficientes de correlación simples que aparecen en la matriz Σ. En ambos casos estimamos unos coeficientes mayores en valor absoluto a los obtenidos en el modelo de regresión múltiple, por las razones que acabamnos de exponer.

Veamos qué implicaciones tiene el coeficiente de correlación estimado, de -0,829. Puesto que,

$$\rho_{Pub,P} = E\left(\frac{Pub - E(Pub)}{\sigma_{Pub}}\frac{P - E(P)}{\sigma_P}\right)$$

tenemos que $\rho_{Pub,P}$ mide el valor medio que toma el producto de las fluctuaciones en Pub y P alrededor de sus respectivas medias. Supongamos que ambas variables están permanentemente en torno a sus valores medios, de los cuales se desvían cada período en una cuantía media igual a sus respectivas desviaciones típicas.[5] Esto significa que un incremento nominal de 6,20 euros en el gasto en publicidad, equivale a un aumento de una desviación típica en dicha variable. La expresión anterior, junto con $\rho_{Pub,P} = -0,829$, sugiere que dicho incremento venga asociado con un descenso de 0,829 desviaciones típicas en el precio.[6] Teniendo en cuenta que $\sigma_P = 5,22$, dicho descenso equivale a una reducción en el precio

[5] Estrictamente hablando, este supuesto es apropiado únicamente en situaciones en que las desviaciones respecto del valor medio en períodos sucesivos son independientes. Esto no ocurre en presencia de comportamientos tendenciales como los de las variables en este ejemplo.

[6] Esta interpretación es estrictamente válida si entendemos que el nivel de precios se fija por la empresa en respuesta al gasto en publicidad acometido, y no al revés; es decir, si interpretamos la alta correlación entre estas variables en el sentido $Publicidad \rightarrow \mathrm{Pr}\,ecio$.

del producto de 4,327 euros. La estimación del modelo de regresión múltiple sugiere que el mayor gasto en publicidad eleva las ventas en 13,665 euros, mientras que el descenso en el precio aumenta las ventas en 6,335 euros, siendo la suma de ambos efectos de 20,00 euros. Este es el efecto que sobre las ventas tiene un incremento de 6,20 euros en el gasto en publicidad, teniendo en cuenta la relación que existe a lo largo de la muestra entre esta variable y el precio unitario del producto. Si consideramos un incremento de 100 euros en el gasto en publicidad, como $100 = (6,20)(16,13)$, tendríamos un efecto estimado sobre las ventas de: $(20,00)(16,13) = 322,6$ euros en ventas, aproximadamente igual a lo obtenido al estimar el modelo (1.2).[7] Por tanto, los coeficientes estimados en las regresiones simples incorporan el efecto que simultáneamente se produce en la variable omitida cuando cambia el valor numérico de la variable incluida en la regresión simple.

1.2.1. Sesgo de variables omitidas

El razonamiento que hemos hecho en el párrafo anterior es, exactamente, la aplicación práctica de las expresiones sobre el sesgo que se produce en el estimador de mínimos cuadrados cuando se omiten del modelo variables relevantes. Consideremos un modelo de regresión múltiple con dos variables explicativas, x_i, x_e, cuyos subíndices denotan que una se incluye en la regresión simple, y otra queda excluida de dicha regresión. Es decir, la regresión múltiple es: $y = \beta_{im}x_i + \beta_{em}x_e + u$, mientras que la regresión simple es, $y = \beta_{is}x_i + v$. En el razonamiento previo hemos descompuesto el efecto β_{is} de una variación unitaria en la variable incluida en una regresión simple x_i, en dos componentes: el efecto directo, medido por el coeficiente de la variable que en el modelo de regresión múltiple tiene la variable incluida β_{im}, y el efecto indirecto. Para estimar éste último, hemos calculado la variación unitaria en términos de desviaciones típicas, $\frac{1}{\sigma_i}$. A continuación, hemos utilizado la definición de coeficiente de correlación para inferir que, en media, esta variación irá acompañada de una variación de $\frac{1}{\sigma_i}\rho_{ie}$ desviaciones típicas en la variable excluida, x_e. Esto equivale a una variación nominal de $\frac{1}{\sigma_i}\rho_{ie}\sigma_e$ en dicha variable. Utilzando las estimaciones de la regresión múltiple, su impacto sobre la variable dependiente será $\frac{1}{\sigma_i}\rho_{ie}\sigma_e\beta_{em}$. Pero esto es igual a $\frac{\sigma_{ei}}{\sigma_i^2}\beta_{em}$, y $\frac{\sigma_{ei}}{\sigma_i^2}$ no es

[7] El lector puede repetir el ejercicio partiendo de un descenso de una desviación típica en el precio del producto. Comprobará que el efecto global que obtiene sobre las ventas debido a un descenso de 100 euros en el precio del producto es el que estimaría a partir del modelo de regresión simple (1.3).

sino la estimación de mínimos cuadrados del coeficiente $\beta_{e/i}$de la regresión simple de la variable omitida x_e, sobre la incluida, x_i. En definitiva, el efecto global de una variación unitaria en la variable incluida en la regresión simple, x_i, sobre la variable dependiente, es,

$$\beta_{im} + \beta_{em}\beta_{e/i}$$

Esta es precisamente la expresión de la esperanza matemática del estimador de mínimos cuadrados del modelo de regresión múltiple $E(\beta_{is})$ que incluye a x_i como única variable explicativa. [**CHEQUEAR**: La cuestión es que si bien el estimador de mínimos cuadrados de β_{is} es un estimador sesgado de β_{im}, es, sin embargo, un estimador insesgado del efecto que una variación en x_i tiene sobre la variable dependiente, siendo, por el contrario, β_{im} un estimador sesgado de tal efecto.]

1.3. La capacidad explicativa de cada variable

Varias son las cuestiones que han de tenerse en cuenta al tratar de evaluar, en términos relativos, el contenido informativo que cada variable explicativa tiene sobre la variable dependiente. En primer lugar, podríamos utilizar el hecho de que el coeficiente estimado para Pub en (1.1) es mayor en valor absoluto que el estimado para P_t para decir que la primera variable es más relevante al explicar las ventas de la empresa. Esto sería incorrecto por dos razones: una de ellas ha sido explicada en la sección anterior, donde hemos visto que en un contexto de colinealidad, un coeficiente individual no puede interpretarse como el impacto que sobre la variable dependiente tiene una variación unitaria en la variable que acompaña a dicho coeficiente estimado. La segunda razón es que, en todo caso, los coeficientes individuales medirían el impacto que sobre las ventas tiene una variación unitaria, positiva o negativa, en cada una de las variables explicativas; el problema es que una variación de una unidad o de 100 unidades puede ser muy grande para una variable, y muy pequeña para otra. Ello dependerá de las variaciones medias que cada una de las variables experimenta a lo largo de la muestra, lo que nos lleva al siguiente epígrafe,

1.3.1. La volatilidad de la variable explicativa

Para afirmar que la publicidad es más importante que las ventas porque el valor absoluto del coeficiente estimado para la primera en (1.1) es mayor que el de la

segunda, deberíamos tener en cuenta el tamaño medio de la variación anual media en ambas variables, medido por sus respectivas desviaciones típicas.[8] El efecto promedio de una variable explicativa sobre la variable endógena se obtendría multilpicando el coeficiente estimado por la variación media enla variable explicativa. Utilizando los coeficientes estimados en el modelo de regresión múltiple (1.1), tendríamos,

$$\begin{aligned} Efecto(Pub &\rightarrow Ventas) = 2,204 * 6,20 = 13,665 \\ Efecto(\mathrm{Pr}ecio &\rightarrow Ventas) = -1,464 * 5,22 = -7,642 \end{aligned}$$

resultando superior en valor absoluto para los gastos en publicidad que para el precio del producto. Los gastos en publicidad tienen asociado un mayor coeficiente y, además, sus variaciones anuales son de mayor tamaño; ambos efectos se agregan, en este caso, para sugerir que esta variable es la más relevante para explicar las cifras de ventas. Sin embargo, esta impresión está sujeta a la limitación que impone la fuerte colinealidad entre ambas variables explicativas del modelo. [**CHEQUEAR**: Si utilizásemos los coeficientes estimados en las regresiones simples (1.3) , (1.2) para estimar el efecto que sobre las Ventas produce una variación de una desviación típica en cada una de las dos variables , explicativas, tendríamos un resultado similar, ya que en dichas regresiones, la estimación numérica del coeficiente incorpora la correlación existente entre la variable incluida y la variable excluida,

$$\begin{aligned} Efecto(Pub &\rightarrow Ventas) = 3,224 * 6,20 = 19,99 \\ Efecto(\mathrm{Pr}ecio &\rightarrow Ventas) = -3,637 * 5,22 = -18,98 \end{aligned}$$

que como se ve son, efectivamente, muy similares].

1.3.2. Comparación de residuos

También podemos examinar la matriz de correlaciones entre la variable V_t y los residuos procedentes de las dos regresiones simples anteriores. El criterio es que cuanto más se distinga el residuo de la variable dependiente, mayor es la capacidad

[8] De nuevo, esta interpretación es correcta únicamente si las variaciones anuales son independientes entre sí. Sólo en tal caso pueden interpretarse como fluctuaciones alrededor del valor promedio de la variable.

explicativa del modelo. En la tabla de correlaciones [$CORRELACIONES$] se aprecia que las ventas tienen menor correlación con el residuo de la regresión sobre publicidad (0,312) que con el residuo de la regresión sobre el precio del producto (0,434), lo que sugiere nuevamente el mayor contenido informativo de los gastos en publicidad para explicar las Ventas. En todo caso, la correlación con los residuos de la regresión que utiliza ambas variables como explicativas es sensiblemente inferior (0,238), sugiriendo que ambas variables tienen contenido informativo no trivial.

Es asimismo importante comparar los residuos del modelo de regresión múltiple con los que se derivan de la estimación de cada uno de los modelos de regresión simple. La misma tabla de correlaciones en el fichero de trabajo nos muestra que si excluimos los gastos en publicidad del modelo de regresión múltiple, el coeficiente de correlación entre ambos conjuntos de residuos es de 0,548, mientras que si excluimos la variable precio de la regresión múltiple, el coeficiente de correlación entre residuos es de 0,761. Esto significa que la exclusión de los gastos en publicidad altera más los residuos del modelo que la omisión de la variable Precio, lo que sugiere que la primera es la variable más relevante, igual conclusión a la que alcanzamos en el párrafo anterior.

1.3.3. Comparación de estadísticos tipo t de Student

Es muy habitual en el trabajo empírico en Economía juzgar la capacidad explicativa de cada variable en términos relativos, de acuerdo con los valores numéricos de sus respectivos estadísticos tipo-t. Ello parece deberse a que la comparación entre el valor absoluto de dicho estadístico y el umbral crítico de 2,0 es el procedimiento habitual para analizar si una variable contiene capacidad explicativa sobre la variable endógena. Es decir, la significación estadística de una variable se interpreta directamente como su capacidad de explicar los valores numéricos de la variable dependiente. De dicho procedimiento parece inferirse que cuanto mayor sea el valor absoluto del estadístico tipo-t, mayor es la capacidad explicativa de la variable en cuestión.

Este procedimiento es generalmente inapropiado porque el economista está interesado en el impacto cuantitativo que cambios en una variable explicativa implican sobre la variable dependiente, y la significación estadística de una variable explicativa puede ser simultánea con un efecto cuantitativo muy reducido de dicha variable sobre la variable dependiente. Esta confusión entre significación estadística y relevancia cuantitativa ha sido muy dañina en la ionterpretación de

las estimaciones de modelos de regresión en Economía.

Hay diversas razones por las que una variable explicativa que tiene un notable efecto cuantitativo sobre la variable dependiente puede resultar estadísticamente no significativa. Dada la estructura del estadístico tipo-t, éste conjuga la estimación del impacto numérico de cambios en la variable explicativa, con la precisión con que dicho impacto se estima. Así, la ausencia de significación estadística puede surgir bien porque el impacto numérico de dicha variable es muy reducido, o porque, siendo importante, no se mide con suficiente precisión, es decir, con una varianza suficientemente pequeña. De este modo, no puede interpretarse un valor pequeño del estadístico tipo-t como evidencia de una reducida capacidad explicativa; en particular, un valor de dicho estadístico inferior a 2 no necesariamente significa que la variable explicativa en cuestión no tenga contenido informativo sobre la variable dependiente. Ignorar el papel que la precisión en la estimación de un coeficiente del modelo de regresión tiene sobre los contrastes de significación estadística es la segunda fuente tradicional de error en la interpretación de las estimacioens de modelos de relación entre variables económicas.

1.4. ¿Cuánta información aporta cada variable explicativa?

1.4.1. Coeficientes de determinación y de correlación parcial

En este ejemplo podemos definir el coeficiente de correlación parcial entre Ventas y Publicidad, $\rho_{V,Pub.P}$ como el coeficiente de correlación simple entre las variables transformadas que resultan al extraer de ambas variables el efecto común que sobre ellas tiene el Precio del producto. Análogamente, podemos definir el coeficiente de correlación parcial entre Ventas y Precio, $\rho_{V,P.Pub}$ como el coeficiente de correlación simple entre las variables transformadas que resultan al extraer de ambas variables el efecto común que sobre ellas tiene el gasto en Publicidad.

En función de los coeficientes de correlación simples habituales, demostramos en la Sección XX las expresiones,

$$\rho_{V,Pub.P} = \frac{\rho_{V,Pub} - \rho_{V,P}\rho_{Pub,P}}{\sqrt{\left(1-\rho_{V,P}^2\right)\left(1-\rho_{Pub,P}^2\right)}} = \frac{0,950-(-0,901)(-0,829)}{\sqrt{1-(-0,901)^2}\sqrt{1-(-0,829)^2}} = 0,837$$

$$\rho_{V,P.Pub} = \frac{\rho_{V,P} - \rho_{V,Pub}\rho_{Pub,P}}{\sqrt{\left(1-\rho_{V,Pub}^2\right)\left(1-\rho_{Pub,P}^2\right)}} = \frac{-0,901-0,950(-0,829)}{\sqrt{1-0,950^2}\sqrt{1-(-0,829)^2}} = -0,650$$

que nos proporcionan los valores numéricos de ambnos coeficientes de correlación simples.

Vimos asimismo en el texto expresiones para el cálculo de los coeficientes de determinación y correlación parcial en función de Sumas Residuales,

$$\begin{aligned} R^2_{V/Pub.P} &= 1 - \frac{SR^2_{V./Pub,P]}}{SR^2_{V/P}} \Rightarrow \rho_{V,Pub.P} = \sqrt{R^2_{V/Pub.P}} \\ R^2_{V/P.Pub} &= 1 - \frac{SR^2_{V/[Pub,P]}}{SR^2_{V/Pub}} \Rightarrow \rho_{V,P.Pub} = \sqrt{R^2_{V/P.Pub}} \end{aligned}$$

Teniendo en cuenta que la relación entre la Suma Residual de un modelo de regresión que tiene como variable dependiente a y y el coeficiente de determinación de la misma es: $SR^2 = T.S^2_y(1 - R^2)$, tenemos,[9]

$$\begin{aligned} SR^2_{V/[Pub,P]} &= T.S^2_V(1 - R^2_{V/[Pub,P]}) = 10(443,5)\,(1 - 0,943) = 252,80 \\ SR^2_{V/Pub} &= T.S^2_V(1 - R^2_{V/Pub}) = 10(443,5)\,(1 - 0,902) = 434,63 \\ SR^2_{V/P} &= T.S^2_V(1 - R^2_{V/P}) = 10(443,5)\,(1 - 0,812) = 833,78 \end{aligned}$$

Utilizando estos valores numéricos, tenemos,

$$\begin{aligned} R^2_{V/Pub.P} &= 1 - \frac{SR^2_{V/[Pub,P]}}{SR^2_{V/P}} = 1 - \frac{252,80}{833,78} = 0,697 \Rightarrow \rho_{V,Pub.P} = \sqrt{0.697} = 0,835 \\ R^2_{V/P.Pub} &= 1 - \frac{SR^2_{V/[Pub,P]}}{SR^2_{V/Pub}} = 1 - \frac{252,80}{434,63} = 0,418 \Rightarrow \rho_{V,P.Pub} = \sqrt{0.418} = -0,647 \end{aligned}$$

donde hemos asignado un signo negativo a $\rho_{V,P.Pub}$ debido a ser una correlación entre ventas y nivel de precios. La ligera variación observada en los resultados proporcionados por ambos enfoques se debe a la aproximación numérica en las diferentes operaciones realizadas en cada caso.

[9] $SR^2_{V/[Pub,P]}$ denota la Suma Residual que resulta cuando Publicidad y Precio explican Ventas, mientras $SR^2_{V/Pub}$ denota la suma de cuadrados de residuos en una regresión de Ventas sobre gastos en publicidad. $SR^2_{V/Pub.P}$ sería la Suma Residual que se tendría en una regresión de ventas sobre gastos en publicidad, después de excluir de ambas variables el componente común que tienen por incorporar información sobre el precio. $SR^2_{V/P.Pub}$ se interpreta análogamente.

El coeficiente de correlación simple entre ventas y gastos en publicidad es de 0,950, reduciéndose a 0,835 si excluimos la simultaneidad que ambas variables muestran debido a su correlación común con el nivel de precios. El coeficiente de correlación simple entre ventas y nivel de precios es de -0,901, reduciéndose en valor absoluto al caer el coeficiente a -0,647 cuando excluimos la información común que sobre los valores anuales de ambas variables tiene el gasto en publicidad. Por tanto, el gasto en publicidad explica casi un 30% de la correlación entre venta y nivel de precios, mientras que el nivel de precios es responsable de sólo un 12% de la correlación entre ventas y gastos en publicidad. De ello concluimos que el gasto en publicidad es la variable más importante para explicar la evolución temporal de la cifra de ventas.

Tal conclusión coincide con la que alcanzamos al examinar las correlaciones entre Ventas y los residuos de las regresiones simples, así como al comparar los residuos de la regresión múltiple con los que se obtienen en cada una de las regresiones simples. Estos son los procedimientos que sugerimos utilizar para el análisis de este tipo de cuestiones. En este ejemplo se alcanzaría la misma conclusión por los procedimientos habituales de comparar el valor numérico de los coeficientes individuales estimados en el modelo de regresión múltiple, o el valor absoluto de los estadísticos tipo-t asociados a ambos coeficientes. Sin embargo, ya hemos indicado como ninguna de tales comparaciones está justificada y la coincidencia es casual. Veremos en otros ejemplos que los resultados no son siempre coincidentes[10].

Debe recordarse, de la discusión teórica en la Sección XX, que 0,835 es asimismo el coeficiente de correlación que obtendríamos entre los residuos de regresiones que explican las ventas y los gastos en publicidad, respectivamente, por la variable precio ($RES\_V\_PREC, RES\_PUB\_PRECIO$). De modo similar, -0,647 es el coeficiente de correlación entre los residuos de regresiones que explican las ventas y el precio, respectivamente, utilizando los gastos en publicidad como única variable explicativa (RES_V_PUB, RES_PREC_PUB). Ambos resultados pueden comprobarse utilizando las variables descritas, que se contienen en el fichero de trabajo.

[10]W. Kruskal, The American Statisticiam (1987), propuso utilizar el promedio de los cuadrados de los coeficientes de correlación simple y parcial entre Y y cada variable explicativa para evaluar la proporción de la fluctuación en Y que es explicada por cada una de éstas. En nuestro ejemplo tendríamos para los gastos en publicidad: $\frac{0.950^2+0.837^2}{2} = 0,801$, y para el nivel de precios: $\frac{(-0.901)^2+(-0.650)^2}{2} = 0,617$ alcanzando la misma conclusión [Johnston y DiNardo pág.80].

Esto nos recuerda el significado de los coeficientes de correlación parcial: al estimar las regresiones de ventas y precios sobre publicidad, estamos extrayendo de estas variables la información común con los gastos en publicidad, y luego correlacionamos los componentes así medidos, obteniendo el grado de asociación entre ventas y precio, excluyendo aquella correlación que pueda estar debida al hecho de que ambas se relacionan con el gasto en publicidad.

¿Qué explica cada variable? En sus estudios sobre los ciclos económicos, Tinbergen (1939) propuso un interesante método para reflejar la información contenida en cada variable explicativa a lo largo de la muestra.[11] Trabajando en desviaciones respecto de la media, Tinbergen sugería mostrar un gráfico representando simultáneamente los valores observados de y y los ajustados por el modelo, un gráfico para cada producto $\hat{\beta}_i x_i$, y un gráfico de residuos. Para ello se utilizan los coeficientes estimados en el modelo de regresión llineal múltiple. Hemos optado por presentar en Ventas_niveles.doc:[12] un gráfico de los valores anuales observados de y, junto con los valores anuales ajustados por el modelo; dos gráficos que confrontan los valores anuales observados para y con los valores explicados por cada una de las variables explicativas por separado, y un último gráfico que representa los valores observados de y frente a los residuos del modelo.

1.5. Colinealidad entre las variables explicativas: publicidad y precio

1.5.1. Regresiones simples cruzadas

Un investigador podría estimar asimismo dos modelos que tratan de recoger la correlación existente entre las variables explicativas,

$$Pub_t = \underset{()}{107,11} \underset{(0,235)}{-0,986}\ P_t,\ R^2_{Pub.P} = 0,687,\ \hat{\sigma}_u = 3,88 \tag{1.4}$$

$$P_t = \underset{()}{103,52} \underset{(0,166)}{-0,697}\ Pub_t,\ R^2_{P.Pub} = 0,687,\ \hat{\sigma}_u = 3,26 \tag{1.5}$$

en las que:

[11] Este procedimiento, olvidado por mucho tiempo, ha sido recordado por Johnston y DiNardo (1997).

[12] Recordamos nuevamente que utilizar desviaciones respecto de la media muestral en presencia de tendencias temporales puede no tener mucho significado. Puede conducir, además, a conclusiones erróneas.

- el investigador detecta fuerte correlación entre ambas variables, con un coeficiente de determinación que es igual, por supuesto, al cuadrado del coeficiente de correlación simple entre ambas variables. Por eso el coeficiente de determinación es el mismo en ambas regresiones, puesto que el coeficiente de correlación no encierra ninguna idea de causalidad y es, independiente, por tanto, de qué variable tomemos como dependiente y cuál como independiente.

- sin embargo, sería imposible concluir, utilizando estas regresiones estimadas, si la correlación entre precios y gastos en publicidad es fruto o no de una política explícita de comercialización.

- a pesar de la coincidencia entre coeficientes de determinación, las desviaciones típicas del término de error no son iguales, sin embargo, ya que las regresiones explican variables dependientes diferentes. Sin embargo, nuestros ratios habituales coinciden,

$$Ratio\,(Pub_t/P_t) = \frac{3.88}{6.20} = 0,626;\ Ratio\,(P_t/Pub_t) = \frac{3.26}{5.22} = 0,625$$

donde la leve diferencia se debe exclusivamente a los redondeos a tres decimales.

- las pendientes estimadas en ambas regresiones no son iguales. Su producto es: $(-0,986)(-0,697) = 0,687$ que es, precisamente, el coeficiente de determinación entre ambas variables, gastos en publicidad y nivel de precios. Esta es una propiedad de la regresión lineal simple: si se estiman por mínimos cuadrados regresiones de Y sobre X y de X sobre Y, el producto de las pendientes resultantes es siempre igual al cuadrado del coeficiente de correlación lineal simple entre ambas variables.

1.5.2. Tratamiento de la colinealidad

La regresión auxiliar entre nivel de precios y gastos en publicidad (1.5), nos permite estimar el componente de la evolución temporal del nivel de precios que no está explicado por las fluctuaciones que anualmente experimenta el gasto en publicidad,

$$P\backslash Pub_t = P_t - 103,52 + 0,697\ Pub_t \tag{1.6}$$

que no es sino el residuo de la regresión (1.5). Las propiedades del estimador de mínimos cuadrados garantizan que dicho residuo tiene correlación nula con los gastos en publicidad, por ser ésta la variable explicativa en la regresión a partir de la cual se han generado los residuos. Por tanto, $Corr(Pub_t, P\backslash Pub_t) = 0$.

Si ahora estimamos una regresión que pretende explicar las ventas mediante los gastos en publicidad y el componente de precios no explicado por estos, tenemos,

$$V_t = \underset{(5,26)}{95,99} + \underset{(0,305)}{3,224}\ Pub_t - \underset{(0,649)}{1,464}\ P\backslash Pub_t,\ R^2 = 0,943,\ \bar{R} = 0,927,\ \hat{\sigma}_u = 5,983 \tag{1.7}$$

donde puede observarse que el coeficiente estimado para $P\backslash Pub_t$ es el mismo que obtuvimos en la regresión inicial (1.1), y se estima con la misma precisión.

Sin embargo, el coeficiente estimado para los gastos en publicidad es ahora mayor que en (1.1); la razón es que en (1.1), al impacto directo sobre las ventas de un aumento en los gastos en publicidad había que añadir el impacto de la reducción en precios que usualmente acompaña el mayor gasto en publicidad. El efecto global es superior al medido por el coeficiente $2,204$ que los gastos en publicidad reciben en (1.1), y eso aparece claro en (1.7). De hecho, el coeficiente estimado para Pub_t en (1.7) es el mismo que obtuvimos en la regresión simple con esta variable, sólo que ahora lo estimamos con una mayor precisión.[13] Aunque numéricamente es mayor, también se estima dicho coeficiente con mayor precisión (menor desviación típica) en (1.7) que en la regresión inicial (1.1), gracias a que la ausencia de correlación entre las variables explicativas en (1.7) permite discriminar mejor el efecto de cada variable.

Podría pensarse que una limitación del modelo (1.7) es el hecho de que en él no aparece el precio del producto, sino tan sólo el componente del mismo que no está explicado por los gastos en publicidad. Esto es, en cierta forma, sólo aparente, pues si combinamos (1.7) con (1.6) se recuperan exactamente los mismos coeficientes estimados en la regresión original (1.1), excepto por el hecho de que el coeficiente de los gastos en publicidad se ha estimado con una precisión superior.

Alternativamente, podríamos utilizar (1.4) para estimar el componente de los gastos en publicidad no explicado por el nivel de precios,

$$Pub\backslash P_t = Pub_t - 107,11 + 0,986\ P_t,$$

[13] Que ambas estimaciones numéricas coincidan no es sino reflejo del resultado teórico que afirma que las estimaciones numéricas del coeficiente de mínimos cuadrados de una variable explicativa no cambia si se excluyen o se incluyen en el modelo variables explicativas incorrelacionados con la primera.

con $Corr(Pub\backslash P_t, P_t) = 0$. A continuación, estimaríamos la regresión de ventas sobre el nivel de precios y $Pub\backslash P_t$, obteniendo,

$$V_t = \underset{(5,26)}{95,99} + \underset{(0,545)}{2,204}\ Pub\backslash P_t - \underset{(0,363)}{3,637}\ P_t,\ R^2 = 0,943,\ \bar{R} = 0,927,\ \hat{\sigma}_u = 5,983 \tag{1.8}$$

siendo ahora el coeficiente de la variable auxiliar $Pub\backslash P_t$ el que coincide con el obtenido en el modelo original (1.1) para Pub_t, mientras que el coeficiente estimado para el nivel de precios es ahora mayor en valor absoluto que en (1.1), por las mismas razones antes descritas. Coincide con el obtenido en la regresión simple (1.3), aunque se estima con mayor precisión que en dicha regresión, y también con mayor precisión que en la regresión inicial. El investigador debería quedarse con una de las dos regresiones (1.7) o (1.8), dependiendo de la dirección de causalidad en la que interprete la correlación existente entre nivel de precios y gasto en publicidad.

La incorporación del componente del precio no relacionado con los gastos en publicidad eleva el coeficiente de determinación de la regresión de ventas sobre gastos en publicidad desde 0,902 a 0,943. De modo similar, la inclusión del componente del gasto en publicidad no relacionado con el precio en la regresión de ventas sobre precios, eleva el coeficiente de determinación de 0,812 al mismo nivel citado, 0,943. Esto sugiere que el contenido informativo de los gastos en publicidad sobre las ventas es mayor que el que tiene la variable Precio. Sin embargo, la comparación de coeficientes de determinación reduce toda la información muestral relativa a la explicación de las cifras de ventas a una sóla cifra. Preferimos comparar los residuos de modelos que incluyen o excluyen una variable explicativa, pues nos permiten analizar el impacto que dicha variable tiene, observación a observación. Es perfectamente imaginable que tal efecto sea muy notable pero esté concentrado en unas pocas observaciones que tengan alguna característica en común.[14] Ello haría que la comparación de medidas agregadas, como los coeficientes de determinación, no detectase la contribución de la variable explicativa. Si, por el contrario, una comparación detallada de los dos conjuntos de residuos nos detecta variaciones importantes en los residuos correspondientes a ese reducido

[14]Por ejemplo, cinco años consecutivos durante los que se produjo una gran elevación en los precios del petróleo, en una muestra de 60 años. En una muestra de sección cruzada correspondiente a un amplio conjunto de países, podrían ser los residuos correspondientes a los países subsaharianos los que experimentan una variación muy notable al incluir una determinada variable epxlicativa en el modelo.

conjunto de observaciones, podríamos definir una variable ficticia apropiadamente, mejorando con ello la capacidad explicativa del modelo.

2. Ejercicios

1. Compruebe que las expresiones que relacionan los coeficientes de las regresiones simples y múltiple,

$$\hat{\beta}_{Pub} = \frac{\hat{\beta}_{V/Pub} - \hat{\beta}_{V/P}\hat{\beta}_{P/Pub}}{1 - \hat{\beta}_{Pub/P}\hat{\beta}_{P/Pub}};\ \hat{\beta}_{P} = \frac{\hat{\beta}_{V/P} - \hat{\beta}_{V/Pub}\hat{\beta}_{Pub/P}}{1 - \hat{\beta}_{Pub/P}\hat{\beta}_{P/Pub}}$$

donde $\hat{\beta}_{Pub}, \hat{\beta}_{P}$ denotan las estimaciones de mínimos cudrados de los coeficientes de las variables Pub, P en el modelo de regresión múltiple, siendo el resto estimaciones de modelos de regresión simple, se satisfacen en este ejemplo.

1. Compruebe asimismo que las expresiones que relacionan los coeficientes de la regresion múltiple con los coeficientes de correlación simple,

$$\hat{\beta}_{Pub} = \frac{\rho_{V,Pub} - \rho_{V,P}\rho_{P,Pub}}{1 - \rho^2_{Pub,P}}\frac{S_V}{S_{Pub}};\ \hat{\beta}_{P} = \frac{\rho_{V,P} - \rho_{V,Pub}\rho_{Pub,P}}{1 - \rho^2_{Pub,P}}\frac{S_V}{S_P}$$

donde $\hat{\beta}_{Pub}, \hat{\beta}_{P}$ denotan las estimaciones de mínimos cudrados de los coeficientes de las variables Pub, P en el modelo de regresión múltiple, siendo el resto coeficientes de correlación simple, y S_V, S_{Pub}, S_P desviaciones típicas muestrales, también se satisfacen en este ejemplo.

1. Partiendo de un descenso de una desviación típica en el precio del producto, y siguiendo un argumento análogo al utilizado en la Sección XX, compruebe que el efecto global que obtiene sobre las ventas debido a un descenso de 100 euros en el precio del producto es el mismo que estimaría a partir del modelo de regresión simple (1.3) .

2. Al tratar en este ejemplo con datos temporales, el investigador podría optar por presentar en Ventas_variaciones.doc varios gráficos relativos a variaciones anuales:[15] un gráfico de variaciones anuales de y, junto con las variaciones ajustadas por el modelo; dos gráficos que confrontan las varaiciones

[15] En vez de utilizar desviaciones respecto de la media muestral, que en presencia de tendencias temporales no tiene mucho significado

anuales observadas en y con las explicadas por cada una de las variables explicativas por separado, y un último gráfico que representa las variaciones observadas frente a los residuos del modelo. Discuta si la imagen que tiene acerca del contenido informativo de ambas variables explicativas, en términos relativos, es el mismo que el que obtuvo de los gráficos de Tinbergen en niveles que se mostraron en el desarrollo del ejemplo.