

Franklin M. Fisher

Traducción: Enrique A. Bour, 2018

El análisis de regresión múltiple es un dispositivo para hacer estimaciones precisas y cuantitativas de los efectos de diferentes factores sobre alguna variable de interés. No es una herramienta nueva, ya que se remonta en sus orígenes a Carl Friedrich Gauss, un matemático muy importante nacido hace 240 años. Sin embargo, el uso práctico de la regresión múltiple ha crecido de manera muy sustancial en los últimos sesenta años más o menos. Este crecimiento se debe en parte al desarrollo de métodos estadísticos modernos, en parte a una mayor disponibilidad de datos estadísticos decentes, y quizás más que nada al desarrollo de la computadora electrónica. Parte del uso creciente de la regresión múltiple y las técnicas relacionadas se produjo en relación con procedimientos legales de diversa índole, aunque los abogados y los jueces a menudo se han inclinado a considerar dicho uso con desconfianza general (y en ocasiones saludable).



Johann Carl Friedrich Gauß
1777-1855

A la luz de la importancia cada vez mayor del análisis de regresión múltiple, es importante que los abogados comprendan qué es, cómo funciona y para qué se puede utilizar adecuadamente. Quizás el uso legal más importante de la regresión múltiple hasta ahora ha sido el análisis de los efectos disuasorios de la pena de muerte sobre el asesinato, citado por el Procurador General en su escrito *amicus curiae* ante la Corte Suprema en los casos de pena de muerte.¹ El hecho de que los estudios en los que se basó el Procurador General fueran fatalmente defectuosos,² en mi opinión, sólo aumentan la importancia de comprender la metodología involucrada. A un nivel menos importante, los

@Fuente: F.M. Fisher - *Multiple Regression in Legal Proceedings*, 80 Columbia Law Review 702 (1980). Fisher es Profesor de Economía en el Instituto de Tecnología de Massachusetts, Bachelor of Arts 1956, Master of Arts 1957, Ph. D. 1960, Harvard University. Este artículo fue adaptado de un documento entregado a la Asociación del Colegio de Abogados de la Ciudad de Nueva York (Comité Especial sobre Datos Empíricos en la Toma de Decisiones Legales) en mayo de 1979. Estoy en deuda con Michael O. Finkelstein por sus útiles críticas pero mantengo la habitual responsabilidad por los posibles errores.

¹ Fowler v. North Carolina, 428 U.S. 904 (1976); Woodson v. Carolina del Norte, 428 U.S. 280 (1976); Jurek v. Texas, 428 U.S. 262 (1976); Proffitt v. Florida, 428 U.S. 242(1976); Gregg v. Georgia, 428 U.S. 153 (1976). Ver en general *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates* (A. Blumstein, J. Cohen & D. Nagin eds. 1978) [en lo sucesivo citado como *Deterrence and Incapacitation*], que contiene, entre otras cosas, algunas discusiones devastadoras sobre los estudios involucrados. (En particular, ver el artículo de Klein, Forst & Filatov, *The Deterrent Effect of Capital Punishment: An Assessment of the Estimates*, p. 336).

² Ver texto que acompaña a las notas 38-46 infra.

estudios de regresión múltiple han figurado en varios otros procedimientos legales, y mientras que aquellos con los que estoy más familiarizado han sido los procedimientos regulatorios, no hay ninguna razón por la cual la regresión múltiple no deba usarse también en otros litigios.³

Este artículo primero explica, a nivel básico, el concepto de análisis de regresión múltiple, sus propiedades básicas y los supuestos fundamentales sobre los cuales descansa su validez.⁴ También discutiré los métodos para medir la exactitud y confiabilidad de las estimaciones generadas mediante regresión múltiple. La segunda parte del artículo explora con mayor profundidad el uso adecuado de regresión múltiple en procedimientos legales centrándose en tres áreas en las que los estudios de regresión múltiple podrían desempeñar un papel: el examen de la discriminación salarial, la determinación de daños antimonopolio y la evaluación del castigo como disuasorio del crimen.

I. ANÁLISIS DE REGRESIÓN MÚLTIPLE

A. *Usos de la Regresión Múltiple*

Los dos usos principales del análisis de regresión múltiple se ilustran mejor mediante un examen de situaciones reales en las que se emplearon estudios de regresión múltiple. Tomemos los siguientes dos casos:

1. Durante muchos años después de la desaparición de las locomotoras de vapor, hubo una perenne disputa laboral vinculada a preservar el empleo de los bomberos de los ferrocarriles. Independientemente de los méritos de esa disputa (que finalmente se resolvió, creo, a través de la negociación), uno de los temas en cuestión se refería a la cuestión de si la presencia de un bombero en los trenes contribuía a la seguridad del ferrocarril. Un estudio de ese tema, usando regresión múltiple, se presentó en testimonio ante un consejo de emergencia presidencial en 1970.⁵

2. Los sistemas de televisión por cable (CATV) han sido objeto de repetidos procedimientos de reglamentación por parte de la Federal Communications Commission. Entre los temas que intervienen en tales procedimientos se encuentra el efecto de la entrada y la actividad de los CATV sobre los beneficios y el crecimiento de las estaciones de televisión abierta. Este problema involucra cuestiones tales como la influencia de CATV sobre la audiencia a la que llegan determinadas estaciones de transmisión y el efecto de los cambios en la audiencia de una estación sobre los ingresos que recibe.⁶ En general, como cabría esperar, los operadores de cable han pretendido que tales efectos son pequeños y las esta-

³ Para una excelente discusión de procesos que usan estudios de regresión múltiple, ver Finkelstein, *Regression Models in Administrative Proceedings*, 86 Harv. L. Rev. 1442 (1973).

⁴ He sido responsable de varios estudios de regresión múltiple utilizados en procedimientos legales y, como los conozco mejor, de esos estudios extraeré ejemplos para gran parte de este artículo, con la esperanza de agregar más interés a lo que de otra forma podría degenerar en una discusión bastante seca y técnica.

⁵ El estudio puede ser consultado en Fisher & Kraft, *The Effect of the Removal of the Firemen on Railroad Accidents 1962-1967*, 2 Bell J. Econ. & Management Sci. 470 (1971).

⁶ Estas son preguntas importantes para la FCC ya que tienen que ver con el grado en que la regulación de la televisión por cable es necesaria para fomentar el crecimiento de nuevas estaciones de frecuencia modulada o para mantener ganancias que subsidian al servicio público y a otra programación de estaciones de transmisión locales.

ciones de transmisión han insistido que son grandes. El problema ha sido estudiado repetidamente por métodos de regresión múltiple, más recientemente en un estudio de la relación entre el tamaño de la audiencia e ingresos, escrito en parte por mí y presentado a la FCC en 1978-79.⁷

En el primer caso, el problema es si una variable particular (presencia o ausencia de bombros) tiene algún efecto sobre alguna otra variable (seguridad ferroviaria). En el segundo caso (la relación audiencia-ingreso), nadie duda que el tamaño de la audiencia afecte los ingresos de los canales, o sea la atención del espectador, que es lo que los canales venden a los anunciantes, y todas las partes están vitalmente interesadas en las estadísticas de audiencia; el problema es más bien el de medir el efecto. Estos dos usos de la regresión múltiple son lo que los estadísticos llaman **test de hipótesis** por un lado y **estimación de parámetros** por el otro. En el primer caso, **se pretende poder establecer si algo es verdadero o no**. En el segundo, **uno está más interesado en la magnitud precisa de los efectos involucrados**. Obviamente, las dos preguntas están estrechamente vinculadas.



Franklin Marvin Fisher. [Guía de sus trabajos 1958-2009](#)

Existe un tercer uso, pero menos extendido, al que se puede aplicar el análisis de regresión múltiple: **pronosticar los valores de alguna variable**. Un análisis de regresión múltiple muestra cómo ciertas variables independientes afectan a una variable dependiente. A partir de ese análisis, y de un pronóstico de los valores de las variables independientes (obtenido de alguna otra "fuente"), se puede generar un pronóstico de la variable dependiente. Este tipo de **pronóstico incondicional** no siempre es útil, lo cual es una suerte, ya que tales pronósticos incondicionales tienden a ser relativamente inexactos. Mucho más a menudo lo que interesa es un **pronóstico condicional**, una predicción de lo que sucederá a la variable dependiente si es cambiada otra variable o, mirando retrospectivamente, qué habría sucedido a la variable dependiente si el valor de una variable independiente hubiera sido diferente.

⁷ Charles River Associates, *The Audience-Revenue Relationship for Local Television Stations* (1978) (FCC Docket No. 21284); Charles River Associates, *The Value of Different Day Parts in the Audience-Revenue Relationship for Local Television Stations* (1979) (FCC Docket No. 21284). El primer estudio sobre el problema fue, creo, el que presenté como testimonio por escrito a la FCC en 1964. Está presentado en Fisher, Ferrall, Jr., y otros, *Community Antenna Television Systems and Local Television Station Audience*, 80 QJ. Econ. 227 (1966). Ambos estudios fueron preparados para la National Association of Broadcasters.

Tomen los dos ejemplos ya descritos. La pregunta en el caso de los bomberos del ferrocarril en realidad no implicaba predecir el número de accidentes ferroviarios. Más bien, buscaba tratar de decidir si el número de accidentes sería significativamente mayor si los bomberos del ferrocarril ya no estuvieran empleados. De modo similar, si bien la predicción de los ingresos de los canales de TV sería deseable para algunos fines regulatorios, el tema principal del estudio de ingresos de la audiencia era la medición sistemática de los efectos sobre los ingresos de cambios en el tamaño y las características socioeconómicas de la audiencia de un canal.

El ejemplo de los bomberos resalta mejor los problemas que conlleva hacer tales predicciones. Por su naturaleza, los accidentes ferroviarios involucran eventos aleatorios, al azar. Incluso la tasa de accidentes (cualquiera sea su medida) está sujeta a tales fluctuaciones casuales. *Simplemente determinar si la presencia o ausencia de bomberos introduce una diferencia significativa en la tasa de accidentes ferroviarios puede ser más fácil que predecir la tasa en sí con gran precisión.* Una de las características distintivas del análisis de regresión múltiple es que puede proporcionar información sobre los efectos de la variable de interés (en este caso, el empleo de los bomberos) sobre la variable dependiente (aquí, accidentes ferroviarios) sin requerir necesariamente predecir la variable dependiente en sí con gran precisión.

En cierto modo, la *regresión múltiple se podría describir como un método utilizado para extraer una señal sistemática del ruido que presentan los datos.* Hay dos problemas principales involucrados en la extracción de esa señal. En primer lugar, es típico que el factor cuya influencia se desea probar o medir *no sea el único factor* importante que afecta a la variable dependiente; por ejemplo, la cantidad de tráfico en los ferrocarriles también tiene que ver con los accidentes. En segundo lugar, incluso si se puede de alguna manera dar cuenta de los efectos de los otros factores sistemáticos importantes, generalmente *quedan componentes aleatorios.*

Si pudiéramos hacer experimentos controlados, sería relativamente fácil cuantificar la relación que está siendo investigada. Un experimento controlado en el caso de ingresos derivados de la audiencia, por ejemplo, variaría las audiencias de la estación y otras variables que se espera que influyan en los ingresos una por una, manteniendo todo lo demás constante y observando los ingresos resultantes. Obviamente, esto es imposible; no hay forma de que podamos jugar con las audiencias de la estación. Esto significa que debemos contentarnos con analizar, por así decirlo, los *experimentos realizados por la naturaleza*, en los que más de una de las variables consideradas que pueden afectar los ingresos se mueven al mismo tiempo.

Además, aunque pudiéramos controlar las audiencias del canal y mantener constantes las variables que consideramos importantes, no sabríamos lo suficiente sobre la relación audiencia-ingresos para asegurarnos de mantener constantes todas las variables que afectan realmente los ingresos de un canal individual. Puede ser, por ejemplo, que la personalidad y la eficacia de los representantes de ventas de los canales o las políticas publicitarias o la calidad de publicación de los periódicos competidores afec-

ten los ingresos. Estas variables son difíciles de medir, y mucho más de mantener constantes.⁸

La incapacidad de realizar experimentos controlados no es excepcional. De hecho, ocurre incluso cuando se hacen experimentos controlados en ciencias naturales. La diferencia es que aquí se puede estar bastante seguro de que los efectos incontrolados que no se conocen en detalle son extremadamente pequeños. Cuando se trata de observaciones del sistema económico (o, de hecho, de cualquier sistema en donde los experimentos son realizados por la naturaleza y no por el experimentador), es probable que haya un **elemento residual no trivial con efectos desconocidos sobre la variable de interés, incluso después de haber tenido en cuenta los principales efectos sistemáticos.** **La regresión múltiple es una forma de enfrentar estas dificultades.**

B. Cómo funciona la regresión múltiple

1. Una visión general. En regresión múltiple, primero se especifican las principales variables que se cree que influyen sobre la variable dependiente. En nuestros ejemplos, esto significa especificar las influencias importantes o sistemáticas que puedan afectar los ingresos del canal de TV o la seguridad ferroviaria. Inevitablemente, quedarán influencias de menor entidad, cada una quizás muy pequeña, pero creando en combinación un efecto no despreciable. A estas influencias menores se las trata ubicándolas dentro lo que se denomina un término de perturbación aleatorio y suponiendo que su efecto conjunto no está relacionado sistemáticamente con los efectos de las principales variables que son investigadas; en otras palabras, tratando sus efectos como debidos a la probabilidad.⁹ Obviamente, es muy deseable que la parte aleatoria de la relación resulte pequeña, en particular con relación a la parte sistemática. De hecho, el tamaño de la parte aleatoria proporciona una indicación de cuán correctamente uno ha juzgado cuál es la parte sistemática. Por tanto, la regresión múltiple proporciona un medio no sólo para extraer los efectos sistemáticos de los datos, sino también para evaluar qué tan bien se ha logrado hacerlo en presencia de los efectos aleatorios restantes.

La relación entre la variable dependiente y la variable independiente de interés se estima extrayendo los efectos de las otras variables principales (que conforman la parte sistemática). **Cuando esto ha sido hecho, se tiene el mejor sustituto disponible de una experimentación controlada.** Los resultados de las regresiones múltiples se pueden leer para mostrar los efectos de cada variable sobre la variable dependiente, manteniendo las otras constantes. Además, esos resultados permiten hacer afirmaciones sobre la probabilidad de que el efecto descrito haya sido observado meramente como resultado de una fluctuación aleatoria.

2. Estimación de regresiones múltiples. Supongamos que la relación a examinar incluye sólo dos variables, la variable dependiente (Y) y una variable independiente (X). Supongamos además (por simplicidad de exposición) que se piensa que la relación entre

⁸ Del mismo modo, en el caso de los bomberos, incluso si pudiéramos experimentar con el empleo de bomberos, no podríamos mantener constante el tráfico ferroviario. Además, otras variables que afectan la seguridad (lo que llamamos "azar") nunca se conocen con precisión.

⁹ Las perturbaciones (parte aleatoria o no sistemática) afectarán la dispersión de los valores verdaderos de la variable dependiente alrededor de los valores que se predecirían solamente a partir de la parte sistemática.

estas variables es una línea recta.¹⁰ Tal relación podría ser expresada matemáticamente como:

$$(1) \quad Y = a + b X$$

o bien en un diagrama como en la Figura 1. El problema para el investigador es descubrir los valores de los parámetros, a y b (es decir, la *intersección al origen* y la *pendiente* de la línea). Si la relación fuera realmente exacta, si no hubiera influencias aleatorias, sería extremadamente fácil de hacer. Sólo se requeriría observar dos puntos con valores diferentes de X . Dado que dos puntos determinan una línea, solo se requeriría un cálculo aritmético de rutina para encontrar la línea que determinan.

En la vida real, sin embargo, las relaciones a ser ajustadas no son exactas. Más bien hay influencias aleatorias en la variable dependiente, como se describió anteriormente. Por lo tanto, la relación correcta no suele ser la ecuación (1) sino más bien:

$$(2) \quad Y = a + b X + u$$

donde u representa las *influencias aleatorias*. Diferentes valores de u producirán diferentes valores de Y que estarán por encima o por debajo de la línea; de hecho, darán lugar a una dispersión de puntos como se muestra en la Figura 1.

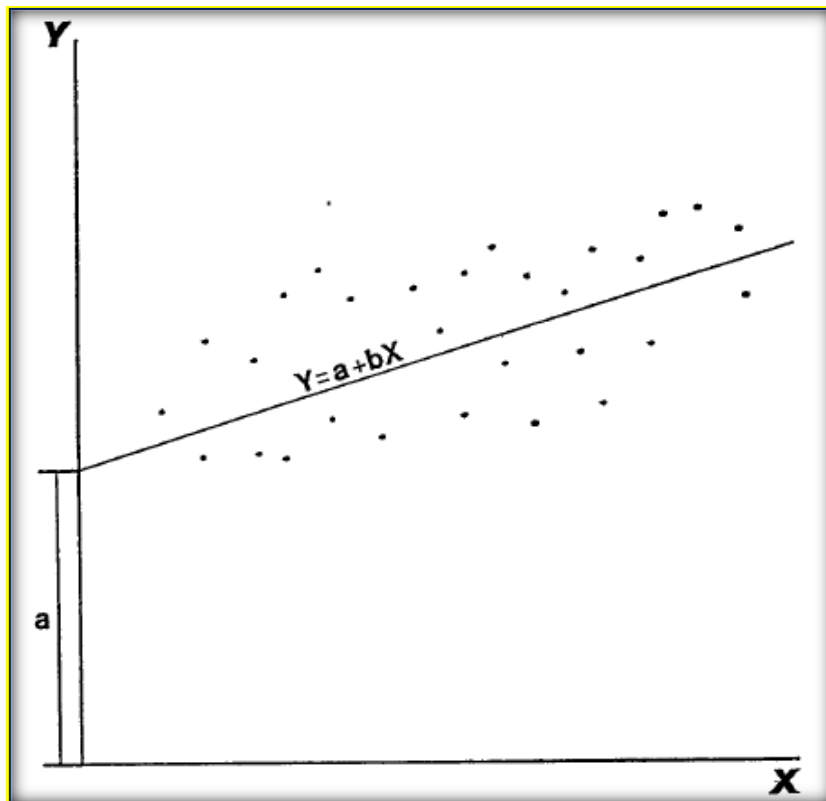


Figura 1

¹⁰ Elegí el caso de línea recta como el más fácil de entender, pero la teoría no es tan restringida. No hay nada que impida que una o más de las variables en la ecuación (1) sean un cuadrado, un logaritmo o el cociente de otras dos variables. Muchas (no todas) las relaciones matemáticas se pueden convertir en la forma de la ecuación (1) transformando o redefiniendo las variables. Además, la mayoría de las relaciones no lineales pueden ser aproximadas al menos mediante líneas rectas.

La tarea del investigador es traspasar el ruido generado por estas influencias aleatorias y extraer la "señal" $a + bX$, es decir, la línea alrededor de la cual se dispersan los puntos. Esto se hace seleccionando la línea que se adapta mejor a la dispersión de puntos en el sentido de que la suma de los desvíos cuadrados entre los valores Y predichos y reales es minimizada:¹¹ a esto se llama "regresión por mínimos cuadrados". (El adjetivo "múltiple" es utilizado cuando hay más de una X .)

3. Supuestos de la Regresión por Mínimos Cuadrados. En la práctica, la regresión por mínimos cuadrados no es realizada en forma de diagrama sino numéricamente (generalmente mediante computadora), lo que da como resultado estimadores numéricos de a y b . La relación que tales estimadores probablemente tengan con los valores verdaderos de a y b depende de los supuestos que uno está dispuesto a hacer sobre el término de perturbación aleatoria, u , y su relación con la variable independiente, X , cuyo efecto sobre Y (representado por b) debe ser medido.

En esencia, aquí hay tres supuestos principales implicados: (a) que los efectos del término aleatorio son independientes de los efectos de la variable independiente; (b) que los valores del término aleatorio de distintas observaciones no están sistemáticamente relacionados y que el tamaño cuadrado promedio del efecto aleatorio no tiene una tendencia sistemática a cambiar a lo largo de las observaciones; y (c) que la suma de los efectos aleatorios incorporados en el término de perturbación se distribuye normalmente, como una "curva de campana" que es una característica general de la distribución de la suma de efectos aleatorios independientes.¹²

La validez de estos supuestos tiene que ver con la efectividad y fiabilidad del análisis de mínimos cuadrados. Varias propiedades de la regresión múltiple dependen de la precisión de los supuestos, y diferentes supuestos implican diferentes propiedades. Además, la dependencia es acumulativa: si los supuestos iniciales son inválidos, las propiedades asociadas a los supuestos posteriores probablemente no estén presentes. En situaciones donde los supuestos pueden fallar, es probable que el uso del análisis de regresión múltiple sea inapropiado.¹³

a. Independencia del término de perturbación. El supuesto fundamental de la regresión de mínimos cuadrados es que los efectos no controlados de la perturbación aleatoria (u) son, en un sentido técnico apropiado, independientes de los efectos controlados de la variable independiente (X). (Alternativamente, esto puede expresarse como el supuesto de que el término de perturbación tiene una media cero cualquiera que sea el

¹¹ Usar la suma de las desviaciones al cuadrado da idéntico peso a los desvíos positivos y negativos. Además, en un diagrama multidimensional (no dibujado) se puede demostrar que en un sentido, minimizar la suma de los desvíos al cuadrado equivale a minimizar la distancia entre el punto que representa los valores reales de la variable dependiente y el punto que representa los valores que uno predeciría a partir de la regresión. El promedio de los valores cuadrados es la medida estadística estándar de dispersión.

¹² La palabra normal aquí es una jerga referida a la forma de la distribución. El nombre indica que la distribución involucrada es característica de muchas variables aleatorias. Lo más importante, si una variable aleatoria se compone de la suma de otras variables aleatorias que actúan independientemente, esa suma tiende a distribuirse normalmente. Esto hace que la hipótesis de normalidad sea obvia a menos que haya una razón convincente para apartarse de ella.

¹³ Como regla general, existen métodos para probar y resolver la falla de tales supuestos, pero implican usar herramientas más avanzadas de la econometría en lugar de la regresión por mínimos cuadrados.

valor de X . En muestras repetidas, el término de perturbación para cualquier X dada no es en promedio ni positivo ni negativo). Si así no fuera, entonces intentar determinar los efectos de X sobre Y no podría hacerse simplemente observando diferentes X e intentando promediar los efectos de u . En tal caso, los movimientos en X estarían sistemáticamente asociados con los movimientos en u y, si no se pone mucho cuidado, los estimadores de b incluirían no sólo los efectos directos de X sobre Y , sino también los efectos asociados de los movimientos en el término de perturbación, u .

¿Cuándo es probable que falle este supuesto? El caso más simple de entender ocurre cuando algún factor grande y sistemático, aparte de X , quedó fuera del análisis; esto es llamado **error de especificación**. En el estudio de la audiencia por niveles de ingresos, por ejemplo, resulta que el ingreso promedio de los hogares, así como el tamaño de la audiencia, afectan los ingresos de los canales de televisión. Supongamos, sin embargo, que no hemos pensado en esto, sino que simplemente hemos tratado de estimar el efecto del tamaño de la audiencia sobre los ingresos. (Aquí, los ingresos serían Y y el tamaño de la audiencia sería X .) En efecto, esto significaría que estamos colocando los ingresos del hogar en el término de perturbación. Sin embargo, si el ingreso de los hogares en los mercados de TV está positivamente asociado con el tamaño de la audiencia, entonces parte de lo que atribuiríamos a un tamaño de audiencia más amplio sería de hecho atribuible a mayores ingresos. En otras palabras, habríamos fallado en controlar por nivel de ingreso y la falta de dicho control sería importante.

Obviamente, el supuesto de que se han controlado todas las influencias importantes es básico para cualquier intento de medir esas influencias correctamente. Sin embargo, hay otras maneras en que se puede violar el supuesto de independencia entre la perturbación aleatoria y los factores incluidos. En general, esto ocurrirá cuando existan relaciones entre las variables dependientes e independientes además de la relación que se estima. Discutiré ejemplos específicos de tales casos en la parte II.

Si se justifica el supuesto de independencia entre u y X , entonces los estimadores mínimo-cuadráticos de los parámetros (a y b) tendrán ciertas propiedades deseables. Primero, **los estimadores serán insesgados**: serán correctos **en promedio**. Lo cual significa que si uno hiciera los cálculos para una muestra de un tamaño particular, y luego repitiera el procedimiento en numerosas muestras del mismo tamaño, obteniendo cada vez estimadores diferentes de a y b , el promedio de los estimadores así obtenidos serían los verdaderos valores de a y b . Para decirlo de otra manera, los estimadores mínimo-cuadráticos no tienden a errar en forma sistemática, ni por el lado alto ni por el bajo.

Además, si el supuesto de independencia es correcto, los estimadores mínimo-cuadráticos serán **consistentes**. La propiedad de consistencia significa que, a medida que aumenta el tamaño de la muestra, la probabilidad de obtener estimadores de mínimos cuadrados que difieran de los valores reales en más de una ϵ dada tiende a cero. Por lo tanto, a medida que haya más datos disponibles será más fácil extraer los verdaderos valores de a y b del ruido que presenta la parte aleatoria.

b. *Comportamiento del término de perturbación.* La consistencia es la propiedad mínima que uno pretende de un estimador. Pero hay muchos estimadores consistentes y, en algunas situaciones, incluso muchos insesgados. Además, la falta de sesgo sólo asegura que el estimador sea correcto en promedio; no indica qué tan lejos puede estar

en una muestra determinada. De manera similar, la consistencia garantiza sólo que uno se acercará a los valores verdaderos de los parámetros si uno sabe lo suficiente; no puede determinar cuánto necesita saber para acercarse. Es claramente deseable tener medidas de **confiabilidad**, es decir, medidas de cuán lejos uno puede esperar que se encuentren los estimadores. Además, dentro de la clase de estimadores insesgados o consistentes, es obviamente deseable elegir el que sea más confiable.

Con un supuesto adicional, la regresión de mínimos cuadrados resulta generar tales estimadores y provee estimaciones de su confiabilidad. Este supuesto se refiere a la naturaleza del término de perturbación aleatorio (u), más que a un supuesto relativo a su relación con X . El supuesto puede dividirse en dos partes.

En primer lugar, se supone que si se tiene información sobre el valor de u para ciertas observaciones, no se ganaría nada consiguiendo información sobre su valor para otras observaciones. Por ejemplo, si las observaciones se refieren a variables a lo largo del tiempo (años), un valor inusualmente alto y positivo de u no debe ser seguido por una tendencia a que sea alto el próximo año. Por el contrario, los valores sucesivos de u deberían ser independientes entre sí. Se puede ver por qué es probable que esto importe. La regresión de mínimos cuadrados es una forma generalizada de promediar. Promediar es una forma excelente de tomar en cuenta el ruido aleatorio, siempre que se promedie sobre eventos independientes. Sin embargo, si las perturbaciones aleatorias de diferentes observaciones no son mutuamente independientes, el promedio involucrado en la regresión por mínimos cuadrados no desactivará los efectos aleatorios. En tal caso, se podría hacerlo mejor suponiendo expresamente que un alto término de perturbación en un período indica algo sobre el valor del término de perturbación en el período siguiente, y luego usar esta información para intentar excluir la perturbación de la ecuación.

En segundo lugar, se supone que no hay una tendencia sistemática para que la perturbación aleatoria (u) sea grande o pequeña.¹⁴ Dicho de otra manera, **se asume que las posibilidades de un efecto aleatorio grande versus uno pequeño son las mismas para todas las observaciones.**¹⁵ De nuevo, se puede ver por qué esto importará. Si algunas observaciones tienden a tener efectos aleatorios más grandes que otras, entonces las observaciones con grandes efectos aleatorios contendrían información menos confiable que las observaciones con pequeños efectos aleatorios. En cualquier procedimiento de

¹⁴ Ya hemos supuesto que el término de perturbación aleatoria no tiene tendencia sistemática a ser alto o bajo, es decir, que tiene un valor medio o esperado de 0 para todos los valores de X (el "valor esperado" debe considerarse como la media de la población. A grandes rasgos, es el valor promedio que se espera obtener si se toma una muestra suficientemente grande). Ese supuesto implica el signo algebraico del término de perturbación aleatoria. El presente supuesto, por otro lado, tiene que ver con la magnitud absoluta del término de perturbación, independientemente del signo. Dicho de forma más precisa, la dispersión de una variable aleatoria se mide por el valor medio o esperado de la desviación al cuadrado de su media. Esto es llamado la "varianza". Su raíz cuadrada es llamada "desvío estándar". El supuesto previo del texto era que la media del término de perturbación aleatoria no está relacionada sistemáticamente con X . El que se está haciendo ahora es que la varianza o desvío estándar del término de perturbación no guarda tal relación y es, de hecho, el mismo para todas las observaciones.

¹⁵ Técnicamente, ésta es la propiedad de que la varianza del término de perturbación debe ser la misma para todas las observaciones.

promediado, se querría darle más peso a estas últimas. Como la regresión de mínimos cuadrados trata a todas las observaciones por igual, no tomará esto en cuenta.¹⁶

Estos supuestos serán violados si, al tratar con una serie de observaciones a lo largo del tiempo, el término de perturbación incluye los efectos de variables que se comportan sistemáticamente a lo largo del tiempo. Ciertamente, ésta es una posibilidad seria en modelos econométricos. De modo similar, si las observaciones son de entidades individuales, como empresas, bien podría ser que los efectos de eventos particulares no controlados (como eventos políticos) sean mayores para las empresas grandes que para las pequeñas. En tal caso, la segunda parte del supuesto estaría siendo violado. Sin embargo, al igual que con todos los supuestos de la regresión mínimo-cuadrática, uno querría estar seguro de que las violaciones son realmente importantes antes de abandonar el análisis de regresión. En los casos planteados anteriormente, pequeños desvíos de los supuestos tendrían pequeños efectos. Además, las propiedades de los mínimos cuadrados asociadas con los supuestos son tan fuertes como para hacer que la regresión por mínimos cuadrados resulte superior a estimadores alternativos que buscarían tratar de subsanar tales pequeños desvíos.

Dada la validez de los supuestos en discusión, los estimadores de mínimos cuadrados serán eficientes. Esto significa que, dentro de una amplia clase de estimadores insesgados y consistentes, los estimadores mínimo-cuadráticos tendrán la menor varianza. Por lo tanto, si se pudiera tomar muestras repetidas, la varianza de los estimadores de mínimos cuadrados alrededor de los valores verdaderos de a y b sería menor que la varianza de otros estimadores insesgados y consistentes; en resumen, los estimadores de mínimos cuadrados serían más fiables.

c. *Normalidad de la distribución.* El último supuesto de mínimos cuadrados impone mayores restricciones a la perturbación aleatoria, u , que las que ya hemos discutido. El supuesto es que u , para todos los valores de X , sigue la distribución normal (curva de campana),¹⁷ con una media de cero, como ya se asumió. Esto, sin embargo, no es tan restrictivo como pueda parecer. Como proposición general de estadísticas, la distribución normal es característica de grandes promedios de efectos aleatorios independientes. En la medida en que el término de error esté compuesto por la suma de pequeños efectos aleatorios, esa suma tenderá a distribuirse normalmente.¹⁸

El supuesto de normalidad, además de reforzar la propiedad de eficiencia de mínimos cuadrados, implica la capacidad de hacer enunciados de probabilidad precisos con respecto a cuán lejos pueden estar los estimadores de mínimos cuadrados.¹⁹

4. *Múltiples variables independientes.* En la práctica, generalmente no se trabaja con relaciones que involucren sólo dos variables, sino con relaciones en las que una variable

¹⁶ Hay formas de tener en cuenta esta falla del supuesto: no es sorprendente que la técnica involucrada se llame mínimos cuadrados ponderados, una variedad de mínimos cuadrados generalizados.

¹⁷ Véase nota 12 *supra*.

¹⁸ Ver nota 12 *supra*. La distribución "normal" está completamente caracterizada por su media y varianza. Es difícil construir ejemplos prácticos en los que uno esté inclinado a cuestionar la normalidad sin cuestionar los otros supuestos sobre el término de perturbación aleatoria. Por lo tanto, aunque hay contrastes del desvío de normalidad, casi nunca se utilizan.

¹⁹ Ver el texto que acompaña a las notas 24-28 *infra*.

dependiente está influida por muchas independientes (tráfico ferroviario y empleo de bomberos, ingresos de la audiencia y tamaño de la audiencia). Denotando las variables independientes como X_1, X_2, \dots, X_k , la relación a estimar (suponiendo linealidad)²⁰ se puede expresar como:

$$(3) \quad Y = a + b_1 X_1 + b_2 X_2 + \dots + b_k X_k + u$$

Si sólo hay una variable independiente, éste es el caso ya considerado, el caso de una línea recta. Cuando hay dos variables independientes, se está ajustando un plano a una dispersión de puntos en el espacio. Cuando hay más de dos variables independientes, se está ajustando un hiper-plano (la generalización de un plano a más de tres dimensiones), pero los principios siguen siendo los mismos, aunque la visualización ya no sea inmediata. Los cuadrados mínimos aún conservan todas las propiedades enumeradas para el caso simple anterior.

La regresión por cuadrados mínimos aprovecha el hecho de que las variables independientes rara vez se mueven perfectamente juntas, sino que se mueven (como su nombre lo sugiere) independientemente. Al determinar cómo cambia la variable dependiente cuando las variables independientes se mueven en una variedad de formas diferentes, se extrae el efecto de cada una de las variables independientes.

Este tipo de extracción sistemática de los efectos de cada variable es importante. El examen de los datos brutos conduce a conclusiones fáciles y, a veces, erróneas. Con el tiempo, por ejemplo, se produjo la eliminación de bomberos y un mayor número de

²⁰ De nuevo, he elegido una forma lineal aquí. La teoría de mínimos cuadrados se ejecuta principalmente en términos de tales formas, pero esto no es tan restrictivo como podría parecer, ya que muchas formas no lineales se pueden convertir en una forma lineal similar a la ecuación (3) mediante transformaciones apropiadas de las variables. El supuesto básico involucrado de linealidad es que el efecto de cada variable independiente sobre la variable dependiente es independiente del nivel de las otras variables independientes. Por lo tanto, en el ejemplo de los bomberos, la linealidad implicaría que el efecto de la presencia de bomberos sobre el número de accidentes ferroviarios sería el mismo a niveles elevados de tráfico que en los bajos. También implicaría que el efecto sería el mismo independientemente de si había otros miembros de la tripulación sustituyendo a los bomberos. Obviamente, estos no son supuestos sobre los que necesariamente cabría confiar. Afortunadamente, no es necesario confiar en ellos. Si uno pensara, por ejemplo, que dos de las variables -por ejemplo, X_1 y X_2 - interactuaron, entonces uno podría definir una nueva variable X_3 como el producto de X_1 y X_2 . La regresión de mínimos cuadrados procedería entonces como si X_3 fuera simplemente una variable diferente, pero su coeficiente diría algo acerca de la importancia de dicha interacción. Para tomar un ejemplo diferente, a menudo no es plausible suponer (como lo hace la linealidad) que el efecto sobre la variable dependiente de cambiar una variable independiente en una unidad debe ser el mismo en términos absolutos para todos los niveles de la variable independiente. Con frecuencia es más plausible suponer que un cambio del uno por ciento en una variable independiente tiene un efecto de porcentaje constante en la variable dependiente. Dichos casos pueden tratarse en el marco de la linealidad al ingresar en la ecuación (3), no las variables originales en sí, sino sus logaritmos. Esto se hace con frecuencia y también tiene la ventaja de suponer que el efecto del error aleatorio en diferentes observaciones es probablemente del mismo tamaño en términos porcentuales más que absolutos, un asunto que apareció arriba en la discusión de uno de los supuestos de mínimos cuadrados. Ver el texto que acompaña a las notas 14-16 supra. En general, la elección de la forma en la que se ingresan las variables o, más en general, la forma de la relación requiere una reflexión seria sobre la forma en que la relación estimada puede funcionar. Al igual que con la decisión de qué variable incluir en la relación, en primer lugar, esto debe hacerse en gran parte pensando en el problema en lugar de esperar que los datos proporcionen la respuesta. En cualquier caso, relaciones como la ecuación (3) son sustancialmente más generales de lo que podría parecer a primera vista.

accidentes. Que estos eventos estuvieran causalmente conectados no puede ser concluido si ambos también estaban asociados con aumentos en una tercera variable (tráfico ferroviario) que afecta plausiblemente los accidentes ferroviarios. Sólo utilizando sistemáticamente el hecho de que el tráfico ferroviario, si bien está asociado al empleo de bomberos en los datos, no está perfectamente asociado, se puede averiguar el efecto independiente de los bomberos. No controlar por tráfico ferroviario lo ubicaría en el término de perturbación de la ecuación (3) y violaría la hipótesis fundamental de mínimos cuadrados de que los términos de perturbación y las variables independientes son independientes.

Como sugiere esta descripción, es muy importante que las variables de hecho se muevan de manera un tanto independiente. Supongamos, por ejemplo, que en el estudio de audiencia de ingresos se desea investigar los efectos por separado sobre los ingresos del canal de audiencias cercanas a un canal (X_1) y audiencias ubicadas más lejos (X_2). Supongamos, como no es el caso, que cada vez que la audiencia cercana aumenta en un diez por ciento, a medida que uno pasa de canal en canal, la audiencia lejana también aumenta en un diez por ciento. Entonces, si bien se podría determinar la influencia en los ingresos de la audiencia total, no se podría descubrir los efectos por separado en los ingresos de las dos subdivisiones de esa audiencia. Ningún "experimento realizado por la naturaleza" habría separado esos efectos de ninguna manera.

Tal situación extrema generalmente no se encuentra en la práctica; más bien, lo que se encuentra es algo cercano a eso. Supongamos que cada vez que la audiencia cercana sube un diez por ciento, la audiencia lejana aumenta en cantidades que varían solo ligeramente hacia arriba o hacia abajo desde el diez por ciento. En ese caso, sería posible estimar los efectos separados generados por cada subdivisión del tamaño de la audiencia, pero el estimador sería muy incierto. **La naturaleza no realizaría experimentos calculados para separar esos efectos con un alto grado de precisión.** Tal circunstancia se llama **multicolinealidad**, así llamada porque implica una relación lineal entre las variables en el lado derecho de la ecuación.

Obviamente, cuanto menos multicolinealidad esté presente, mejor será la capacidad de separar los efectos de interés. Sin embargo, a menos que la multicolinealidad sea perfecta, la regresión múltiple podrá separar los efectos hasta cierto punto y, nuevamente, lo hará con mayor precisión que cualquier otro método, produciendo estimaciones con las propiedades discutidas anteriormente, así como medidas de confiabilidad de estas estimaciones. Los efectos de la multicolinealidad aparecerán en tales medidas de confiabilidad (errores estándar), como se detalla más abajo.²¹

5. ***Inclusión o exclusión errónea de variables.*** La discusión hasta ahora presume que la verdadera relación sistemática es la que se está estimando. Para decirlo de otra manera, ya vimos al discutir la carencia de sesgo que la regresión múltiple conserva las propie-

²¹ Ver el texto que acompaña a las notas 24-28 infra. Téngase en cuenta que el problema aquí se produce cuando dos de las variables independientes se mueven juntas de manera **aproximadamente lineal**. Si se mueven juntos de forma no lineal, no habrá un problema tan grave. Si lo que está involucrado no es otra relación entre dos o más de las variables independientes sino otra relación entre la variable dependiente y una variable independiente, entonces se violará la hipótesis básica de los mínimos cuadrados y tendremos una situación que involucra ecuaciones simultáneas como se explica más adelante. Ver el texto que acompaña a las notas 34, 35 y 43-45 infra.

dades deseables asociadas con ella sólo si uno incluyó todas las variables que probablemente tengan un efecto importante sobre la variable dependiente y puede suponerse con seguridad que los restantes efectos no están correlacionados con las variables independientes incluidas. En el estudio de ingresos del público, por lo tanto, fue necesario controlar el ingreso del hogar y no ubicarlo dentro del término de perturbación. Por lo tanto, es importante incluir en algún momento todas las variables que se podría pensar que tienen un efecto significativo sobre la variable dependiente. En general, esto se hace primero examinando las variables que se piensa que son realmente importantes y luego preguntando qué sucede cuando se incluyen variables adicionales.

Téngase en cuenta que esto debe hacerse especificando de antemano qué variables se consideran importantes. Si se procede primero mirando los datos y luego incluyendo aquellos factores que aparecen correlacionados con la variable dependiente se tiene una receta para resultados espurios. Lleva a que no se pueda hacer un verdadero test de la relación estimada. Además, es probable que se omitan las variables que realmente pertenecen a esa relación y, por lo tanto, se llegue a resultados inválidos y no contrastados. La medición que proporciona la regresión por mínimos cuadrados es una forma de hacer supuestos teóricos precisos o de testarlos; no es un sustituto del pensamiento.

Menciono esto enfáticamente porque hay diversos programas de computación empaquetados usados a veces que involucran lo que se conoce como regresión paso a paso. Estos programas generan regresiones múltiples mediante el siguiente procedimiento. Primero, el programa halla la variable independiente de la lista más correlacionada con la variable dependiente y obtiene una regresión que la involucra. Luego observa los desvíos muestrales de la regresión (las diferencias entre los valores reales y predichos) y se pregunta si esos desvíos están correlacionados con alguna otra variable independiente. Si es así, se agrega la variable más correlacionada con esos errores, etc. Esto no es recomendado. En primer lugar, incluso si ninguna de las variables independientes tiene algo que ver con la variable dependiente, es muy probable que procediendo de esta manera se presente una alta correlación en una muestra particular. En segundo lugar, las variables que de hecho pertenecen a la relación pero que están correlacionadas con las variables independientes utilizadas al inicio del procedimiento tienden a no ser seleccionadas nunca. En general, dichos programas informáticos sufren los mismos problemas que los intentos de mirar relaciones bilaterales que se observan de hecho que involucran la influencia de muchas variables: es probable que atribuyan los efectos de las variables omitidas a las incluidas y den lugar a estimaciones sesgadas.

Lo contrario de construir regresiones de una variable a la vez es poner muchas variables y luego ver si algunas de ellas deberían ser quitadas. Este es un método algo mejor. Si bien hay un efecto importante si se excluye una variable cuyo verdadero coeficiente es muy diferente de cero, el efecto de incluir erróneamente una variable cuyo verdadero coeficiente sea cero tiene muy poca importancia. A tal variable se la puede considerar como realmente estando presente en la relación, con un coeficiente cero que simplemente indica que la variable tiene poco o ningún efecto. La técnica de regresión múltiple luego estima ese coeficiente juntamente con los otros coeficientes verdaderos; por lo tanto, la técnica de regresión debe extraer un parámetro más del mismo número de observaciones. Esto es equivalente a tener una observación menos con la cual ex-

*traer parámetros distintos de cero.*²² Si el tamaño de muestra es grande (hay más de quinientos canales de televisión en Estados Unidos, por ejemplo), habrá un efecto muy pequeño sobre el estimador de los coeficientes restantes y de la predicción de la variable dependiente (a menos que la inclusión de la variable adicional agregue un problema de multicolinealidad). Las medidas de confiabilidad y las medidas de *bondad del ajuste*²³ tendrán plenamente en cuenta la leve reducción de información involucrada. *Donde sea posible, por lo tanto, puede ser mejor comenzar con un modelo más complejo y desmantelarlo.*

Empero, es importante darse cuenta de que no se puede desmantelar sin una teoría previa; no se recomienda el uso de programas informáticos que realicen una regresión regresiva paso a paso. Sin alguna teoría sobre qué variables pueden ser importantes, arrojar un gran número de variables a la tolva es probable que conduzca a resultados espurios. Si se prueban suficientes combinaciones de variables, entonces, en una muestra particular, podría haber alguna relación que parezca encajar bien. *Por lo tanto, un estudio bien hecho comienza con una idea teórica decente sobre qué variables es probable que sean importantes.* Luego puede procederse a contrastar hipótesis bien definidas sobre variables adicionales. Pero es muy probable que un estudio que busque una relación atractiva intentando todo tipo de posibilidades genere relaciones donde no exista ninguna.

Esto me lleva directamente a dos comentarios relevantes para los *abogados*. *Primero, cuando un experto hace un estudio, no se debe insistir demasiado en cubrir todas las posibilidades a la vez.* Más bien, uno debe asegurarse de que el experto proceda estimando un modelo razonable que incluya las variables principales y luego pruebe otras posibilidades. Si se insiste en que todas las variables posibles tengan la misma importancia, es probable terminar con un resultado dudoso.

Segundo, cuando uno se enfrenta a un experto opositor que ha realizado un estudio de regresión, cabe *averiguar cómo decidió el experto qué variables incluir y cuántas combinaciones diferentes de variables y modelos probó antes de decidirse por la que se presenta.* Si el modelo básico fue probado relativamente temprano y luego se intentaron las variantes simplemente para ver si algo más parecía importar, el estudio puede ser correcto. Sin embargo, si el modelo básico que se presenta es el resultado final de grandes cantidades de trabajo en la computadora, de trabajo sin sentido y mecánico, entonces uno puede tener un punto de disputa legítimo.

C. *Midiendo la “bondad de ajuste”*

²² Esto sucede porque, con el fin de evaluar la confiabilidad del estimador de la regresión, lo que importa es el número de *grados de libertad*: el exceso de la cantidad de observaciones sobre el número de parámetros a estimar. Lo siguiente transmite una idea de qué se trata. Siempre es posible ajustar una recta a dos observaciones, pero no hay grados de libertad y no hay forma de evaluar la confiabilidad del resultado. Si se tiene una tercera observación, entonces no siempre se podrá ajustar exactamente una recta, pero se podrá adquirir cierta noción de confiabilidad observando cuán cerca se está de hecho. Si se agrega otra variable con un coeficiente a ser estimado, en tal caso se está estimando un plano que se puede ajustar con precisión a tres observaciones. *Por lo tanto, la adición de otro coeficiente a ser estimado tiene el mismo efecto que la eliminación de una observación.*

²³ Véase el texto que acompaña a las notas 24-28 *infra*.

Como ya mencioné varias veces, la regresión por mínimos cuadrados no sólo estima los efectos de las variables involucradas en el modelo sino que también mide la certeza o exactitud de tales estimaciones. Además, proporciona medidas generales de qué tan bien el modelo se ajusta a los datos como un todo. Hay varias medidas diferentes involucradas y dado que cada una mide cosas diferentes, es importante tener claro las diferencias entre ellas.

1. *Error Estándar de los Coeficientes y Estadísticos-t*. Asociado con el valor estimado de cada coeficiente de regresión (a y b en las ecuaciones anteriores) hay un número conocido como el error estándar de ese coeficiente,²⁴ que mide la fiabilidad del coeficiente. En general, cuanto mayor sea el error estándar, menos confiable o menos preciso será el valor estimado del coeficiente.

Hablando en términos generales, en muestras grandes las chances son diecinueve sobre veinte de que el verdadero coeficiente se encuentre en un entorno de aproximadamente dos errores estándar del coeficiente estimado. *Las posibilidades son noventa y nueve sobre cien que se encuentre dentro de aproximadamente dos y medio errores estándar del coeficiente.*²⁵ (En muestras pequeñas, los límites tienden a ser más amplios.) Así, por ejemplo, si el coeficiente estimado es diez con un error estándar de dos, las posibilidades son diecinueve sobre veinte que el verdadero coeficiente se encuentre entre seis y catorce, y noventa y nueve sobre cien que se encuentra entre cinco y quince. Decir que las chances son diecinueve sobre veinte de que el verdadero coeficiente se encuentre entre seis y catorce, sin embargo, no significa que sea igualmente probable que el verdadero coeficiente esté en cualquier parte de ese rango. La cifra más probable es diez. La probabilidad de igualar la cifra correcta disminuye a medida que uno se aleja de diez y, como lo indica la pequeña diferencia entre los rangos de seis a catorce y cinco a quince, esa probabilidad disminuye muy rápidamente a medida que se aleja sustancialmente del estimador promedio.

Es convencional usar el error estándar de un coeficiente estimado para hacer un contraste estadístico de la hipótesis de que el coeficiente verdadero es realmente cero, es decir, que la variable a la que corresponde realmente no tenga efecto sobre la variable dependiente. Esencialmente, tales enunciados se construyen preguntando qué tan probable es que los rangos del tipo que acabamos de describir incluyan el cero. *Esto se hace tomando la razón del coeficiente estimado a su error estándar. Tal relación es llamada estadístico-t.*

En muestras grandes, un estadístico-t de aproximadamente dos significa que hay una chance menor que uno sobre veinte de que el coeficiente verdadero sea cero y que estamos observando un coeficiente mayor por casualidad. En tal caso, se dice que el co-

²⁴ Como se explica en la nota 14 supra, las dos medidas básicas de dispersión de una variable aleatoria son su varianza, que es el desvío cuadrado promedio alrededor de su media y su desvío estándar, la raíz cuadrada de la varianza. *El error estándar de un estadístico (aquí, el error estándar de un coeficiente de regresión) es, en un sentido aproximado, su desvío estándar esperado.* Más precisamente, es la raíz cuadrada del desvío cuadrático promedio que cabría esperar obtener si se usara el mismo procedimiento de estimación una y otra vez. Es una medida conveniente de la fiabilidad del estadístico con el que está asociado, ya que la probabilidad de que el estadístico difiera del valor verdadero en una cantidad dada depende directamente del número de errores estándar que representa la cantidad.

²⁵ Esto dependerá del supuesto de *normalidad*, discutido en el texto que acompaña a las notas 17-19 supra.

eficiente es **significativo al nivel del cinco por ciento**. Un estadístico-t de aproximadamente dos y medio significa que las chances son de sólo una en cien que el coeficiente verdadero sea cero; en ese caso, el coeficiente es "significativo al nivel del uno por ciento".²⁶ (En muestras pequeñas, el estadístico-t debe ser mayor para un nivel de significancia dado.) En el ejemplo numérico dado, el estadístico-t sería cinco (diez dividido por dos) y la probabilidad de que el coeficiente verdadero sea cero es extremadamente pequeña. El coeficiente sería significativo a un nivel muy superior al de uno por ciento.

Los estadísticos utilizan generalmente niveles de significación de cinco por ciento y uno por ciento para contrastar hipótesis. Es decir, dado un nivel de significación del cinco por ciento (o uno por ciento para un investigador más estricto) es prudente suponer que el verdadero coeficiente no es cero y que, por lo tanto, la variable que se prueba tiene algún efecto sobre la variable dependiente en cuestión. Algunos abogados pueden cuestionar que el uso de tales niveles imponga un estándar demasiado severo. ¿Por qué rechazar la hipótesis de que un cierto coeficiente sea cero sólo si la probabilidad de que los resultados obtenidos se deban al azar es del cinco por ciento o menos? Cuando la hipótesis involucrada tiene importancia legal (por ejemplo, cuando un coeficiente distinto de cero indicaría la presencia de discriminación por sexo en los salarios), **¿no tendría más sentido utilizar una norma de "preponderancia de la evidencia" y requerir sólo una significación del cincuenta por ciento?**

Tal enfoque, sin embargo, reflejaría una comprensión errónea de lo que realmente significan los niveles de significación. En particular, un nivel de significación del cincuenta por ciento no correspondería a un estándar de "preponderancia de la evidencia". El nivel de significación nos dice sólo la probabilidad de obtener el valor del coeficiente medido si el valor verdadero es cero; no da la probabilidad de que el valor verdadero del coeficiente sea cero, como tampoco restar el nivel de significación del cien por ciento da la probabilidad de que la hipótesis no sea verdadera. Debido a que, incluso con una muestra grande, es bastante posible obtener resultados que difieran del verdadero valor de un coeficiente, convencionalmente se piensa que debe haber una probabilidad muy alta de que el coeficiente no sea cero antes de que pueda afirmarse de manera concluyente que la variable asociada con el coeficiente tiene un efecto definido sobre la variable dependiente.

Esto no significa que sólo se presenten o consideren resultados significativos al nivel del cinco por ciento. Resultados menos significativos pueden ser sugestivos, aún si no son probatorios, y la evidencia sugestiva ciertamente vale la pena. En las regresiones múltiples, **nunca se debe eliminar una variable que tenga bases teóricas firmes para**

²⁶ Los ejemplos de significancia dados en el texto son para lo que se conoce como un *contraste de dos colas*. Por ejemplo, el nivel de significación del cinco por ciento asociado con un estadístico t de aproximadamente dos es la probabilidad de obtener un coeficiente estimado tan grande como el que realmente se obtuvo, ya sea positivo o negativo, si el coeficiente verdadero es realmente cero. En muchas situaciones, por ejemplo, no hay problemas en cuanto a si un coeficiente en particular es positivo o negativo; más bien, el único problema puede ser si es positivo o cero. En tal circunstancia, la prueba apropiada es un *contraste de una cola* en el que el cinco por ciento representaría la probabilidad de observar algún coeficiente positivo si el valor verdadero fuera realmente cero. El estadístico-t requerido para la significancia a un nivel dado en un contraste de una cola es menor que el requerido para el mismo nivel en un contraste de dos colas. En el caso del cinco por ciento, por ejemplo, lo que se requiere es aproximadamente 1.6 en lugar de 2.

ser incluida sólo porque su coeficiente estimado no resulta significativo en una muestra particular.

Con todo, el cálculo de los errores estándar de los coeficientes o los estadísticos-t correspondientes es una cuestión de considerable importancia. Es realizado de manera rutinaria por todos los profesionales, con los niveles de significación generalmente aceptados de cinco y uno por ciento como el punto en el que se rechaza la hipótesis nula. El hecho de no informar tales medidas de confiabilidad es una señal clara de que el estudio es sospechoso.

2. El Error Estándar de Regresión. Otra estadígrafo informado a menudo con los resultados de la regresión por mínimos cuadrados es el "error estándar de estimación" o "error estándar de regresión". No debe confundirse con los errores estándar de los coeficientes. El error estándar de estimación es una de las medidas resumidas que reflejan el grado en que la línea o el plano de regresión estimados se ajustan a los datos. En términos de la discusión previa, es una estimación de cuán ampliamente los puntos están dispersos alrededor de la línea (o el plano).

Más precisamente, el error estándar de estimación describe el desvío promedio de los valores reales de la variable dependiente en la muestra de los valores que se predecirían a partir de la regresión.²⁷ Por lo tanto, un error estándar de cero correspondería a un ajuste perfecto. Cuanto mayor sea el error estándar de estimación, más pobre será el ajuste, en el sentido de que será más importante el componente aleatorio no explicado.

El tamaño del error estándar de estimación dependerá de las unidades de medida de las variables. Por ejemplo, si midiéramos a la variable dependiente en centavos en lugar de pesos, el error estándar de estimación también estaría en centavos. Para juzgar si el error estándar de estimación es elevado o reducido, por lo tanto, se lo debe comparar con algo más. Una comparación posible implica calcular el coeficiente de correlación, tratado más abajo. Otra comparación posible consiste en evaluar el promedio de la variable dependiente y determinar qué porcentaje de ese valor constituye el error estándar. En general, el error estándar de estimación puede utilizarse para formular enunciados probabilísticos acerca de cuán alejados están los pronósticos del modelo. Si la muestra es de tamaño considerable, los pronósticos estarían alejados de la media en más de dos errores estándar solamente una vez sobre veinte veces.²⁸

²⁷ De hecho, no es calculado como un promedio aritmético. Más bien, es la raíz cuadrada del desvío cuadrático promedio en la muestra (ajustado por grados de libertad, ver nota 22 *supra*).

²⁸ Vinculado al error estándar de estimación, pero no idéntico, está el error estándar de pronóstico. Mide cuán confiables serían los pronósticos realizados a partir de la ecuación de regresión. Más precisamente, es la raíz cuadrada de las diferencias esperadas al cuadrado entre los valores reales de la variable dependiente y los valores pronosticados. El error estándar de pronóstico y el error estándar de estimación difieren por el motivo siguiente. En tanto que el error estándar de estimación mide la magnitud de los desvíos en el período muestral alrededor de la relación estimada, los errores de pronóstico involucrarán no solamente los desvíos de la relación estimada sino también el hecho de que la relación estimada podría diferir de la relación verdadera. Es bastante instructivo considerar por qué difieren estos dos errores estándar. En general, se espera estar más seguro sobre la ubicación exacta de la verdadera relación en puntos que caen dentro del rango de puntos ya observados en la muestra; se estará más inseguro sobre otros puntos que no resulten típicos de la muestra. El error estándar de pronóstico depende de cuán alejados de los valores muestrales típicos están los valores de la variable independiente en el período de pronóstico. Será mayor cuanto más alejados de la muestra estén dichos valores. Dada la locali-

Con todo, es importante tener en cuenta que un amplio error estándar de estimación no dice nada sobre la precisión con la que son medidos los efectos de las variables independientes. En forma semejante, un amplio error estándar de estimación no dice nada sobre la probabilidad de que los efectos de todas esas variables sean realmente nulos y que sólo están siendo observados efectos estocásticos. (Estas propiedades son evaluadas mediante los errores estándar de los coeficientes y los estadísticos-t ya descritos.) El error estándar de estimación es una forma de evaluar cuán importante es el componente estocástico del modelo; no dice nada acerca de los efectos de esta aleatoriedad sobre nuestra capacidad de medir la parte sistemática no aleatoria.

Un ejemplo puede aclarar esto. Supóngase que a un grupo de trabajadores se les paga a todos el mismo salario, w , por hora trabajada. Supóngase, además, que los trabajadores son empleados por diferentes números de horas. Ahora supóngase que a fines de cada semana cada trabajador cobra su salario y participa en un juego de ruleta con apuestas elevadas. En tal caso, el ingreso de cada trabajador vendrá dado por la suma del salario percibido en su empleo y sus ganancias o pérdidas en el juego de ruleta.

Ahora supóngase que se trata de estimar el salario ordinario por hora trabajada, w , a partir de los datos sobre la cantidad de horas trabajadas y el ingreso total, pero que no se puede observar en forma directa la paga salarial. Lo podremos hacer mediante una regresión donde la variable dependiente sea el ingreso total y la variable independiente sea el número de horas trabajadas; el coeficiente resultante de las horas trabajadas sería nuestro estimador del salario horario, w . La influencia del juego de ruleta sería, por supuesto, la parte aleatoria del modelo.

¿Cómo podríamos medir la precisión del estimador del salario horario? Éste sería medido en términos del error estándar *del coeficiente estimado* (w). Si tuviéramos una muestra amplia, el error estándar sería muy pequeño. (Ésta es la propiedad de *consistencia* de los mínimos cuadrados.) A pesar de todo, hallaríamos un elevado error estándar del *estimador* porque, sin importar lo que se haga, seríamos incapaces de estimar los efectos del juego de ruleta no sistemático. En semejante circunstancia, podríamos intentar concluir que hubo efectos no sistemáticos muy amplios que afectaron nuestra capacidad de predecir el ingreso total. Sin embargo, en ningún caso podríamos concluir a partir de ello que obtuvimos un estimador sesgado o no fiable del salario horario. Aún menos cabría extraer la conclusión de que el cambio del número de horas trabajadas no tiene efecto alguno sobre el ingreso (es decir, que el verdadero salario sería cero) o, para tomar el caso más extremo, que los trabajadores estarían indiferentes entre ser despedidos y no serlo. Enunciados de esta naturaleza estarían señalizados mediante errores estándar muy grandes del estimador del salario horario, el coeficiente de regresión de la cantidad de horas trabajadas, no por elevados errores estándar de estimación de la regresión.

Por lo tanto, un amplio error estándar de estimación de la regresión nos está diciendo que no sabemos *todo*. Esto no es lo mismo que nos digan que no sabemos *nada*. Lo cual es importante en la práctica. En el caso de los bomberos lo que está involucrado es la diferencia entre predecir bien la cantidad de accidentes y saber que el empleo de

zación de las variables independientes en el período de pronóstico, empero, el error estándar de pronóstico es proporcional al error estándar de estimación, que no depende de dicha localización.

bomberos tuvo un efecto sobre ese número. Si bien son conceptos relacionados, no se trata de lo mismo y su medición es diferente.

3. *El Coeficiente de Correlación*. El modo usual de normalizar el error estándar de estimación para distintas unidades es compararlos (o mejor, comparar su cuadrado) con una medida de variación total de la variable dependiente. Lo que esta comparación permite es separar la variación de la variable dependiente en torno a su media en una parte explicada por los movimientos de la variable independiente (la parte sistemática) y otra parte que no es explicada (la parte no sistemática). El coeficiente de correlación múltiple al cuadrado, R^2 , mide el porcentaje de variación explicada por la parte sistemática.²⁹

¿Cómo cabría interpretar los valores de R^2 ? Es obvio que un valor cero significa que los cambios en las variables independientes no explican en absoluto los movimientos de la variable dependiente. A mayor R^2 , mayor resulta la asociación entre los movimientos de la variable dependiente y las variables independientes. Un valor unitario significa que toda la variación de la variable dependiente resulta explicada por el modelo.³⁰ Más allá de estas observaciones, esta medida usada comúnmente debe ser tomada con mucha precaución, porque R^2 puede ser afectado por cambios que podrían ser triviales en la forma que el problema ha sido planteado.³¹

II. USO ADECUADO E INADECUADO DE LA REGRESIÓN MÚLTIPLE EN PROCEDIMIENTOS LEGALES

Hasta este punto, este artículo sobre la regresión múltiple en procedimientos legales ha versado primariamente sobre Regresión Múltiple. Es tiempo ahora de hablar sobre los procedimientos legales. Lo haré discutiendo tres áreas donde ha figurado el análisis de regresión múltiple: el análisis de discriminación salarial contra las mujeres; la determinación de daños en los casos de fijación de precios; y la evaluación del castigo como disuasorio del crimen. Estos tres ejemplos ilustrarán cierto número de puntos técnicos que ya fueron planteados y facilitarán algunas lecciones sobre lo que la regresión

²⁹ No nos detendremos aquí sobre por qué el coeficiente de correlación es un cuadrado.

³⁰ Cuán elevado debería ser el R^2 esperado depende del número de grados de libertad (Véase nota 22 *supra*). Si se tiene dos observaciones que son ajustadas mediante una línea recta, por ejemplo, tal ajuste siempre será exacto y R^2 será siempre igual a uno. Si la línea debe ajustar varias observaciones, un R^2 próximo a uno sería la evidencia más rotunda de que los movimientos de la variable dependiente son explicados por los movimientos de las variables independientes.

³¹ Por ejemplo, supóngase que en la relación entre audiencia e ingresos, se ha decidido que la verdadera relación fuera logarítmica, siendo la variable dependiente el logaritmo de los ingresos. Supóngase además que una de las variables independientes fuera el logaritmo del tamaño de la audiencia. Supóngase ahora que restamos el logaritmo del tamaño de la audiencia miembro a miembro, haciendo de esta forma que la variable dependiente sea el logaritmo del *ingreso por espectador* (que es igual a logaritmo del ingreso menos el logaritmo del tamaño de la audiencia). Obviamente, lo único sustantivo que esto implica sería restar uno del coeficiente del logaritmo de la audiencia. Pero esto también cambiaría el R^2 , que ahora medirá cuánta variación del logaritmo del ingreso *por espectador* estamos explicando, en lugar de cuánta variación explicamos del logaritmo del ingreso en sí. El valor resultante del R^2 podría resultar superior o inferior al valor original.

múltiple puede y no puede hacer. Por mi parte, creo que el análisis de regresión múltiple es un instrumento completamente apropiado para examinar la posible discriminación en los salarios, pero me resulta dudosa su utilidad en casos de fijación de precios y creo que es peligrosamente equívoco en los exámenes de disuasión.

A. Discriminación de Salarios

En este ejemplo, se presenta un caso contra una empresa en nombre de un grupo de sus empleadas mujeres. Se acusa a la firma de discriminar pagando a las mujeres menos que a los hombres. El objetivo del estudio estadístico es contrastar que ello sea así en verdad.

Supongamos que los hechos son tales que así parece ser. El salario pagado a la empleada femenina promedio es menor que el salario pagado al empleado varón promedio. Para simplificar, supongamos que consideramos sólo mujeres y hombres en trabajos similares.³² La empresa se defiende – o es probable que se defienda – diciendo que las mujeres no son en promedio tan calificadas como los hombres. En particular, tienen menos educación y menos experiencia laboral. También registran puntajes más bajos en algunos exámenes de aptitud.

Obviamente se trata de una defensa razonable, si fuera cierta. Sin embargo, para que lo sea debe darse no sólo que las mujeres en promedio sean menos calificadas según diversas mediciones, sino también que la diferencia de calificación dé cuenta de las diferencias en la paga. Si la empresa no paga a las mujeres bien educadas más que a los varones menos educados, en tal caso difícilmente pueda decirse que ésta sea la base de la diferencia entre los salarios de hombres y mujeres.

La regresión múltiple es apropiada para responder a este tipo de preguntas en forma bastante precisa. Aún más, de no haber un estudio de regresión múltiple es difícil apreciar cómo podría decidirse el caso. Una comparación en bruto de los salarios promedio de hombres y mujeres podría levantar una sospecha, pero no puede decir nada definido. En realidad, podría ser equívoca en ambas direcciones. Por caso, sería completamente posible en un escenario diferente que a las mujeres se les pague en promedio tanto como a los hombres pero que un análisis de regresión múltiple revele que hay efectivamente discriminación porque las mujeres están efectivamente *más calificadas* en las mediciones que toman en cuenta la variación de salarios de los hombres.³³

Regresando al problema original ¿cómo podría ser planteado esto en un contexto de regresión múltiple? Comenzamos haciendo algo que puede parecer innecesariamente alambicado pero que será útil más adelante. Definamos una variable S, del modo siguiente:

$$(4) \quad \begin{aligned} S &= 0, \text{ si el empleado es mujer} \\ &= 1, \text{ si el empleado es hombre.} \end{aligned}$$

³² Controlar por tipo de trabajo es algo que resulta obvio, y que también puede ser realizado mediante regresión múltiple.

³³ Véase Finkelstein, *The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases*, 80 Colum. L. Rev. 737 (1980).

S es llamada variable *dummy*, y es usada en situaciones donde se desea examinar variaciones discretas y no continuas – en particular, la clasificación en categorías. Tomemos la ecuación de regresión siguiente:

$$(5) \quad Y = a + bS + u,$$

donde Y denota el ingreso pagado a un determinado empleado. No es difícil de apreciar que la ecuación de estimación (5) mediante una regresión mínimo-cuadrática es simplemente otra forma de computar la diferencia en los pagos promedio recibidos por los hombres y las mujeres. Si $S = 0$, luego en promedio el pago vendrá dado por a; éste será el pago promedio de las empleadas mujeres. Si, por otra parte, $S = 1$, luego en promedio el pago vendrá dado por $(a + b)$; éste será el pago promedio de los empleados varones. La diferencia de promedios resulta así b, el coeficiente de S, y contrastar que este coeficiente sea significativamente distinto de cero permite contrastar si a los hombres se les paga efectivamente más que a las mujeres.

Pero naturalmente este contraste sólo concierne al enunciado original, que los hombres en promedio reciben pagos superiores a las mujeres y que la diferencia de paga no está explicada más que por fluctuaciones aleatorias. Este contraste es mejor que mirar simplemente la diferencia de pagas, pero aún no hemos abordado el problema de controlar las otras variables, a saber las calificaciones.

Tal control es fácil de realizar. Por ejemplo, supongamos por un momento que sólo hubiera una medida de la calificación (digamos, mediciones de un test de aptitud), que denotamos A. Modifiquemos ahora la ecuación de regresión (5) de la forma siguiente:

$$(6) \quad Y = a + bS + cA + u$$

La estimación de esta ecuación mediante regresión múltiple responde a la pregunta de si el sexo afecta los salarios, manteniendo constante el test de aptitud. Esto puede ser apreciado en la Figura 2.

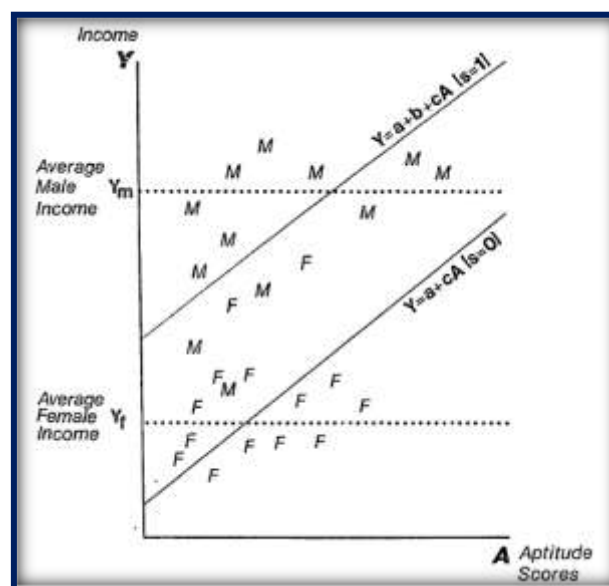


Figura 2

En la Figura 2, el ingreso de los empleados es graficado con relación a un test de aptitud. Los puntos que denotan al personal masculino son indicados con M; los puntos que denotan al personal femenino lo son con F. He dibujado un caso en el que a los hombres se les paga más en promedio que a las mujeres, pero en el que, también en promedio, las empleadas mujeres alcanzan un puntaje inferior en el test de aptitud que el obtenido por los hombres. Un examen de los salarios promedio sin corregir por el test de aptitud – equivalente a estimar por mínimos cuadrados la ecuación (5) – implica dibujar una línea horizontal en el diagrama (horizontal porque se supone que la aptitud no tiene ninguna incidencia en la ecuación (5)) al nivel del ingreso promedio de los hombres, y otra al nivel del ingreso promedio de las mujeres. Se aprecia que están bastante distantes entre sí. Por otro lado, corregir la estimación de la ecuación (6) por las mediciones de los test de aptitud, es equivalente a dibujar dos líneas paralelas que pasan a través de los puntos masculinos y femeninos, respectivamente. El hecho de que sean paralelas indica el supuesto de que los tests de aptitud deberían implicar el mismo efecto sobre el salario de los hombres que sobre el de las mujeres, si no hay discriminación. La diferencia de ordenadas al origen es el coeficiente de S, que mide la diferencia remanente entre los salarios una vez que se ha controlado mediante los tests de aptitud.

Ahora puede ser contrastado el enunciado de que los hombres ganan sistemáticamente más que las mujeres aún luego de controlar mediante el test de aptitud. Esto podría hacerse utilizando el estadístico-t asociado con b (el coeficiente de S) para apreciar si este coeficiente es significativamente distinto de cero. (Como no se supone que las mujeres ganen sistemáticamente más que los hombres, el contraste adecuado es de una sola cola.) La significación al cinco por ciento requeriría un estadístico-t algo por arriba de 1,6.

Este ejemplo puede ser extendido de varias formas que merecen ser discutidas. En primer lugar, no hay ningún motivo por el que debiera controlarse mediante un único test de aptitud. Elegí ese caso porque el diagrama resultante es más fácil de dibujar. Si hubiera diversas medidas posibles de la calificación, todas ellas podrían ser incluidas en la regresión como nuevas variables. Una de las grandes ventajas de proceder así en este problema es que no hay demasiadas variables que puedan explicar los salarios, con lo cual el interés está centrado simplemente en la cuestión de si el sexo es una de esas variables. No hay necesidad de revolver entre varias combinaciones diferentes de variables que podrían ser incluidas. Más bien, si se encuentra un efecto aparente en los datos brutos, la única cuestión es si dicho efecto obedece a no haber controlado mediante otras variables plausibles.

He planteado el problema en la ecuación (6) como si la única pregunta fuera si se le paga más a un hombre con una aptitud dada que a una mujer que tenga la misma aptitud. Esto está indicado en la Figura 2 mediante la diferencia constante entre las dos líneas inclinadas. Con arreglo a la ecuación, las mujeres registran una desventaja constante en pesos cualquiera sea su aptitud, y la pregunta es si esa desventaja es cero o no. Naturalmente, ésta no es la única posibilidad. Hay otra que resulta tan plausible, como que las mujeres tengan una desventaja porcentual constante, con lo que la diferencia en pesos es más elevada para las mujeres con mayor aptitud. Esto puede ser incluido fácilmente dentro del análisis. Sin hacer el diagrama, lo único que sería requerido es usar el logaritmo del ingreso en lugar del propio ingreso, como variable dependiente de la ecuación (6).

Podría probarse una variante algo más sutil. He planteado la ecuación (6) – o su equivalente logarítmico – de tal forma que lo que es contrastado es la hipótesis de que las mujeres están en desventaja, *dado* que las medidas de los tests de aptitud inciden sobre los salarios de hombres y mujeres de la misma forma (las líneas de pendiente positiva de la Figura 2 han sido dibujadas paralelas). Ésa es una buena forma de hacerlo, pero no es la única. Se podría estimar dos ecuaciones separadas de regresión – una para los hombres y otra para las mujeres – en la cuales el ingreso sería explicado por la aptitud. Se podría contrastar entonces *si los coeficientes de regresión de ambas ecuaciones son similares en todos los aspectos*. Después de todo, habría evidencia de discriminación si el efecto de los tests de aptitud sobre los salarios no fuera el mismo para hombres y mujeres. Es posible construir casos en los que b en la ecuación (6) termina siendo cero, pero en los que ecuaciones estimadas por separado proporcionan valores significativamente distintos de b para los hombres y las mujeres. Por otra parte, tratar de examinar varias cosas al mismo tiempo (por ejemplo, si un conjunto completo de coeficientes son los mismos para los hombres y las mujeres) produciría contrastes menos potentes que el examen separado de cada uno.

Hay otras dos características de este ejemplo que merecen ser comentadas. Primero, he usado en forma deliberada medidas de los tests de aptitud como una medida apropiada de la aptitud. Es bien sabido que estos tests no dan mediciones perfectas de la capacidad. Empero, esto no hace diferencia alguna a la validez del modelo de regresión. En la medida que la verdadera aptitud posea distintas dimensiones, lo crudo de las mediciones de los tests de aptitud como una medida válida puede ser corregido por las otras variables introducidas en la ecuación (6) – variables tales como los años de educación o la experiencia laboral. Segundo, lo que interesa en el problema actual es qué puede observar el empleador cuando busca distinguir la aptitud. El abogado defensor en este caso parecerá relativamente débil si anuncia que sólo tuvo acceso a una forma no medible de evaluar la aptitud y que todos los métodos de medición están sujetos a error. En efecto, lo importante de este problema no es una medida subyacente de aptitud sino la medida que el empleador puede observar y gratificar. Un argumento tal como que los tests de aptitud están sujetos a error debería ser desafiado por el querellante en busca de una medida más confiable y objetiva.

Dejando esto de lado, lo crudo de los números de los tests de aptitud podría implicar una diferencia sustancial si la variable verdadera (aptitud) fuera medida solamente mediante números con un error aleatorio. *En tal caso, puede demostrarse que el estimador de c , coeficiente del valor numérico del test de aptitud de la ecuación (6), estará sesgado hacia cero*. Esto es lo que cabe esperar, ya que usar variables que contengan mucho “ruido” es posible que genere estimadores que sugieren que tales variables carecen de efecto sistemático. Aún más importante es que el sesgo no estará restringido al coeficiente de la variable sujeta a error. En el presente caso, la variable S (que representa las diferencias de sexo) está correlacionada con la variable A (que representa mediciones del test de aptitud), reflejando que en la muestra de empleados, las mujeres tendrán una tendencia hacia un número de aptitud más bajo que los hombres. Esta correlación implica que el coeficiente de S también estará sesgado – y justamente este coeficiente es el que resulta interesante. Lamentablemente, sin información adicional no resulta posible indicar en qué dirección será el sesgo. Bajo ciertas circunstancias, hay pasos que pueden ser dados a fin de protegerse de los efectos de los errores de medición, pero discutir esto aquí me apartaría demasiado del campo presente.

El punto final que plantearé acerca de este ejemplo es que una predicción exacta de la variable dependiente, ingreso, no es requerida para una solución exitosa de este problema. Más bien lo que se requiere es un contraste directo de la significación de un coeficiente en particular. Su precisión – técnicamente, su *potencia* – depende del error estándar de ese coeficiente y no de lo bien o mal que la ecuación funciona para predecir la variable dependiente. *En general, contrastes como estos son más exitosos que los contrastes que dependen directamente de predicciones.*

Lo que hace que el ejemplo de discriminación salarial sea tan adaptado a la regresión múltiple es su sencillez y la facilidad con que puede ser adaptado al molde de un test de significación de un coeficiente de regresión particular. Nótese en particular la característica siguiente: *haya o no discriminación, uno esperaría que la versión ampliada de la ecuación (6) se ajuste bien a los puntos observados.* Lo que se hizo aquí es incorporar dentro de una teoría de la determinación del salario la diferencia que posiblemente involucra la discriminación sexual. Al menos a este nivel, la cuestión de qué factores distintos de la discriminación determinan los salarios puede ser considerada sin tener en cuenta si existe o no discriminación. Además, la presencia o ausencia de discriminación implica una diferencia definida claramente en los resultados que uno espera ver. Estas características hacen de este ejemplo uno en marcado contraste con el que veremos a continuación.

B. *Daños Antitrust en casos de Fijación de Precios*

En este ejemplo, el acusado ha perdido el caso sobre responsabilidad en un caso de fijación de precios, y la cuestión a ser decidida es la extensión de los daños. Sus abogados preparan un estudio que busca exhibir que el efecto de la fijación del precio fue mínimo, dado que el precio habría sido el mismo (o aún mayor) de no haber existido la conspiración.³⁴ Hay diversas maneras de hacerlo, pero tengo mis grandes dudas sobre todas ellas.

Una forma de proceder es tomar un atajo a partir del ejemplo de discriminación que hemos visto. En ese ejemplo, el estudio procedió controlando diversas variables y, finalmente, estimando qué ingreso se hubiera producido de no haber discriminación. ¿Por qué no tomar la misma idea estimando en forma sistemática qué precios se hubieran registrado de no haber habido fijación del precio? Pensemos en la siguiente manera de hacerlo. Bajo competencia, el precio queda determinado por la intersección de las curvas de oferta y demanda. Supongamos, para simplificar, que no hay sustitutos próximos del producto en cuestión, por lo que la demanda depende sólo del ingreso de los consumidores (o de la producción de los usuarios industriales) y del precio. La oferta dependerá del precio y los costos, los que a su vez dependen de los precios de los factores de producción. *Esto sugiere que podríamos explicar el precio mediante una ecuación de regresión que incluye el ingreso de los consumidores y los precios factoriales.*

Si bien podría suponerse que se debería incluir la *cantidad* como una de las variables con impacto sobre el precio, es más apropiado tratar al precio y a la cantidad en forma

³⁴ Dado que, bajo la regla *per se*, no constituye defensa la carencia de efectividad de una conspiración para fijar precios, tal exhibición sería irrelevante al efecto de la responsabilidad.

independiente, ya que, de hecho, las mismas fuerzas de mercado controlan a ambos. Esto resulta evidente examinando las ecuaciones específicas que determinan la oferta y la demanda del mercado.

Tanto la cantidad como el precio son determinados por la intersección de las curvas de oferta y demanda. Suponiendo linealidad, para simplificar, podemos escribir la curva de demanda como:

$$(7) \quad Q = a + bP + cY + u.$$

Aquí, Q indica cantidad, P denota precio e Y indica ingreso de los consumidores. Como antes, u es una perturbación aleatoria. En forma similar, podemos escribir la curva de oferta como:

$$(8) \quad Q = d + eP + fW + v$$

Aquí, W es una medición de los precios factoriales y v es otra perturbación aleatoria.

Las ecuaciones (7) y (8) conforman las así llamadas las *ecuaciones estructurales* de un *sistema de ecuaciones simultáneas*. Tal sistema involucra la interacción de más de una ecuación – ecuaciones que deben ser resueltas en forma simultánea. El hecho de que el precio sea determinado por la intersección de la oferta y la demanda está reflejado en el hecho de que P y Q deben tener el mismo valor en ambas ecuaciones. Podemos entonces resolver ambas ecuaciones para estas dos variables, igualando los valores que cada ecuación predice para la variable “cantidad”. Para hacerlo, introducimos nuevos coeficientes (π_0 , π_1 , etc.) que dependen de todos los coeficientes de las curvas de oferta y demanda. Al hacerlo así, la solución del precio aparecerá como lo siguiente:

$$(9) \quad P = \pi_0 + \pi_1 Y + \pi_2 W + u^*,$$

donde u^* es la perturbación aleatoria, que depende de algunos coeficientes de las curvas de oferta y demanda, así como de u y v. (Más precisamente, es igual a $(u-v)/(e-b)$). No nos detengamos en el álgebra exacta de estas expresiones. Habrá otra solución similar para Q.

La ecuación (9) y su compañera para Q son llamadas la *forma reducida* del modelo. Muestran las variables precio y cantidad en términos de variables determinadas por fuerzas distintas de las que están siendo modeladas (Y, W, u y v). Tales formas reducidas pueden ser estimadas mediante regresiones mínimo-cuadráticas.

Sería erróneo, empero, incluir Q en la ecuación de P. No aparecerá en la ecuación (9) por el simple motivo de que cantidad y precio son conjuntamente determinados por las mismas fuerzas, y no puede decirse que una determine a la otra. Una regresión que incluya cantidad de un lado y precio del otro debería interpretarse como un intento de estimar ya sea la ecuación (7) o la ecuación (8) directamente, pero esto no puede hacerse en forma consistente mediante mínimos cuadrados. La forma más fácil de verlo es la siguiente. Un movimiento del término de perturbación en la ecuación (7), u, afecta a la cantidad, Q; se trata de un cambio aleatorio de la curva de demanda. Pero los cambios aleatorios de la curva de demanda afectan no sólo a la cantidad, sino también al precio. Luego, cambios de u estarán asociados con movimientos de P, como puede ser obser-

vado directamente en la ecuación (9) y por el hecho de que u^* depende de u . Esto implica que, al estimar la ecuación (7), el supuesto fundamental mínimo-cuadrático (a saber, que las perturbaciones aleatorias se desplazan en forma independiente de las variables independientes) es violado. La ecuación (7) puede ser estimada, pero mínimos cuadrados no es la manera de hacerlo.

Luego, tratar de determinar el precio que regiría en un mercado competitivo mediante una regresión del precio versus un conjunto de variables que incluyen la cantidad está condenado al fracaso. Supóngase, empero, que somos más inteligentes y practicamos una regresión del precio contra el ingreso y los precios factoriales (Y y W), estimando así en forma directa la ecuación (9) y usamos esta ecuación para predecir el precio si no hay un acuerdo de fijación del precio.

Esto es mejor, pero aún no es adecuado. Aquí el problema es que no habrá una distinción clara entre los resultados que serían conseguidos si el mercado fuera afectado por un esquema de fijación del precio y los resultados que se darían en caso contrario. Si el mercado no fuera competitivo, y en su lugar fuera afectado por una fijación del precio, el precio no quedaría determinado por la intersección de las curvas competitivas de oferta y demanda. Más bien, sería determinado por los fijadores del precio. Que no fijarían un precio arbitrario, sino maximizando beneficios en la mayor medida posible.

Sin entrar en gran detalle, no es difícil ver que la maximización del beneficio habría requerido considerar la posición y la forma de la curva de demanda (7), como así también de los costos de producción. En términos habituales de los economistas, la maximización del beneficio requiere igualar el ingreso marginal y el costo marginal. El ingreso marginal dependerá directamente de la demanda, y el costo marginal dependerá directamente de los precios factoriales. El precio y la cantidad que igualan el ingreso marginal y el costo marginal dependerán, como en la ecuación (9), del ingreso y de los precios factoriales. En efecto, en cuanto al precio, se puede terminar con una ecuación idéntica a la ecuación (9); la diferencia que involucra la fijación del precio es que los coeficientes de la ecuación (9) serán distintos bajo la fijación del precio y bajo competencia.

Esto significa, empero, que no tiene sentido estimar la ecuación (9) directamente y usarla para predecir el precio. La ecuación (9) sería válida hubiera o no fijación del precio y no se podría decir si las predicciones que genera son competitivas o no. El caso del ejemplo de discriminación salarial fue diferente. Allí la cuestión fue definida marcadamente como si cierto coeficiente era o no nulo. Aquí la cuestión podría describirse como involucrando diferencias en cierto conjunto de coeficientes (los π en la ecuación (9)), pero sólo podremos estimar a estos coeficientes una vez y no habrá manera de que puedan ser comparados los valores que obtengamos con los valores que habrían sido obtenidos – ya sea en el caso competitivo como bajo la hipótesis no competitiva.

¿Significa esto que nada puede hacerse? No, pero casi. Podríamos operar de forma algo más sofisticada y tratar de estimar (7), la curva de demanda, que es la misma bajo ambos regímenes. También podríamos buscar saber cómo hubiera sido la curva de oferta competitiva. Esto podría hacerse teóricamente, pero en la práctica es probablemente imposible. La curva de demanda de la ecuación (7) podría ser estimada. Como se ha visto, no puede ser estimada mediante mínimos cuadrados bajo la hipótesis de compe-

tencia, pero hay otros métodos de estimarla, y estos métodos seguirán siendo válidos en general, aún bajo un esquema de fijación del precio.³⁵ Empero, para averiguar qué precio hubiera regido bajo condiciones competitivas, es necesario estimar la curva de oferta competitiva. No se lo puede hacer directamente porque hacerlo así requiere suponer que las observaciones fueron generadas por una demanda y una oferta competitiva. ¡Pero esto es justamente lo que se tiene que probar! Luego, habrá que buscar otra forma. En general esto implicará estimar la curva de costo de los productores y calcular el costo marginal. Aún si los acusados están dispuestos a entregar la información requerida para su cálculo, es probable que resulte extraordinariamente difícil hacerlo. Una vez que se dejan de lado los ejemplos simples de producto único, los cálculos de costo (y también la estimación de varias curvas de demanda) se tornan muy complicados. Se requieren grandes volúmenes de datos, los que en su mayoría no estarán presentados de manera útil, con una tenue promesa en la otra punta. En realidad, si se va a analizar directamente información de costos, podría resultar mejor hacer una exhibición directa de que los precios están aproximados a los costos marginales. Para eso, no se requeriría usar la demanda.

Queda una posibilidad dentro de esta área que luce levemente más prometedora. Muchos de los problemas que se ha descrito aparecen porque se desea saber cómo hubiera sido el mundo competitivo sin suponer que la competencia existió en los hechos. Sin embargo, si hay acuerdo en que la conspiración para la fijación de precios fue efectiva sólo por un lapso limitado, se podría considerar la opción de estimar la ecuación en forma reducida del precio (9) y la ecuación compañera de la cantidad, usando sólo datos del período competitivo. Luego estas ecuaciones podrían ser usadas para proyectar precios en el período de fijación del precio y estudiar las diferencias en los resultados.

Este tipo de programas luce factible, al menos en principio.³⁶ Lamentablemente, es improbable que dé resultados en la práctica. Uno sería utilizar las ecuaciones estimadas para pronosticar fuera del período muestral. Si han cambiado las condiciones (y éstas suelen hacerlo a lo largo del tiempo) esto involucra pronosticar más allá de la media de la muestra. Aún si el modelo es completamente correcto, no se podrá hacer este tipo de pronóstico con gran certidumbre. Es probable que se encuentre que el precio en un momento determinado durante el período de fijación del precio no sea significativamente mayor que el precio predicho por el modelo competitivo, pero que el error estándar de esa predicción sea amplio. Luego, aunque sea posible contrastar si la diferencia de precio es significativa, será posiblemente muy difícil decidir cuánto de esa diferencia es atribuible al error aleatorio.³⁷ Además, cambios de precios en cualquier

³⁵ Si se estuviera dispuesto a admitir que la fijación del precio tuvo un impacto sustancial sobre el precio – lo cual, presuntamente, no es el caso – la estimación mínimo cuadrática de la curva de demanda podría resultar mucho más fácil, porque esencialmente los precios habrían sido determinados en forma controlada. Sobre este punto, véase mi estudio de la demanda de aluminio, F.M. Fisher, *A Priori Information and Time Series Analysis* 93-117 (1962).

³⁶ Puede haber algunos problemas técnicos sobre la forma de estimar la ecuación (9) directamente mediante regresión múltiple o usar técnicas sofisticadas de ecuaciones simultáneas para estimar las ecuaciones (7) y (8) directamente, pero no nos detendremos en ellos.

³⁷ Esto sería contrastado típicamente mediante el llamado test de "Chow". Ver Fisher, *Tests of Equality Between Sets of Coefficients in Two Linear Regressions: An Expository Note*, 38 *Econometrica* 361-66 (1970). Este sería también el contraste utilizado para determinar si la regresión completa del ingreso vis-a-vis la aptitud fuera la misma para los hombres y mujeres en el ejemplo de discriminación que se analizó más arriba.

dirección pueden ser explicados – ya sea por los querellantes o los acusados – en términos de cambios en las condiciones de demanda y costo. Luego, si lo que está involucrado es una predicción a largo plazo, es posible que sea válido realizarla, pero es probable que no resulte útil. Como forma alternativa de encarar el análisis, tiene empeño el mérito de facilitar una comparación clara de las dos hipótesis involucradas.

C. *El Castigo como Disuasorio del Crimen*

El último tópico que deseo abordar es la utilización de estudios que persiguen examinar el efecto del castigo como un factor disuasorio general del crimen – es decir, como una disuasión a otras personas que las que son castigadas. Ya he mencionado los estudios sobre pena capital referidos por el Procurador General. Además, hay varios estudios de otras categorías de crímenes y tipos de castigo. Esta no es una oportunidad apropiada para discutirlos en gran detalle; tales discusiones pueden ser consultadas en otros lados.³⁸ Sin embargo, consideraré algunos motivos por los que estos estudios son insatisfactorios, ilustrando así puntos que son generalmente aplicables al uso del análisis de regresión múltiple.

A primera vista, el problema parece eminentemente apropiado al análisis de regresión. Casi todos los análisis de los datos en los que cambian los castigos demuestran que el crimen cambia en la dirección contraria. Los datos anuales sobre asesinatos cometidos en Estados Unidos (una *serie temporal*) indican que la cantidad de asesinatos aumentan en los años en los que no hay ejecuciones. Con respecto a otros crímenes, los datos de sección cruzada muestran que las jurisdicciones con castigos menos severos son aquellas con mayores tasas de criminalidad. Pareciera que existe en forma clara una correlación negativa entre severidad del castigo y tasa de criminalidad y que el problema es meramente el de evaluar la magnitud del efecto disuasorio.

Lamentablemente, si bien estoy de acuerdo con que en los datos aparece algo significativo, el problema de medición resulta ser muy severo. Lo cual es cierto por más de una razón. Primero, hay un problema debido a que carecemos de una buena teoría de lo que causa el crimen, y porque consiguientemente ignoramos qué otras variables cabría controlar para derivar una ecuación del crimen. Segundo, debemos controlar no solamente otras variables que aparecen en la misma ecuación sino además la presencia de relaciones adicionales entre estas variables y el crimen. Sumemos a esto el carácter dudoso de muchos datos y tenemos un problema serio.

Voy a comenzar considerando los estudios sobre la pena de muerte.³⁹ El estudio primario⁴⁰ utilizó series de tiempo de Estados Unidos en conjunto para los años 1933-1969. Se trata de una muestra de treinta y siete observaciones, aunque tuvieron que construirse los datos de algunas variables. Sin embargo, los resultados dependieron casi en su totalidad de los años posteriores a 1962. Lo cual, tal vez, no sea sorprendente; fue en esos años y a principios de los 1970s que varias jurisdicciones experimentaron con la abolición de la pena capital. Con todo, significa que hay una cantidad limitada de datos a ser usados para controlar otros efectos, a pesar del tamaño aparentemente grande de

³⁸ Véase, por ejemplo, *Deterrence and Incapacitation*, *supra* nota 1.

³⁹ Para una discusión y referencias más detalladas, véase Klein, Forst & Filatov, *supra* nota 1.

⁴⁰ Ehrlich, *The Deterrent Effect of Capital Punishment: A Question of Life and Death*, 65 Am. Econ. Rev. 397 (1975).

la muestra. Además, los mismos años coinciden con un florecimiento general del crimen, no sólo de los crímenes sujetos a la pena capital. Por consiguiente, no podemos estar seguros de que los resultados del estudio no dependan simplemente de fenómenos que son poco comprendidos sobre las causas del crimen.

Aquí hay lecciones que podemos aprender. Primero, cuando encaramos un estudio de regresión múltiple, se debería buscar establecer si los resultados dependen en forma crucial de algunos años escogidos o si sólo se trata de variaciones dentro de la muestra. Si es cierto que los resultados dependen de ciertos años, se debería tratar de decidir si hay alguna otra característica asociada a estos años que podría haber afectado los resultados. Segundo, y tal vez más importante, se debería tratar de determinar si se sabe lo suficiente sobre el fenómeno investigado (aquí, las causas del crimen) para estimarlo en términos del modelo seleccionado. De no ser así, probablemente habrá otras explicaciones plausibles de los resultados alcanzados.

También resulta que el estudio de la pena capital depende en forma bastante crucial de la forma de la ecuación utilizada. Hay una gran diferencia de los resultados según que la ecuación sea estimada en forma lineal o logarítmica.⁴¹ Naturalmente, si se tuviera una razón para creer que la forma correcta de la ecuación fuera una u otra, sencillamente se usaría esa forma. Pero no se conoce cuál es la forma “correcta”. Aquellos resultados que dependen del uso de una versión particular de la ecuación pueden no ser válidos; dependen de un supuesto no contrastado.⁴² Cuando hay que ejecutar a un ser humano, debería lograrse una máxima concentración mental. En tales cuestiones, los estudios sobre los que cabe apoyarse no deberían depender de un período muestral particular o de una elección particular de la especificación.

Muchos de los problemas que hay con la pena capital surgen por la naturaleza limitada de los datos disponibles. Un conjunto de experimentos que sería una alternativa obvia implicaría usar datos concernientes a distintos tipos de crimen y extraídos de distintas jurisdicciones, a efectos de conseguir una muestra amplia y suficiente variabilidad.⁴³ El problema es el siguiente.

Es obvio que hay razones no vinculadas al castigo por las cuales las tasas de criminalidad varían a través de las jurisdicciones. Luego, es necesario controlarlas. Algunas posibilidades son la tasa de desempleo, la proporción de población urbana, etc. La regresión múltiple podría hacerlo.

Lamentablemente, también hay razones por las que los niveles de castigo cambian a lo largo de las jurisdicciones. Una de las razones que se ha sugerido en la literatura tiene que ver con la tasa de criminalidad. Es sencillo ver por qué esto podría ser así. Las ju-

⁴¹ Véanse notas 10 & 20 *supra*.

⁴² Hay métodos para contrastar si una forma es mejor que otra. Sin embargo, a menudo es difícil extraer una conclusión de los resultados.

⁴³ Ehrlich también realizó un análisis de sección cruzada del asesinato, pero estoy menos familiarizado con éste que con su estudio de los crímenes no-capitales. Véase Ehrlich, *Participation in Illegitimate Activities: A Theoretical and Empirical Investigation*, 81 J. Pol. Econ. 521 (1973). Los siguientes comentarios han sido ampliados en Nagin, *General Deterrence: A Review of the Empirical Evidence*, in *Deterrence and Incapacitation*, *supra* nota 1, en 95, y Fisher & Nagin, *On the Feasibility of Identifying the Crime Function in a Simultaneous Model of Crime Rate and Sanction Levels*, *id.* en 361.

jurisdicciones con tasas de criminalidad más elevada pueden adoptar la política de actuar en forma severa. En forma alternativa – y esto fue sugerido por mucha literatura – las jurisdicciones con elevada tasa de criminalidad pueden tener una *sobrecarga* de sus dispositivos de castigo y pueden terminar tolerando más delitos relativamente comunes que otras jurisdicciones con tasas de criminalidad más bajas. En todo caso, hay una posibilidad de que el cambio en el nivel de los castigos sea explicado, al menos en parte, por cambios en la tasa de criminalidad. En tales circunstancias – como en parte del ejemplo de oferta y demanda enunciado más arriba – el problema no es que se deba controlar otras variables, sino que se tiene que controlar por la presencia de otra ecuación. Para apreciar el tipo de problemas que aparece, tomemos el siguiente ejemplo enormemente simplificado.

Supongamos, por el momento, que lo único que afecta la tasa de criminalidad es el castigo. Asumiendo una relación lineal por conveniencia, la ecuación de la tasa de criminalidad a ser estimada podría escribirse de la siguiente forma:

$$(10) \quad C = a + bS + u$$

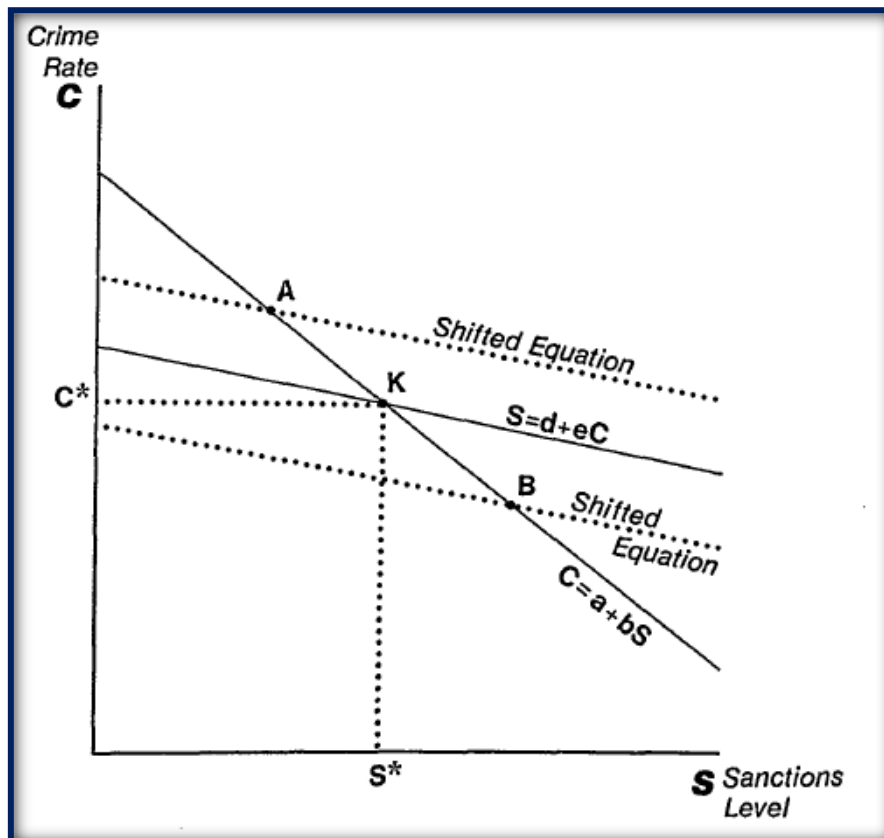


Figura 3

Aquí, C es la medida de la tasa de criminalidad, y S es una medida de castigo o nivel de la sanción. El coeficiente b representaría el efecto disuasorio de las crecientes sanciones.

Supongan, empero, que las sanciones también dependen de la tasa de criminalidad y sólo de ella. Entonces, la ecuación que indica cómo son determinados los niveles de sanción puede ser escrita (suponiendo de nuevo una forma lineal):

$$(11) \quad S = d + eC + v.$$

En estas ecuaciones, u y v son perturbaciones aleatorias.

Dada la interrelación entre ambas ecuaciones, no se podrá estimar efectivamente la ecuación del crimen (10) mediante una regresión por mínimos cuadrados. El supuesto fundamental de mínimos cuadrados es que el término de perturbación aleatoria actúa independientemente de la variable independiente. Todas las propiedades de los mínimos cuadrados dependen de esto. En el presente caso, eso requiere que u y S estén no correlacionados. Lo cual no puede ser el caso, porque el mismo modelo (como en el modelo de oferta y demanda mencionado más arriba) implica que esto no sea así. Un desplazamiento hacia arriba de u , con arreglo a la ecuación (10), implicará un traslado hacia arriba de la tasa de criminalidad C . Pero un cambio positivo de la tasa de criminalidad C , según la ecuación (11), dará lugar a un cambio del nivel de las sanciones S . Luego, **cambios en u no pueden ser independientes de los cambios de S y esto hará fallar los mínimos cuadrados.** (Esto también puede visualizarse resolviendo las ecuaciones (10) y (11) en términos de C y S a fin de obtener la forma reducida del sistema, como fue hecho en el ejemplo de oferta y demanda.)

Pero el problema es aún peor. Para verlo, ignoremos las perturbaciones aleatorias por un momento, y supongamos que las ecuaciones (10) y (11) fueran exactas. Véase la Figura 3. En tal situación, la tasa de criminalidad y el nivel de sanciones sería completamente determinado por la solución simultánea de las dos ecuaciones no estocásticas – produciendo la intersección de ambas líneas en K . (La semejanza con un gráfico de oferta y demanda no es accidental.) Si así fuera el caso, el único punto que observaríamos correspondería al nivel de sanciones indicado como S^* y la tasa de criminalidad indicada como C^* . Pero de ser éste el único punto observado, no habría forma alguna de recuperar las ecuaciones (10) y (11). En términos del gráfico, no podríamos distinguir la verdadera función del crimen (la línea más empinada) de la función de las sanciones (la línea menos empinada) o, incluso, de cualquier otra línea que pase a través del mismo punto; cada línea podría variar, de infinitas formas, alrededor del punto K .

Aún incorporando de nuevo las perturbaciones, no llegaríamos a ningún lado. El efecto de reintroducir las perturbaciones aleatorias sería producir un **cluster** de puntos en torno a la intersección marcada en el gráfico, pero, de nuevo, sería imposible recuperar las dos líneas subyacentes que generaron este **cluster** o distinguir de qué línea se trata si pudiéramos recuperarlas. **En tal circunstancia, las ecuaciones del crimen y de las sanciones se dice que no son identificables.**

Este problema, conocido como de **identificación**, es un tema muy estudiado en econometría.⁴⁴ Elegí en forma deliberada un caso extremo. Lamentablemente, el problema de identificación continúa estando en los estudios de disuasión aún cuando sean relajados estos supuestos extremos.

Supóngase, por ejemplo, que hay alguna variable que traslada los niveles de sanción a lo largo de las jurisdicciones pero que no afecta la tasa de criminalidad. Esto implicaría que hay otra variable significativa en la ecuación (11) que no era también una variable de la versión expandida de la ecuación (10). Dejando la ecuación (10) tal como está, el

⁴⁴ Véase F.M. Fisher, *The Identification Problem in Econometrics* (1966) (re-impreso 1976).

efecto sería trasladar la ecuación de las sanciones de la Figura 3 hacia arriba y abajo. (Esto se ilustra mediante líneas a rayas paralelas a la línea sólida que corresponde a la ecuación de las sanciones en la Figura 3 e indicada como “*shifted equation*” [ecuación trasladada]). Si esto sucediera, observaríamos no solamente una intersección de la ecuación de las sanciones y la ecuación del crimen, sino varias intersecciones, como los puntos A y B por ejemplo. Todos estos puntos yacerían en la ecuación del crimen, y en realidad, a medida que la ecuación de las sanciones cambiara hacia adelante y atrás por la presencia de la variable adicional, los puntos de intersección trazarian la ecuación del crimen.

En esa situación, como sugiere el diagrama, hay una técnica para recuperar la ecuación del crimen a partir de los datos. Pero esa técnica no es la de regresión por mínimos cuadrados, porque la correlación entre el término de perturbación y la variable independiente de la ecuación (10) daría lugar a resultados inválidos. Más aún, no será posible recuperar la propia ecuación de las sanciones.

A raíz del problema de identificación es necesario hallar una variable que traslade una ecuación del modelo, pero no la ecuación a ser identificada. Empero, no sólo es mala práctica tratar de hallar tales variables a partir de los datos, sino que es literalmente imposible. Ninguna manipulación de los datos generados por el modelo podrá revelar tales variables; la selección de tales variables debe ***surgir a partir de una teoría previa.***

Es fácil ver en la Figura 3 por qué debe ser así. Si hay una variable que traslada la ecuación de las sanciones pero no la ecuación del crimen, entonces los puntos observados serán tales como los puntos A, B y K del diagrama. Pero este patrón de intersecciones también podría ser producido por una variable que traslada la ecuación del crimen pero no la ecuación de las sanciones. Más en general, podría ser producido por traslado de ambas ecuaciones. ***Pero sólo si sabemos a partir de consideraciones teóricas, no generadas por los datos, que se trata de la ecuación de sanciones la que se traslada, podremos estar seguros de que lograremos identificar la ecuación del crimen.***

En muchas situaciones pueden hallarse estas consideraciones teóricas (Por ejemplo, el ingreso del consumidor actúa sobre la curva de demanda, pero no la de oferta; los costos factoriales afectan la oferta pero no la demanda.) Pero en el caso presente no sucede lo mismo. Si bien hay un gran número de variables que podrían intervenir en la ecuación de las sanciones, resulta difícil, si no imposible, pensar que semejante variable no intervendría también en la ecuación del crimen.⁴⁵ Los estudios existentes han buscado sortear el problema suponiendo que variables tales como el desempleo influyen sobre las sanciones pero no sobre el crimen. Esto es simplemente erróneo. En el estado presente de nuestro conocimiento, no sabemos lo suficiente sobre la estructura del sistema generador de las observaciones como para que sea válidamente usado para estimar la ecuación del crimen.

⁴⁵ Por otra parte, es difícil pensar en variables que intervengan en la ecuación del crimen pero no tengan influencia directa sobre la elección de sanciones. ***El desempleo, por ejemplo, es mucho más probable que influya sobre la tasa de criminalidad que sobre las sanciones.*** Otros ejemplos podrían incluir medidas de disparidad del ingreso o los gastos en sistemas de seguridad. Si tales variables influyen realmente sobre la tasa de criminalidad, pero no sobre las sanciones, al incluirlas en la ecuación del crimen se trasladaría esa función con respecto a la ecuación de las sanciones. Los puntos de intersección trazados yacerían todos en la ecuación de las sanciones, que sería entonces identificable y podría ser estimada (aunque no por mínimos cuadrados).

De este problema surgen algunas implicancias generales para el uso del análisis de regresión. Primero, resulta importante ser muy cuidadoso no solamente en controlar las variables adicionales, sino también la posibilidad de tener que controlar también la existencia de relaciones adicionales entre las variables dependiente y las independientes. Si las hubiera mínimos cuadrados no será un estimador apropiado, e incluso hasta es posible que no exista ningún estimador apropiado (si bien esto no es lo común). Segundo, si existe otra ecuación involucrada, hay que hallar cómo hizo el experto su estimación. Si lo que hizo fue explorar los datos mediante regresión múltiple y luego, habiendo decidido el modelo a emplear, lo alteró mediante otra técnica de estimación, los resultados serán hartos sospechosos.⁴⁶

Finalmente, cabría estar seguro de que el modelo usado ha sido elaborado en base a hipótesis sólidas basadas en consideraciones teóricas generadas desde afuera del modelo. *Si la regresión múltiple y las técnicas econométricas asociadas son instrumentos poderosos para analizar los datos, su uso apropiado presupone una teoría subyacente de la estructura que los genera.* Si bien algunas hipótesis sobre esa estructura pueden ser contrastadas mediante estas herramientas, la propia teoría no puede ser descubierta mediante corridas de computadora y experimentando con los datos. Luego, el experto que realiza el estudio debe entender no sólo el uso apropiado de las herramientas estadísticas, sino además debe conocer y aprender algo sobre el fenómeno y las hipótesis investigadas.

CONCLUSIÓN

El análisis de regresión múltiple puede desempeñar un rol vital en procedimientos legales. Usado de forma adecuada, se trata de un método exacto y confiable para determinar las relaciones entre dos o más variables, y puede resultar una herramienta indispensable para resolver las disputas sobre los hechos. Para que esto suceda, empero, la regresión múltiple debe ser mejor comprendida por la comunidad legal; en particular, es preciso que haya una comprensión tanto de su potencial como de los límites de la técnica.

No es necesario que los abogados entiendan la mecánica de la regresión múltiple en términos de lo que sucede adentro del computador. Sin embargo, se requiere que entiendan el modelo de regresión y los supuestos utilizados en cualquier estudio de regresión, cómo los resultados de la regresión juegan con la hipótesis que está siendo contrastada, y cómo los resultados permiten distinguir esa hipótesis particular de otras. El experto responsable del análisis debe explicar todo esto al abogado que lo consultó, y un experto que no pueda explicarlas probablemente sea marginado en un examen cruzado.

⁴⁶ Consideren el siguiente procedimiento, que es muy habitual. Como la regresión múltiple es simple de llevar a cabo, uno experimenta mediante regresión múltiple hasta que se llega a una versión de la ecuación estimada que corresponde a lo que uno prefiere. Entonces se reestima la ecuación mediante una técnica apropiada de ecuaciones simultáneas. Si los resultados difieren demasiado de la versión por mínimos cuadrados se continúa explorando. Ésta no es una forma de producir resultados consistentes.