Introducción al análisis de regresión

Alan O. Sykes

Alan O. Sykes, *An Introduction to Regression Analysis* (Coase-Sandor Institute for Law & Economics Working Paper No. 20, 1993)

El análisis de regresión es una herramienta estadística para investigar relaciones entre variables. Por lo general, el investigador trata de determinar el efecto causal de una variable sobre otra: el efecto de un aumento del precio sobre la demanda, por ejemplo, o el efecto de los cambios en la oferta monetaria sobre la tasa de inflación. Para explorar estas cuestiones, el investigador reúne datos sobre las variables subyacentes de interés y emplea la regresión para estimar el efecto cuantitativo de las variables causales sobre la variable que afectan. El investigador también suele evaluar la "significación estadística" de las relaciones estimadas, es decir, el grado de confianza de que la relación verdadera esté próxima a la relación estimada.

Las técnicas de regresión han sido durante mucho tiempo fundamentales en el campo de la estadística económica (*econometría*). Cada vez son más importantes para los abogados y los encargados de la formulación de políticas legales. El análisis regresión fue ofrecido como prueba de responsabilidad bajo el Título VII de la Ley de Derechos Civiles de 1964,[1] como evidencia de sesgo racial en litigios por pena de muerte,[2] como evidencia de daños en acciones contractuales,[3] como evidencia de violaciones bajo la Ley de Derechos Electorales,[4] y como evidencia de daños y perjuicios en litigios antimonopolios,[5] entre otras cosas.

En esta conferencia, daré una visión general de las técnicas más básicas del análisis de regresión: cómo funcionan, qué suponen y cómo pueden fallar cuando no se cumplen los supuestos clave. Para que la discusión sea más concreta, emplearé una serie de ilustraciones que implicarán un análisis hipotético de los factores que determinan los ingresos individuales en el mercado de trabajo. Las ilustraciones tendrán un sabor más legal en la última parte de la conferencia, donde se incorporará la posibilidad de que los ingresos estén inadmisiblemente afectados por el género en violación de las leyes federales de derechos civiles.[6] Quiero enfatizar que esta conferencia no es un tratamiento integral de las cuestiones estadísticas que surgen en los litigios del Título VII y que la discusión sobre la discriminación de género es simplemente un vehículo para exponer ciertos aspectos de la técnica de regresión.[7] También, por necesidad, hay muchos temas

---

[1] Véase, por ejemplo, *Bazemore v. Friday, 478 US 385, 400 (1986).*
[2] Véase, por ejemplo, *McClesky v. Kemp, 481 U.S. 279 (1987*).
[3] Véase, por ejemplo, *Cotton Brothers Baking Co. v. Industrial Risk Insurers, 941 F.2d 380 (5th Cir. 1991)*.
[4] Véase, por ejemplo, *Thornburgh v. Gingles, 478 U.S. 30 (1986)*.
[5] Véase, por ejemplo, *Sprayrite Service Corp. v. Monsanto Co., 684 F.2d 1226 (7th Cir. 1982)*.
[6] Véase *42 U.S.C. §2000e-2 (1988),* modificada.
[7] Los lectores con un interés particular en el uso del análisis de regresión según el Título VII tal vez deseen consultar las siguientes referencias: Thomas Campbell, *Regression Analysis in Title VII Cases—Minimum Standards, Comparable Worth, and Other Issues Where Law and Statistics Meet*, 36 Stan. L. Rev. 1299 (1984); Catherine Connolly, *The Use of Multiple Regression Analysis in Employment Discrimination Cases*, 10 Population Res. and Pol. Rev. 117 (1991); Michael Finkelstein, *The Judicial Reception of Multiple Regression Studies in Race and Sex*

importantes que omitiré, incluyendo los modelos de ecuaciones simultáneas y los mínimos cuadrados generalizados. La conferencia se limita a los supuestos, la mecánica y las dificultades usuales con la regresión de una ecuación única con mínimos cuadrados ordinarios.

1 ¿Qué es una regresión?

A efectos de ilustración, supongamos que deseamos identificar y cuantificar los factores que determinan los ingresos en el mercado de trabajo. Un momento de reflexión sugiere una miríada de factores que podrían estar asociados con las variaciones en los ingresos entre los individuos - ocupación, edad, experiencia, logros educativos, motivación y capacidad innata vienen a la mente, tal vez junto con factores como raza y género que pueden ser de interés particular de los abogados. Por el momento, vamos a restringir la atención a un solo factor - que llamamos educación. El análisis de regresión con una sola variable explicativa se denomina "regresión simple".

A. *Regresión simple*

En realidad, cualquier esfuerzo para cuantificar los efectos de la educación sobre los ingresos sin una atención cuidadosa a los otros factores que los afectan podría crear serias dificultades estadísticas (denominadas "sesgo de variables omitidas"), las cuales analizaré más adelante. Pero por ahora vamos a ignorar este problema. También suponemos, de nuevo de forma muy poco realista, que la "educación" se puede medir con un solo atributo: los años de escolaridad. Suprimimos así el hecho de que un número dado de años en la escuela puede representar programas académicos muy diversos.

Al inicio de cualquier estudio de regresión, se formula alguna hipótesis sobre la relación entre las variables de interés, aquí, la educación y los ingresos. La experiencia común sugiere que las personas mejor educadas tienden a ganar más dinero. Además, sugiere que la relación causal probablemente va de la educación a los ingresos y no al revés. *Por lo tanto, la hipótesis tentativa es que niveles más altos de educación causan mayores niveles de ingresos, siendo iguales las demás cosas*.

Para investigar esta hipótesis, imaginen que recopilamos datos sobre educación e ingresos de varias personas. Sea E la educación en años de escolaridad de cada individuo, e I los ingresos de ese individuo en dólares por año. Podemos graficar esta información para todos los individuos de la muestra usando un

---

Discrimination Cases, 80 Colum. L. Rev. 737 (1980); y Franklin M. Fisher, *Multiple Regression in Legal Proceedings*, 80 Colum. L. Rev. 702 (1980), p. 721 - 25.

diagrama bidimensional, llamado convencionalmente diagrama de *dispersión*. Cada punto en el diagrama representa un individuo de la muestra.

El diagrama de hecho sugiere que valores más altos de E tienden a dar valores más altos de I, pero la relación no es perfecta - parece que el conocimiento de E no es suficiente para una predicción completamente precisa sobre I.[8] Podemos deducir entonces que el efecto de la educación sobre los ingresos difiere entre los individuos, o que otros factores distintos de la educación influyen sobre los ingresos. El análisis de regresión suele adoptar esta última explicación.[9] Así, a continuación de la discusión sobre el sesgo de las variables omitidas, ahora hacemos la hipótesis de que los ingresos de cada individuo se determinan por la educación y por un agregado de factores omitidos que denominamos "*ruido*".

Para refinar aún más la hipótesis, es natural suponer que la gente en la fuerza de trabajo sin educación, ganan pese a todo alguna cantidad positive de dinero, y que la educación aumenta los ingresos por encima de esta base. También suponemos que la educación afecta al ingreso de manera "*lineal*" - es decir, que cada año adicional de escolaridad agrega la misma cantidad al ingreso. Este supuesto de linealidad es común en los estudios de regresión, pero no es esencial para la aplicación de la técnica y puede ser relajado cuando el investigador tiene razones para suponer *a priori* que la relación en cuestión es no lineal.[10]

Entonces, la relación hipotética entre educación e ingreso puede ser escrita como sigue:

$$I = \alpha + \beta E + \varepsilon$$

donde

α = una cantidad constante (lo que se gana con cero educación);

β = el efecto en pesos de un año adicional de escolaridad sobre el ingreso, en la hipótesis de que es positivo; y

ε = el término "ruido" que refleja otros factores que influyen sobre el ingreso.

---

[8] Más precisamente, lo que se puede deducir del diagrama es que si el conocimiento de E fuera suficiente para predecir I perfectamente, entonces la relación entre ambos sería compleja, no lineal. Debido a que no tenemos razones para sospechar que la verdadera relación entre la educación y los ingresos sea de esa forma, es más probable que concluyamos que el conocimiento de E no es suficiente para predecir I perfectamente.

[9] La posibilidad alternativa - que la relación entre dos variables sea inestable - corresponde al problema de coeficientes "aleatorios" o "variables en el tiempo" y plantea problemas estadísticos algo diferentes. Véase, por ejemplo, H. Theil, *Principles of Econometrics* 622 - 27 (1971); G. Chow, *Econometrics* 320 - 47 (1983).

[10] Cuando se cree que están presentes relaciones no lineales, los investigadores suelen tratar de modelarlas de manera que les permita ser transformadas en relaciones lineales. Por ejemplo, la relación $y = cx^a$ puede ser transformada en la relación lineal $\log y = \log c + a \log x$. La razón para modelar las relaciones no lineales de esta forma es que la estimación de regresiones lineales es mucho más simple y sus propiedades estadísticas son mejor conocidas. Sin embargo, cuando este enfoque es inviable, se han desarrollado técnicas para la estimación de regresiones no lineales. Véase, por ejemplo, G. Chow, supra nota 9, 220 - 51.

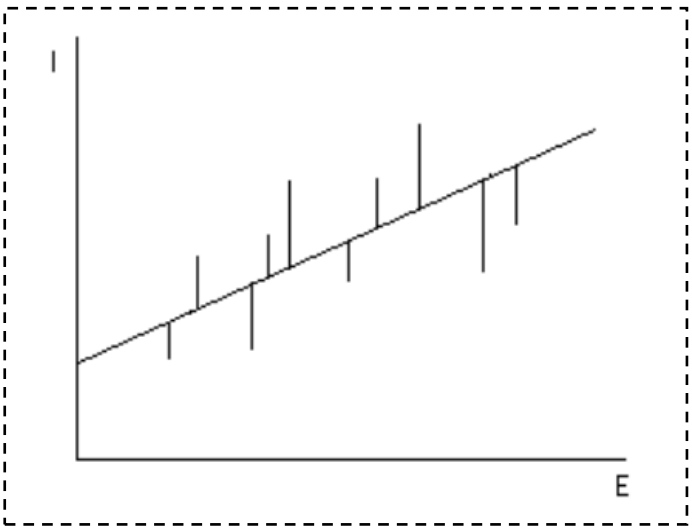
A la variable I se la denomina variable *dependiente* o *endógena*; E es denominada variable *independiente*, *explicativa* o *exógena*; α es el *término constante* y β el *coeficiente* de la variable E.

Recuérdese lo que es observable y lo que no. El conjunto de datos contiene observaciones de I y E. El componente de ruido e está compuesto de factores que no son observables, o al menos no observados. Los parámetros α y β también son inobservables. La tarea del análisis de regresión es producir una *estimación* de estos dos parámetros, basada en la información contenida en el conjunto de datos y, como se verá, sobre algunos supuestos sobre las características de ε.

Para entender cómo se generan las estimaciones de parámetros, obsérvese que si ignoramos el término de ruido ε, la ecuación anterior de relación entre I y E es la ecuación de una recta - una recta con una *ordenada al origen* = α en el eje vertical y una *pendiente* = β. Volviendo al diagrama de dispersión, la relación hipotética implica por tanto que en algún punto del diagrama se puede encontrar una recta con la ecuación I = α + βE. La tarea de estimar α y β es equivalente a la tarea de estimar dónde está ubicada esta línea.

¿Cuál es la mejor estimación con respecto a la ubicación de esta línea? La respuesta depende en parte de lo que pensemos acerca de la naturaleza del término de ruido ε. Si creyéramos que ε es en general un número negativo grande, por ejemplo, querríamos escoger una línea que quede por encima de la mayoría o de todos nuestros puntos - la lógica es que si ε es negativo, el valor verdadero de I (que observamos) dado por I= α+ βE + ε, será menor que el valor de I en la recta I = α + βE. Del mismo modo, si creyéramos que ε es sistemáticamente positivo, sería apropiada una línea situada por debajo de la mayoría de los puntos de datos. El análisis de regresión supone, sin embargo, que el término de ruido no tiene tal propiedad sistemática, sino que es en promedio igual a cero - formularé los supuestos sobre el término de ruido más precisamente en un momento. El supuesto de que el término de ruido es cero en general sugiere una estimación de la línea que pasa aproximadamente en medio de los datos, con algunas observaciones por debajo y otras observaciones por encima.

Pero hay muchas líneas de este tipo, y queda por elegir una línea en particular. El análisis de regresión lo hace adoptando un criterio que se relaciona con el término de ruido estimado o "error" de cada observación. Para ser exactos, defínase el *error estimado* de cada observación como la distancia vertical entre el valor de I a lo largo de la línea estimada I = α + βE (generada colocando el valor real de E en esta ecuación) y el valor verdadero de I para la misma observación. Superponiendo una línea candidata en el diagrama de dispersión, los errores estimados para cada



observación se pueden ver como en la figura.

Con cada línea posible que se pueda superponer a los datos, resultará un conjunto diferente de errores estimados. El análisis de regresión elige entonces entre las líneas posibles seleccionando aquella para la que la suma de los cuadrados de los errores estimados es mínima. Éste es denominado *criterio del mínimo de la suma de errores cuadráticos (mínimo de SEC)*. La ordenada al origen de la línea elegida por este criterio proporciona el estimador de α, y su pendiente proporciona el estimador de β.

No es obvio por qué debamos elegir nuestra línea usando el criterio del mínimo de SEC. Fácilmente se puede imaginar otros criterios que podrían ser utilizados (minimizar la suma de errores en valor absoluto,[11] por ejemplo). Una virtud del criterio SEC es que es muy fácil de emplear computacionalmente. Cuando se expresa matemáticamente la suma de los errores cuadráticos y se emplean técnicas de cálculo para determinar los valores de α y β que la minimizan, se obtienen expresiones de α y β que son fáciles de evaluar con un computador usando sólo los valores observados de E e I de la muestra de datos.[12] Pero la conveniencia computacional no es la única virtud del criterio de mínima SEC, sino que también tiene algunas propiedades estadísticas atractivas bajo supuestos plausibles sobre el término de ruido. Estas propiedades serán discutidas en un momento, después de introducir el concepto de regresión múltiple.

B. *Regresión múltiple*

Claramente, los ingresos se ven afectados por una variedad de factores además de los años de escolaridad, factores que se incorporaron en el término de ruido en el modelo de regresión simple anterior. "Regresión múltiple" es una técnica que permite que entren factores adicionales en el análisis por separado para que el efecto de cada uno pueda ser estimado. Es valiosa para cuantificar el impacto de varias influencias simultáneas sobre una sola variable dependiente. Además, debido al sesgo de variables omitidas con regresión simple, la regresión múltiple es a menudo esencial incluso cuando el investigador sólo está interesado en los efectos de una de las variables independientes.

A fines ilustrativos, considérese la introducción en el análisis de los ingresos de una segunda variable independiente llamada *experiencia*. Manteniendo constante el nivel de educación, esperamos que alguien que ha estado trabajando durante más tiempo gane más. Sea X el número de años de experiencia en la fuerza laboral y, como en el caso de la educación, supongamos que tiene un efecto lineal sobre los ingresos que es estable entre los individuos. El modelo modificado puede escribirse:

$$I = \alpha + \beta E + \gamma X + \varepsilon$$

[11] Debería ser obvio por qué simplemente minimizar la suma de los errores no es un criterio atractivo - los grandes errores negativos y positivos se anularían, por lo que esta suma podría ser mínima, aunque la línea seleccionada se ajuste muy mal a los datos.

[12] La derivación es tan simple en el caso de una variable explicativa que vale la pena incluirla aquí: Continuando con el ejemplo del texto, imaginemos que tenemos datos sobre educación e ingresos para un número de individuos, indexados por j. El valor real de los ingresos del j-ésimo individuo es I_j, y su valor estimado para cualquier línea con la intersección α y pendiente β será $\alpha + \beta E_j$. El error estimado es, por tanto, $I_j - \alpha - \beta E_j$. La suma de errores cuadrados es entonces $\sum_j (I_j - \alpha - \beta E_j)^2$. Minimizar esta suma con respecto a α requiere que su derivada con respecto a α sea cero, o $-2\Sigma_j(I_j - \alpha - \beta E_j) = 0$. Minimizar con respecto a β también requiere $-2\Sigma_j E_i(I_j - \alpha - \beta E_j) = 0$. Ahora tenemos dos ecuaciones en dos incógnitas que pueden ser resueltas para α y β.

donde esperamos que el parámetro γ sea positivo.

La tarea de estimar los parámetros α, β y γ es conceptualmente idéntica a la tarea anterior de estimar sólo α y β. La diferencia es que ya no podemos pensar en la regresión como la elección de una línea en un diagrama bidimensional; con dos variables explicativas necesitamos tres dimensiones, y en lugar de estimar una línea estamos estimando un *plano*. El análisis de regresión múltiple seleccionará un plano de modo que la suma de errores cuadráticos -el error aquí es la distancia vertical entre el valor real de I y el plano estimado- sea mínimo. El punto de intercepción de ese plano con el eje I (donde E y X son cero) implica el término constante α, su pendiente en la dimensión educativa implica el coeficiente β, y su pendiente en la dimensión experiencia implica el coeficiente γ.

El análisis de regresión múltiple puede de hecho tratar con un número arbitrariamente grande de variables explicativas. Aunque la gente carece de la capacidad de visualizar en más de tres dimensiones, la matemática no. Con n variables explicativas, el análisis de regresión múltiple estimará la ecuación de un *hiperplano* en el espacio de *n* dimensiones de modo que la suma de errores cuadrados haya sido minimizada. El punto de intercepción cuando todas las variables son nulas implica el término constante, y su pendiente en cada dimensión implica uno de los coeficientes de regresión. Como en el caso de la regresión simple, el criterio SEC es bastante cómodo computacionalmente. Las fórmulas para los parámetros α, β, γ ... se pueden deducir y evaluar fácilmente en un computador, utilizando de nuevo sólo los valores observados de las variables dependientes e independientes.[13]

La interpretación de los coeficientes estimados en una regresión múltiple merece un breve comentario. En el modelo $I = \alpha + \beta E + \gamma X + \varepsilon$, α capta lo que gana un individuo sin educación ni experiencia, β capta el efecto sobre el ingreso de un año de educación, y γ capta el efecto sobre el ingreso de un año de experiencia. Para decirlo de modo diferente, β es una estimación del efecto de un año de educación sobre el ingreso, manteniendo la experiencia constante. Asimismo, γ es el efecto estimado de un año de experiencia sobre el ingreso, manteniendo la educación constante.

2 *Supuestos Básicos y Propiedades Estadísticas de la Regresión*

Como se indicó, el uso del criterio de mínima SEC puede ser defendido por dos razones: su conveniencia computacional, y sus propiedades estadísticas deseables. Ahora consideraremos estas propiedades y los supuestos necesarios para asegurarlas.[14]

Siguiendo con nuestra ilustración, la hipótesis es que los ingresos en el "mundo real" son determinados de acuerdo con la ecuación $I = \alpha + \beta E + \gamma X + \varepsilon$ - existen los valores verdaderos de α, β, y γ, y deseamos saber cuáles son. Sin embargo, debido al término de ruido ε, sólo podemos estimar estos parámetros.

---

[13] La derivación se puede encontrar en cualquier texto estándar de econometría. Véase, por ejemplo, E. Hanushek y J. Jackson, *Statistical Methods for Social Scientists* 110 - 16 (1977); J. Johnston, *Econometric Methods* 122 - 32 (2ª edición, 1972).

[14] Una discusión accesible y más extensa de los supuestos clave de regresión se puede encontrar en Fisher, *supra* nota 7.

Podemos pensar en el término de ruido ε como una variable aleatoria, extraída por la naturaleza a partir de alguna distribución de probabilidad - la gente obtiene educación y acumula experiencia de trabajo, luego la naturaleza genera un número aleatorio para cada individuo, llamado ε, que aumenta o disminuye los ingresos en consecuencia. Una vez que consideramos el término de ruido como una variable aleatoria, queda claro que las estimaciones de α, β y γ (a diferencia de sus valores reales) también serán variables aleatorias, porque las estimaciones generadas por el criterio SEC dependerán del valor particular de ε extraído por naturaleza para cada individuo en el conjunto de datos. Del mismo modo, debido a que existe una distribución de probabilidad a partir de la cual se extrae cada ε, también debe existir una distribución de probabilidad a partir de la cual se extrae cada estimación de parámetros, siendo la última distribución una función de las distribuciones anteriores. Las propiedades estadísticas atractivas de la regresión se refieren a la relación entre la distribución de probabilidad de los parámetros estimados y los valores reales de esos parámetros.

Comenzamos con algunas definiciones. El criterio de mínima SEC se denomina *estimador*. Los criterios alternativos para generar estimaciones de parámetros (como minimizar la suma de errores en valor absoluto) también son estimadores.

Cada parámetro estimado que produce un estimador, como se ha indicado, puede ser visto como una variable aleatoria extraída de alguna distribución de probabilidad. Si la media de esa distribución de probabilidad es igual al valor verdadero del parámetro que estamos tratando de estimar, entonces el estimador se dice ser *insesgado*. En otras palabras, para volver a nuestra ilustración, imagínense crear una secuencia de conjuntos de datos que contengan a los mismos individuos con los mismos valores de educación y experiencia, diferenciándose sólo porque la naturaleza extrae un ε diferente para cada individuo para cada conjunto de datos. Imaginen además que recalculamos nuestras estimaciones de parámetros para cada conjunto de datos, generando así una gama de estimaciones para cada parámetro α, β, y γ. *Si el estimador es insesgado, hallaríamos que en promedio recuperaríamos el valor verdadero de cada parámetro*.

Un estimador se dice *consistente* si aprovecha los datos adicionales para generar estimaciones más precisas. Más exactamente, un estimador consistente da lugar a estimaciones que convergen al verdadero valor del parámetro subyacente a medida que el tamaño de la muestra se hace más y más grande. Así, la distribución de probabilidad del estimador de cualquier parámetro tiene menor varianza[15] a medida que aumenta el tamaño de la muestra y en el límite (tamaño de muestra infinito) el estimador será igual al valor real.

También es de interés la varianza de un estimador para una muestra de tamaño *dado*. En particular, vamos a restringir la atención a estimadores insesgados. Entonces, es claramente deseable una menor varianza en la distribución de probabilidad del estima-

[15] *Varianza* es una medida de dispersión de la distribución de probabilidad de una variable aleatoria. Considérese dos variables aleatorias con la misma media (mismo valor medio). Si una de ellas tiene una distribución con mayor varianza, entonces, en términos generales, la probabilidad de que la variable asuma un valor alejado de la media es mayor.

dor[16] - se reduce la probabilidad de un estimador que difiera mucho del verdadero valor del parámetro subyacente. Al comparar diferentes estimadores insesgados, el de menor varianza es denominado *eficiente* u *óptimo*.

Bajo ciertos supuestos, el criterio de mínima SEC tiene las características de ausencia de sesgo, consistencia y eficiencia: a continuación se presentan estos supuestos y sus consecuencias:

(1) Si el término de ruido de cada observación, ε, se obtiene de una distribución que tiene una media nula, entonces el criterio de mínima suma de errores cuadráticos SEC da lugar a estimaciones insesgadas y consistentes.

Es decir, podemos imaginar que para cada observación de la muestra, la naturaleza extrae un término de ruido de una distribución de probabilidad diferente. Siempre y cuando cada una de estas distribuciones tenga una media cero (incluso si las distribuciones no son las mismas), el criterio de mínima SEC es insesgado y consistente.[17] Este supuesto es lógicamente suficiente para asegurar que se verifique otra condición - a saber, que cada variable explicativa del modelo no está correlacionada con el valor esperado del término de ruido.[18] Esto resultará importante más adelante.

(2) Si las distribuciones de las que se extraen los términos de ruido para cada observación tienen la misma varianza y los términos de ruido son estadísticamente independientes entre sí (de modo que si hay un término de ruido positivo para una observación, por ejemplo, no hay razón para esperar un término de ruido positivo o negativo para cualquier otra observación), entonces el criterio de la suma de errores cuadráticos nos proporciona las mejores o más eficientes estimaciones disponibles de cualquier estimador lineal (definido éste como un estimador que calcula las estimaciones de los parámetros como función lineal del término de ruido, que es lo que hace el criterio SEC).[19]

Si se violan los supuestos (2), el criterio de SEC sigue siendo insesgado y consistente, pero es posible reducir la varianza del estimador tomando en cuenta lo que sabemos acerca del término de ruido. Por ejemplo, si sabemos que la varianza de la distribución a partir de la cual se extrae el término de ruido es mayor para ciertas observaciones, entonces es *probable* que el tamaño del término de ruido para esas observaciones sea mayor. Y, debido a que el ruido es mayor, podríamos querer dar a esas observaciones menos peso en nuestro análisis. El procedimiento estadístico para tratar este tipo de problema se denomina *mínimos cuadrados generalizados*, lo cual está fuera del alcance de esta conferencia.[20]

---

[16] Una variación más baja, por sí misma no es necesariamente una propiedad atractiva de un estimador. Por ejemplo, podríamos emplear un estimador para β de la forma "β = 17" independientemente de la información del conjunto de datos. Este estimador tiene varianza cero.
[17] Véase, por ejemplo, P. Kennedy, A Guide to Econometrics 42 - 44 (2ª edición, 1985).
[18] Si el valor esperado del término de ruido es siempre cero independientemente de los valores de las variables explicativas para la observación a la que está asociado el término de ruido, entonces por definición el término de ruido no puede estar correlacionado con ninguna variable explicativa.
[19] P. ej, id. en 44; J. Johnston, supra nota 13, en 126-27.
[20] Véase, por ejemplo, id. en 208 - 66.

3. *Ilustración - Discriminación sobre la Base del Género*

Para ilustrar las ideas hasta este punto, así como para sugerir cómo el análisis de regresión puede tener aplicaciones útiles en un procedimiento legal, imagínese una firma hipotética que ha sido demandada por discriminación salarial en base al género. Para investigar estas acusaciones, se han recopilado datos de todos los empleados de la empresa. Las preguntas a contestar son (a) si la discriminación está ocurriendo (responsabilidad), y (b) cuáles son sus consecuencias (daños). Las abordaremos utilizando una versión modificada del modelo de ingresos desarrollado en la sección 1.

La utilidad de la *regresión múltiple* aquí debe ser intuitivamente evidente. Supongamos, por ejemplo, que según los datos, las mujeres en la empresa, en promedio, ganan menos que los hombres. ¿Es esto suficiente para establecer una discriminación procesable? La respuesta es no si la diferencia surge porque las mujeres en esta empresa están menos bien educadas, por ejemplo (y, por tanto, por inferencia son menos productivas), o porque son menos experimentadas.[21] *En resumen, la cuestión jurídica es si las mujeres ganan menos después de tomar en cuenta todos los factores permisibles que la empresa pueda considerar*.

Para generar los datos de esta ilustración, postulo un "mundo real" hipotético en el cual los ingresos se determinan mediante la ecuación (1):

(1) Ingresos = 5000 + 1000 • Escuela + 50 • Aptitud + 300 • Experiencia - 2000 • Género + Ruido

donde "Escuela" es años de escolaridad; "Aptitud" es una puntuación entre 100 y 240 en una prueba de aptitud; "Experiencia" es años de experiencia en la fuerza de trabajo; y "Género" es una variable igual a 1 para las mujeres y cero para los hombres (hablaré más sobre esta variable en un momento). Para producir el conjunto de datos artificiales, hice cincuenta observaciones (correspondientes a cincuenta individuos ficticios) para cada una de las variables explicativas, mitad hombres y mitad mujeres. Al elaborar los datos, deliberadamente intenté introducir alguna correlación positiva entre las variables de escolaridad y de aptitud, por razones que quedarán claras más adelante. A continuación, empleé un generador de números aleatorios para producir un término de ruido extraído de una distribución normal, con un desvío estándar (raíz cuadrada de la varianza) igual a 3.000 y una media igual a cero. Este desvío estándar fue elegido de modo más o menos arbitrario para introducir una cantidad considerable pero no abrumadora de ruido en proporción a la variación total de los ingresos. Las variables del lado derecho se utilizaron entonces para generar el "valor real" de los ingresos para cada uno de los cincuenta "individuos".

El efecto del género en los ingresos en esta empresa hipotética entra a través de la variable Género. Género es una variable "*ficticia*" en la jerga econométrica porque su valor numérico es arbitrario y simplemente captura algún atributo no numérico de la población de la muestra. Por la construcción aquí, los hombres y las mujeres ganan los mismos retornos a la educación, la experiencia, y la aptitud, pero manteniendo estos

[21] Véase, por ejemplo, Miller v. Kansas Electric Power Cooperative, Inc., 1990 WL 120935 (D. Kan.).

factores constantes los ingresos de las mujeres son $ 2.000 más bajos. En efecto, el término constante (ingresos de base) es menor para las mujeres, pero en los otros aspectos las mujeres son tratadas de igual manera. En realidad, por supuesto, la discriminación por razón de género puede surgir de otras maneras (como la menor rentabilidad de la educación y de la experiencia de las mujeres, por ejemplo), y postulo aquí esta forma, que sólo se usa para fines ilustrativos.

Téngase en cuenta que el generador de números aleatorios que he empleado aquí proporciona términos de ruido con un valor esperado de cero, cada uno extraído de una distribución con la misma varianza. Además, los términos de ruido para las distintas observaciones son estadísticamente independientes (el valor realizado del término de ruido de cada observación no tiene influencia sobre el término de ruido utilizado para ninguna otra observación). Por lo tanto, los términos de ruido satisfacen los supuestos necesarios para asegurar que el criterio de mínima SEC produzca estimaciones insesgadas, consistentes y eficientes. El valor esperado de la estimación de cada parámetro es igual al valor verdadero, por lo tanto, y ningún otro estimador lineal que no sea el criterio de mínima SEC hará un mejor trabajo de recuperar los parámetros verdaderos. Sin embargo, es interesante ver qué tan bien se desempeña el análisis de regresión. Utilicé un programa de computación estándar para estimar el término constante y los coeficientes de las cuatro variables independientes de los valores "observados" de Ingresos, Escuela, Aptitud, Experiencia y Género para cada uno de los cincuenta individuos hipotéticos. Los resultados se reproducen en la tabla 1, en la columna denominada "Valor Estimado" (Las tres últimas columnas y el estadístico R^2 los discutiremos en la siguiente sección).

Tabla 1 – Término de ruido con desvío estándar = 3,000

Variable	"Valor verdadero"	Valor estimado	Error estándar	t-estadístico	Prob (2 colas)
Constante	5,000	4136,7	3781,8	1,094	0,280
Escuela	1,000	1584,6	288,1	5,500	0,000
Aptitud	50,0	6,4	27,3	0,236	0,814
Experiencia	300,0	241,7	80,8	2,992	0,004
Género	-2000,0	-1470,4	1402,2	-1,049	0,300
R^2 = 0,646					

Nótese que todos los parámetros estimados tienen el *signo* correcto. Por casualidad, resulta que la regresión sobrestima los retornos de la escolaridad y subestima los otros parámetros. El coeficiente estimado para la aptitud se aleja mucho en proporción de su verdadero valor, y en una sección posterior ofreceré una hipótesis sobre cuál es el problema. Las otras estimaciones de los parámetros, aunque obviamente diferentes del valor real del parámetro subyacente, están mucho más cerca del blanco. Con referencia particular al coeficiente de Género, los resultados de la regresión sugieren correctamente la presencia de discriminación de género, aunque su magnitud está subestimada en alrededor del 25 por ciento (recuérdese que una sobreestimación de la misma magnitud era igual de probable ex ante, es decir, antes de que se generaran valores reales para los términos de ruido).

La fuente del error en las estimaciones de coeficientes es, por supuesto, la presencia de ruido. Si el término de ruido fuera igual a cero para cada observación, los valores verdaderos de los parámetros subyacentes podrían ser recuperados en esta ilustración con exactitud perfecta a partir de los datos de sólo cinco individuos hipotéticos - sería una

simple cuestión de resolver cinco ecuaciones en cinco incógnitas. Además, si el ruido es la fuente de error en las estimaciones de los parámetros, la intuición sugiere que la magnitud del ruido afectará la exactitud de las estimaciones de regresión, con más ruido conduciendo a menos precisión en promedio. Vamos a precisar esta intuición en la siguiente sección, pero antes de proceder quizás sea útil repetir el experimento de estimación de parámetros para una empresa hipotética en la que los datos contengan menos ruido. Para ello, tomé los "datos" de las variables independientes utilizadas en el experimento anterior y de nuevo generé valores de ingresos para los cincuenta individuos hipotéticos usando la ecuación (1), cambiando solamente los términos de ruido. Esta vez, los términos de ruido fueron obtenidos por el generador de números aleatorios de una distribución normal con un desvío estándar de 1.000 en lugar de 3.000 (una reducción significativa). La reestimación de los parámetros de regresión de este conjunto de datos modificados produjo los resultados en la tabla 2:

Tabla 2 – Término de ruido con desvío estándar = 1,000

Variable	"Valor verdadero"	Valor estimado	Error estándar	t-estadístico	Prob (2 colas)
Constante	5,000	4784,2	945,4	5,060	0,000
Escuela	1,000	1146,2	72,0	15,913	0,000
Aptitud	50,0	39,1	6,8	5,741	0,000
Experiencia	300,0	285,4	20,2	14,131	0,000
Género	-2000,0	-1867,6	350,5	-5,328	0,000
R^2 = 0,964					

No es sorprendente que los parámetros estimados aquí estén considerablemente más cerca de sus verdaderos valores. No era cierto que lo serían, porque después de todo sus valores esperados son iguales a sus verdaderos valores independientemente de la cantidad de ruido (el estimador es insesgado). Pero en promedio esperaríamos mayor exactitud, y una mayor precisión surge aquí. Dicho de manera más formal, las distribuciones de probabilidad de las estimaciones de parámetros tienen mayor varianza, cuanto mayor sea la varianza del término de ruido. La varianza del término de ruido afecta así el grado de confianza que tenemos en la exactitud de las estimaciones de regresión.

En aplicaciones reales, por supuesto, el término de ruido es no observable, así como la distribución a partir de la cual se obtiene. La varianza del término de ruido es luego desconocida. Sin embargo, puede calcularse utilizando la diferencia entre los valores predichos de la variable dependiente para cada observación y el valor real (los "errores estimados" definidos anteriormente). Esta estimación, a su vez, permite al investigador evaluar el poder explicativo del análisis de regresión y la "significación estadística" de sus estimaciones de los parámetros.

4 *Inferencia estadística y bondad del ajuste*

Recordemos que las estimaciones de los parámetros son en sí mismas variables aleatorias, que dependen de las variables aleatorias ε. Por lo tanto, cada estimación puede ser pensada como una extracción de alguna distribución de probabilidad subyacente, cuya naturaleza aún no está especificada. Con un supuesto adicional, sin embargo, podemos calcular la distribución de probabilidad de las estimaciones, y utilizarla para docimar hipótesis sobre ellas.

A. *Inferencia Estadística*

La mayoría de los lectores están familiarizados, al menos de paso, con una distribución de probabilidad llamada *normal*. Su forma es la de una "curva en campana", indicando entre otras cosas que si se toma una muestra de la distribución, los valores más probables de las observaciones de la muestra son aquellos próximos a la media y los menos probables son los más alejados de la media. Si asumimos que los términos de ruido ε son extraídos todos de la misma distribución normal, se puede demostrar que las estimaciones de parámetros también tienen una distribución normal.[22]

Sin embargo, la varianza de esta distribución normal depende de la varianza de la distribución a partir de la cual se extraen los términos de ruido. Esta varianza es desconocida en la práctica y sólo puede estimarse usando los errores estimados de la regresión a fin de obtener una estimación de la varianza del término de ruido. La varianza estimada del término de ruido a su vez puede usarse para construir una estimación de la varianza de la distribución normal de cada coeficiente. La raíz cuadrada de esta estimación es denominada *error estándar* del coeficiente – a esta medida la llamaremos *s*.

También es posible mostrar[23] que si la estimación del parámetro, llamémosla *p*, se distribuye normalmente con una media de *m*, entonces *(p - m) / s* tiene una distribución de *t de Student*. La distribución de la *t* se parece mucho a la normal, sólo que tiene colas "más gordas" y su media es cero. Usando este resultado, supongamos que la hipótesis de que el verdadero valor de un parámetro en nuestro modelo de regresión es *m*. Llamemos a ésta la "hipótesis nula". Como el criterio de mínima SEC es un estimador insesgado, podemos deducir que nuestra estimación del parámetro se obtiene de una distribución normal con media *m* si la hipótesis nula es verdadera. Este estadístico puede ser positivo o negativo, según la estimación del parámetro a partir de la cual se deriva sea mayor o menor que el valor real hipotético del parámetro. Recordando que la distribución *t* es muy parecida a una normal con una media de cero, sabemos que los grandes valores del estadístico *t* (en valor absoluto) serán extraídos considerablemente menos frecuentemente que los pequeños valores de ese estadístico *t*. Y, por construcción del estadístico *t*, se tendrán grandes valores de ese estadístico (en valor absoluto), permaneciendo iguales otras cosas, cuando la estimación del parámetro en la que se basa difiera mucho de su valor real (hipotético).

Este punto de vista se da vuelta para un test de hipótesis. Acabamos de argumentar que un amplio estadístico *t* (en valor absoluto) surgirá con poca frecuencia si es correcta la hipótesis nula. Por lo tanto, cuando se tiene un amplio estadístico *t*, *será tentador concluir que la hipótesis nula es falsa*. *La esencia del test de hipótesis con un coeficiente de regresión, entonces, es formular una hipótesis nula en cuanto a su verdadero valor, y luego decidir si se acepta o rechaza según si el estadístico t asociado a esa hipótesis*

---

[22] Ver, por ejemplo, E. Hanushek y J. Jackson, supra nota 13, 66 - 68; J. Johnston, supra nota 13, 135– 38. El supuesto de que los términos de ruido están normalmente distribuidos resulta a menudo intuitivamente plausible y puede ser en líneas generales justificado recurriendo a "teoremas centrales del límite", que sostienen que el promedio de un gran número de variables aleatorias tiende hacia una distribución normal aún si las variables aleatorias individuales que entran en el promedio no se distribuyen normalmente. Véase, por ejemplo, R. Hogg y A. Craig, *Introduction to Mathematical Statistics* 192 - 95 (4a ed., 1978); W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, 243 - 48 (3d ed., 1968). Luego, si pensamos en el término de ruido como la suma de un gran número de pequeñas perturbaciones independientes, la teoría proporciona una base considerable para suponer que su distribución es aproximadamente normal.

[23] Véanse las fuentes citadas nota 22 *supra*.

nula es lo suficientemente grande como para que la plausibilidad de la hipótesis nula esté suficientemente en duda.[24]

Se puede ser un poco más preciso. Podemos concluir que la hipótesis nula es implausible si el estadístico *t* asociado con nuestra estimación de regresión se encuentra tan alejado en una cola de su distribución *t* que tal valor, o uno aún mayor en valor absoluto, surgirían menos que, digamos, 5 por ciento de las veces si la hipótesis nula es correcta. Dicho de otro modo, rechazaremos la hipótesis nula si el estadístico *t* cae en la parte superior de la cola de la distribución *t*, que contiene el 2,5 por ciento de las extracciones que representan los mayores valores positivos, o en la cola más baja, que contiene el 2,5 por ciento de las extracciones que representan los mayores valores negativos. A esto se llama un *test de dos colas*.

Alternativamente, podríamos tener una fuerte creencia previa sobre el verdadero valor de un parámetro que nos llevaría a aceptar la hipótesis nula, incluso si el estadístico t se encuentra muy lejos en *una* de las colas de la distribución. Considérese el coeficiente de la variable *dummy* de género en la tabla 1 como ilustración. Supongamos que la hipótesis nula es que el verdadero valor de este coeficiente es cero. ¿Bajo qué circunstancias lo rechazamos? Podríamos considerar inverosímil que el verdadero valor del coeficiente sea positivo, reflejando una discriminación *en contra* de los hombres. Entonces, incluso si el coeficiente estimado para la dummy de género es positivo con un amplio estadístico t positivo, todavía aceptaríamos la hipótesis nula de que su valor verdadero sea cero. Sólo una estimación de un coeficiente negativo con un amplio estadístico *t* negativo nos llevaría a concluir que la hipótesis nula era falsa. Cuando rechazamos la hipótesis nula sólo si un estadístico *t* que es amplio en valor absoluto tiene un signo particular, estamos empleando un *test de una cola*.

Para operacionalizar un test a una o a dos colas, es necesario calcular la probabilidad exacta de un estadístico *t* igual o mayor en valor absoluto que la asociada con la estimación del parámetro en cuestión. A su vez, es necesario saber exactamente cómo se "dispersa" la distribución *t* de la que se ha extraído la estimación. Un parámetro adicional que necesitamos para definir la forma de la distribución *t* en cuestión se denomina *grados de libertad*, definido como el número de observaciones en la muestra menos el número de parámetros a estimar. En las ilustraciones de las tablas 1 y 2, tenemos 50 observaciones en la muestra, y estamos estimando 5 parámetros, por lo que la distribución *t* para cualquier estimación de los parámetros tiene 45 grados de libertad. Cuanto menos sean los grados de libertad, más "dispersa" será la distribución *t* y, por lo tanto, mayor será la probabilidad de extraer grandes estadísticos *t*. La intuición es que cuanto mayor sea la muestra, más colapsada es la distribución de cualquier estimación de parámetros (recuérdese el concepto de consistencia anterior). Por el contrario, cuanto más parámetros buscamos estimar a partir de una muestra de tamaño dado, más información estamos tratando de extraer de los datos y menos confianza podemos tener en la estimación de cada parámetro - por lo tanto, la distribución *t* asociada resulta más *esparcida*.[25]

---

[24] Limito aquí la discusión a un test de hipótesis con respecto al valor de un parámetro particular. De hecho, pueden ser fácilmente docimados otros tipos de hipótesis, como la de que *todos* los parámetros del modelo son nulos, la hipótesis de que *algún subconjunto* de parámetros son nulos, y así sucesivamente.

[25] Véanse las fuentes citadas nota 13 *supra*.

Al conocer los grados de libertad de la distribución *t*, el investigador puede calcular la probabilidad de extraer el estadístico *t* en cuestión, o uno mayor en valor absoluto, asumiendo la verdad de la hipótesis nula. Usando el test de una o dos colas (el primero sólo es necesario cuando el estadístico *t* tiene el signo correcto), el investigador rechaza la hipótesis nula si esta probabilidad es suficientemente pequeña.

Pero, ¿qué se quiere decir con *suficientemente pequeña*? La respuesta no es obvia, y depende de las circunstancias. Se ha convertido en convención en la investigación científica social practicar el test de una hipótesis nula particular - a saber, la hipótesis de que el verdadero valor de un coeficiente es cero. Bajo esta hipótesis, μ en nuestra notación anterior es igual a cero, y por lo tanto el estadístico *t* es simplemente p/s, la estimación del coeficiente dividida por su error estándar. También es convención adoptar un "nivel de significación" de 0.10, 0.05 o 0.01, es decir, preguntar si el estadístico *t* que el investigador ha obtenido, o uno aún mayor en valor absoluto, surgiría más de 10 por ciento, 5 por ciento, o 1 por ciento del tiempo cuando la hipótesis nula es correcta. Cuando la respuesta a esta pregunta es negativa, la hipótesis nula es rechazada y el coeficiente en cuestión se dice que es "estadísticamente significativo". Por ejemplo, si el parámetro estimado que se obtuvo está lo suficientemente lejos de cero que una estimación de esa magnitud, o uno aún más allá de cero, se produciría menos del 5 por ciento de las veces, entonces se dice que el coeficiente es significativo al nivel de 0,05.

La cuestión de si las pruebas convencionales de significación de las ciencias sociales son apropiadas cuando se utiliza el análisis de regresión en aplicaciones legales, particularmente en los pleitos, es un tema difícil que pospondré hasta la sección conclusiva de esta conferencia. Simplemente supondré por ahora que estamos interesados en el problema general de docimar alguna hipótesis nula y que la rechazaremos si el parámetro estimado obtenido se encuentra suficientemente alejado en una de las colas de la distribución a partir de la cual se ha sacado la estimación. Dejamos abierta la cuestión de qué significa estar "suficientemente alejado" y simplemente buscamos calcular la probabilidad con un test de una o dos colas para obtener una estimación tan alejada de la media de la distribución como la generada por la regresión si la hipótesis nula fuera cierta.

La mayoría de los paquetes de regresión computarizados no sólo reportan la estimación del parámetro (π en nuestra notación), sino también el error estándar de cada parámetro estimado (*s* en nuestra notación). Este valor, junto con el valor del parámetro verdadero supuesto (μ en nuestra notación), puede ser empleado para generar el estadístico *t* apropiado para cualquier hipótesis nula. Muchos paquetes de regresión también reportan un número llamado "estadístico-t", que invariablemente se basa en la hipótesis nula convencional de las ciencias sociales, de que el verdadero valor del parámetro es cero. Por último, algunos paquetes informan de la probabilidad de que el estadístico *t* en cuestión pueda haberse generado a partir de una distribución *t* con los grados de libertad adecuados en un test de una o dos colas.[26]

[26] Si el paquete de regresión no informa estas probabilidades, pueden encontrarse fácilmente en otra parte. Se ha convertido en práctica habitual incluir en los libros de estadística y econometría tablas de probabilidades para distribuciones-t con diferentes grados de libertad. Conociendo los grados de libertad asociados a un estadístico *t*, por lo tanto, se puede consultar dicha tabla para determinar la probabilidad de obtener un estadístico t tan alejado de cero como el genera-

Volviendo a las tablas 1 y 2, toda esta información es reportada para cada uno de los cinco parámetros estimados: el error estándar, el valor del estadístico *t* para la hipótesis nula de que el verdadero valor del parámetro es cero y la probabilidad de obtener un t-estadístico igual o mayor en valor absoluto bajo un test de dos colas con 45 grados de libertad. Para interpretar esta información, considérese el coeficiente estimado de la dummy de género en la tabla 1. El coeficiente estimado de - 1470.4 tiene un error estándar de 1402.2 y por lo tanto un estadístico *t* de - 1470.4 / 1402.2 = - 1.049. La probabilidad asociada en un test de dos colas es reportada como .30. Esto significa que si el valor real del coeficiente de la dummy de género fuera cero, sin embargo un coeficiente mayor o igual a 1470.4 en valor absoluto surgiría 30 por ciento del tiempo dados los grados de libertad de la distribución *t* de la cual se extrae el coeficiente estimado. Un rechazo de la hipótesis nula sobre la base de un parámetro estimado igual o mayor a 1470.4 en valor absoluto, por lo tanto, será erróneo tres veces sobre diez cuando la hipótesis nula sea verdadera. Por lo tanto, según los estándares convencionales en ciencias sociales, el nivel de significación aquí es demasiado bajo para rechazar la hipótesis nula, y el coeficiente de la dummy de género no es considerado estadísticamente significativo. Es de notar que en este caso (en contraste con las aplicaciones del mundo real), conocemos el verdadero valor del parámetro, a saber - 2000.0. Por lo tanto, si empleamos un test de significación convencional de dos colas, se nos induce erróneamente a rechazar la hipótesis de que la discriminación de género esté presente.

Como se ha señalado, se puede considerar que el test a dos colas es inadecuado para el coeficiente de la dummy de género porque consideramos que la posibilidad de discriminación contra los hombres es poco plausible. Es sencillo construir un test alternativo de una cola: la Tabla 1 indica que un coeficiente estimado de 1470,4 o mayor en valor absoluto ocurrirá un 30 por ciento del tiempo si el valor verdadero del coeficiente es cero. Dicho de otra manera, un coeficiente estimado de género mayor o igual a 1470.4 aparecerá el 15 por ciento del tiempo, y una estimación menor o igual a - 1470.4 aparecerá otro 15 por ciento del tiempo. Se deduce que si sólo estamos interesados en la parte inferior de la distribución *t*, el rechazo de la hipótesis nula (cuando sea verdadera) será erróneo sólo el 15 por ciento del tiempo si se requiere un parámetro estimado de - 1470,4 o menor. El nivel de significación de una cola es, por tanto, 0.15, aún por debajo de los umbrales convencionales de significación estadística.[27] Usando estos niveles de significación, por lo tanto, se nos induce de nuevo a aceptar la hipótesis nula, en este caso erróneamente.

Ofrezco esta ilustración no para sugerir que haya algo malo en los tests de significación convencionales, sino simplemente para indicar cómo se reduce la probabilidad de rechazar erróneamente la hipótesis nula (llámese esto un error de "Tipo I") aumentando la probabilidad de aceptarla erróneamente (llámese esto un error de "Tipo II"). Los tests de significación convencionales dan un gran peso a la importancia de evitar errores de tipo I y menos peso a evitar los de tipo II, exigiendo un alto grado de confianza

---

do por la regresión (el concepto "alejado de cero" es, nuevamente, definido por un test de una o dos colas). Como punto de referencia, cuando los grados de libertad son grandes (digamos, 50 o más), entonces el nivel de significación de 0.05 para un test de dos colas requiere un estadístico *t* aproximadamente igual a 2.0.

[27] El resultado de esta ilustración es general: para cualquier estadístico *t*, la probabilidad de rechazar la hipótesis nula erróneamente bajo un test de una cola será exactamente la mitad de esa probabilidad bajo un test de dos colas.

en la falsedad de la hipótesis nula antes de rechazarla. Esto parece perfectamente apropiado para la mayoría de las aplicaciones científicas, en las que se pide justificadamente al investigador que soporte una carga probatoria considerable antes de que la comunidad científica acepte que los datos establecen una relación causal afirmada. Que el proponente de la evidencia de regresión en un procedimiento legal deba soportar la misma carga probatoria es una cuestión más sutil.

B. *Bondad del ajuste*

Otro estadístico común asociado al análisis de regresión es el R^2. Éste tiene una definición simple: es igual a uno menos la razón entre la suma de los errores estimados al cuadrado (la suma de los cuadrados de los desvíos del valor real de la variable dependiente respecto a la línea de regresión) y la suma de los desvíos cuadráticos respecto a la media de la variable dependiente. Intuitivamente, la suma de los desvíos cuadráticos respecto a la media es una medida de la variación total de la variable dependiente. La suma de los desvíos cuadráticos respecto a la línea de regresión es una medida del grado en que la regresión no explica la variable dependiente (una medida del ruido). Por lo tanto, el estadístico R^2 es una medida del grado en que la variación total de la variable dependiente es explicada por la regresión. No es difícil demostrar que el estadístico R^2 necesariamente toma un valor entre cero y uno.[28]

Un alto valor de R^2, que sugiere que el modelo de regresión explica bien la variación en la variable dependiente, es obviamente importante si se desea utilizar el modelo con fines predictivos o de pronóstico. Es considerablemente menos importante si uno está simplemente interesado en estimaciones de parámetros particulares (como, por ejemplo, si se está buscando evidencia de discriminación, como en nuestra ilustración, y por lo tanto uno se enfoca en el coeficiente de la dummy de género). Por supuesto, una gran variación inexplicada en la variable dependiente aumentará el error estándar de los coeficientes del modelo (que son una función de la varianza estimada del término de ruido), y por consiguiente regresiones con bajos valores de R^2 a menudo (pero no siempre) producirán estimaciones de parámetros con pequeños estadísticos t para cualquier hipótesis nula. Debido a que esta consecuencia de un R^2 bajo se reflejará en los estadísticos *t*, sin embargo, no ofrece ninguna razón para preocuparse por un R^2 bajo por sí.

Como ilustración rápida, volvamos a las tablas 1 y 2. Recuérdese que los términos de ruido del conjunto de datos con el que se generaron las estimaciones en la tabla 1 se obtuvieron de una distribución con un desvío estándar 3.000, mientras que para la tabla 2 los términos de ruido se extrajeron de una distribución con un desvío estándar 1.000. Por lo tanto, la variación no explicada de la variable ingresos es probable que sea mayor en el primer conjunto de datos, y de hecho los estadísticos R^2 así lo confirman (.646 para la tabla 1 y .964 para la tabla 2). Del mismo modo, como la varianza estimada del término de ruido es mayor para las estimaciones en la tabla 1, esperamos que los coeficientes estimados tengan errores estándar mayores y estadísticos *t* más pequeños. Esta expectativa también se confirma inspeccionando ambas tablas. Las variables con coeficientes estadísticamente significativos según tests convencionales en la tabla 2, por eso, como la dummy de género, no son estadísticamente significativas en la tabla 1.

[28] Véase, por ejemplo, E. Hanushek y J. Jackson, supra nota 13, 57 - 58.

En estas ilustraciones, el valor de R^2 simplemente refleja la cantidad de ruido en los datos, y un R^2 bajo no es inconsistente con el criterio de mínima SEC que sirve como un estimador imparcial, consistente y eficiente porque sabemos que los términos de ruido fueron todos extracciones independientes de la misma distribución con media cero. En la práctica, sin embargo, un valor bajo de R^2 puede estar indicando que se han omitido factores importantes y sistemáticos del modelo de regresión. Esta posibilidad plantea de nuevo la preocupación por el sesgo de variables omitidas.

5 *Dos problemas estadísticos comunes en el análisis de regresión*

Gran parte del curso típico de econometría está dedicado a lo que sucede cuando los supuestos necesarios para hacer que el criterio de mínima SEC sea insesgado, consistente y eficiente no se cumplen. No puedo comenzar a dar una razón completa de estos temas en una conferencia tan breve, y simplemente ilustraré dos de las muchas complicaciones que pueden surgir, elegidas porque son comunes y bastante importantes.

A. *Variables omitidas*

Como se observó, la omisión en una regresión de algunas variables que afectan a la variable dependiente puede causar un "sesgo por variables omitidas". El problema aparece porque cualquier variable omitida pasa a formar parte del término de ruido y el resultado puede violar el supuesto necesario para que el criterio de mínima SEC proporcione un estimador insesgado.

Recordemos ese supuesto - que cada término de ruido se extrae de una distribución con una media de cero. Se observó que este supuesto lógicamente implica la ausencia de correlación entre las variables explicativas incluidas en la regresión y el valor esperado del término de ruido (porque cualquiera que sea el valor de cualquier variable explicativa, el valor esperado del término de ruido es siempre cero). Por lo tanto, supongamos que comenzamos con un modelo correctamente especificado en el cual el término de ruido de cada observación tiene un valor esperado de cero. Ahora, omitamos una de las variables independientes. Si el efecto de esta variable sobre la variable dependiente no es cero para cada observación, los nuevos términos de ruido provienen ahora de distribuciones con medias no nulas. Una consecuencia es que la estimación del término constante estará sesgada (parte del valor estimado del término constante es en realidad el efecto medio de la variable omitida). Además, a menos que la variable omitida no esté correlacionada con las incluidas, los coeficientes de las incluidas estarán sesgados porque ahora reflejan no sólo una estimación del efecto de la variable a la que están asociados, sino también parcialmente los efectos de la variable omitida.[29]

Para ilustrar el problema de las variables omitidas, tomé los datos sobre los que se basan las estimaciones reportadas en la tabla 1 y volví a correr la regresión después de omitir la variable de escolaridad. Los resultados se muestran en la tabla 3:

[29] Véase J. Johnston, *supra* nota 13, 168-69; E. Hanushek y J. Jackson, *supra* nota 13, 81– 82. El sesgo es una función de dos cosas: los verdaderos coeficientes de las variables excluidas y la correlación dentro del conjunto de datos entre variables incluidas y excluidas.

Tabla 3 – Ilustración de Variables Omitidas

Variable	"Valor verdadero"	Valor estimado	Error estándar	t-estadístico	Prob (2 colas)
Constante	5,000	9806,5	4653,8	2,107	0,041
Escuela	1,000	omitida			
Aptitud	50,0	107,5	25,6	4,173	0,000
Experiencia	300,0	256,9	103,3	2,487	0,017
Género	-2000,0	-2445,5	1779,0	-1,375	0,176
R^2 = 0,408					

Ustedes observarán que la omisión de la variable de escolaridad disminuye el R^2 de la regresión, lo que no es sorprendente dada la importancia original de la variable. También altera los coeficientes estimados. La estimación del término constante aumenta considerablemente, porque el efecto medio de la escolaridad sobre el ingreso es positivo. No es de extrañar que el término constante sea así estimado como mayor que su verdadero valor. Un efecto aún más significativo de la omisión de la escolaridad está en la estimación del coeficiente de la variable aptitud, que aumenta dramáticamente de estar por debajo de su valor verdadero a estar muy por encima de su valor verdadero y llega a ser altamente significativa. La razón es que la variable de escolaridad está altamente correlacionada (positivamente) con la aptitud en el conjunto de datos -la correlación es 0.69- y que la escolarización tiene un efecto positivo sobre los ingresos. Por lo tanto, con la variable de escolaridad omitida, el coeficiente de aptitud está captando erróneamente algunos de los rendimientos (positivos) de la educación, así como los rendimientos de la "aptitud". La consecuencia es que el criterio de mínima SEC produce una estimación sesgada hacia arriba del coeficiente de aptitud, y en este caso la estimación real está realmente por encima del verdadero valor de ese coeficiente.

El efecto sobre los otros coeficientes es más modesto, aunque no trivial. Obsérvese, por ejemplo, que el coeficiente de género aumenta (en valor absoluto) de manera significativa. Esto se debe a que la escolarización está positivamente correlacionada con el ser masculino en mi conjunto de datos ficticios - sin controlar la escolaridad, el efecto aparente del género es exagerado porque las mujeres son un poco menos educadas en promedio.

El problema de las variables omitidas resulta problemático para los investigadores, no sólo porque exige recopilar datos sobre más variables para evitarlo, sino porque las variables omitidas a menudo no son observables. Los estudios del mundo real sobre la discriminación de género son quizás un buen ejemplo. Se puede imaginar fácilmente que los ingresos dependan de factores tales como la habilidad innata y la motivación, que pueden ser inobservables para un investigador. El sesgo de variable omitida puede convertirse en algo que el investigador no pueda evitar, y comprender cuáles son sus consecuencias se torna doblemente importante. Para un investigador preocupado ante todo por el coeficiente de la *dummy* de género, podría argumentarse que el sesgo de variable omitida causado por la exclusión de la habilidad innata y la motivación debería ser moderado porque razonablemente podría pensarse que la correlación en la muestra entre el género y las variables omitidas debe ser reducida. El problema es serio en caso

contrario, y la utilidad de la regresión convencional como herramienta de investigación se reduce en forma considerable.[30]

Advierto de pasada que el problema de incluir variables sin relación o irrelevantes es menos grave. Su coeficiente esperado es cero y las estimaciones de los demás coeficientes no son sesgadas, aunque se reduce la eficiencia del criterio de mínima SEC.[31]

También noto al pasar otro problema que está estrechamente relacionado con el problema de variable omitida, llamado de "errores en las variables". En muchos estudios de regresión, es inevitable que algunas variables explicativas se midan con error. Tales errores se convierten en parte del término de ruido. Supongamos que, en ausencia de error de medición, los términos de ruido obedezcan al supuesto necesario para que haya ausencia de sesgo y consistencia: todos se extraen de una distribución con una media cero y, por lo tanto, no están correlacionados con las variables explicativas. Si hay error de medición, sin embargo, este supuesto ya no se mantendrá.

Imaginen, por ejemplo, que los ingresos dependen de la educación, la experiencia, etc., como se postuló anteriormente, y de la capacidad innata que se ha sugerido. En lugar de suponer que la capacidad innata es una variable omitida, supongan que la puntuación de una prueba de aptitud incluida en la regresión es una "proxy" de la capacidad innata. Es decir, la consideramos una medición imperfecta de la capacidad, correlacionada pero no perfectamente. Cuando el puntaje de la prueba subestima la capacidad, el término de ruido aumenta, y cuando el puntaje de aptitud sobrestima la capacidad, el término de ruido cae. El resultado es una correlación negativa entre el término de ruido y la variable aptitud / capacidad. Dicho de otro modo, si el término de ruido sin error de medición se obtiene de una distribución con media cero, el término de ruido que incluya el error de medición será extraído de una distribución con media igual a la magnitud de ese error. Una vez más, la consecuencia será un sesgo de los coeficientes estimados del modelo.[32]

B. *Multicolinealidad*

El problema de multicolinealidad no da como resultado estimaciones de coeficientes sesgados, pero aumenta el error estándar de las estimaciones y, por lo tanto, reduce el grado de confianza que uno puede depositar en ellas. La dificultad surge cuando dos variables independientes están estrechamente correlacionadas, creando una situación en la que sus efectos son difíciles de separar.

---

[30] Los econometristas han desarrollado algunas técnicas de regresión más sofisticadas para abordar el problema de las variables inobservables, pero no siempre son satisfactorias debido a ciertos supuestos restrictivos que hay que hacer al utilizarlas. Véase, por ejemplo, Zvi Griliches, *Errors in Variables and Other Unobservables*, 42 Econometrica 971 (1974). Se puede hallar una discusión accesible sobre el problema de variables omitidas y cuestiones relacionadas en P. Kennedy, supra nota 17, p. 69-72.

[31] Íd.

[32] Una técnica estándar para abordar este problema es la de *variables instrumentales*, que reemplaza a la variable contaminada por otra variable que se cree que está estrechamente asociada con ella, pero que también se cree que no está correlacionada con el término de perturbación. Sin embargo, por diversos motivos, la técnica de variables instrumentales no es satisfactoria en muchos casos, y el problema de errores en las variables es, por consiguiente, una de las dificultades más serias en el uso de las técnicas de regresión. Una discusión de la técnica de variables instrumentales y otras posibles respuestas a los errores en el problema de las variables se puede hallar en P. Kennedy, supra nota 17, p. 113-16; J. Johnston, supra nota 13, p. 281-91.

La siguiente ilustración transmite la intuición esencial: supongan que dos profesores de la facultad de derecho (llámenlos Baird y Picker) ofrecen charlas regulares en los almuerzos de los alumnos, en parte con el propósito de estimular los aportes de los alumnos. Supongamos que cada vez que uno pronuncia un discurso en el almuerzo, el otro también lo hace, y que el único dato disponible sobre los aportes de los alumnos es el aporte mensual agregado. O sea que sabemos que cada vez que ambos pronuncian un discurso en un mes, los aportes de los alumnos aumentan en $ 10,000. Cuando cada uno da dos charlas en un mes, los aportes aumentan en $ 20,000, etcétera.

Luego, por hipótesis, conocemos el efecto conjunto de un discurso de Baird y Picker ($ 10,000), pero no hay nada en los datos que nos permita determinar sus efectos individuales. Tal vez cada discurso incremente los aportes en $ 5,000, pero podría ser que un orador provoque $ 10,000 adicionales en aportes y el otro ninguno, o que un orador induzca $ 30,000 adicionales en aportes y el otro los reduzca en $ 20,000. En jerga econométrica, los datos sobre las charlas dadas por Baird y Picker son perfectamente *colineales*: la correlación entre el número de charlas mensuales de Baird y las de Picker es 1.0. Todo intento de calcular el efecto de la cantidad de charlas dada por cada uno sobre los aportes fallaría y daría como resultado un mensaje de error de la computadora (por razones que no necesitamos detallar, estaría tratando de dividir por cero).

El término "multicolinealidad" habitualmente se refiere a un problema en los datos que no implica la colinealidad perfecta como en nuestra ilustración, pero donde los cambios en las dos variables están, sin embargo, altamente correlacionados al punto de que sea difícil separar sus efectos. Debido a que la multicolinealidad no se aplica a ninguna propiedad del término de ruido, el criterio de mínima SEC puede seguir siendo insesgado, consistente y eficiente. Pero la dificultad en separar los efectos de ambas variables introduce una mayor incertidumbre en el estimador, que se manifiesta como un aumento del error estándar de los coeficientes y una reducción de sus estadísticos *t*.

Puede que ya se haya proporcionado una ilustración de los efectos de la multicolinealidad. En nuestra discusión de la tabla 1, notamos que la estimación del coeficiente de la variable aptitud estaba muy por debajo de su valor verdadero. Resulta que la aptitud y la escolaridad están altamente correlacionadas en el conjunto de datos, y esto proporciona una conjetura plausible de por qué el coeficiente de la variable escuela es demasiado alto y el de aptitud insignificantemente pequeño (algunos de los efectos de la aptitud en la muestra son captados por el coeficiente de escolaridad).

Para proporcionar otra ilustración, que por cierto nos permite introducir otro uso de las variables ficticias, supongamos que la discriminación de género en nuestra empresa hipotética afecta los ingresos de las mujeres de dos maneras: a través de un efecto sobre los ingresos iniciales de las mujeres como antes, y a través de un efecto sobre los retornos a la educación de las mujeres. En particular, recuerden que en la ecuación (1) ambos sexos ganaban $ 1,000 por año de escolaridad. Supongamos ahora que los hombres ganan $ 1,000, pero las mujeres ganan sólo $ 800. Este efecto puede ser captado matemáticamente mediante un "término de interacción" que incorpore la *dummy* de género, de modo que los ingresos ahora se determinan de acuerdo con la ecuación (2):

(2) Ingresos = 5000 + 1000 • Escuela + 50 • Aptitud + 300 • Experiencia - 2000 • Género – 200 • Género • Escuela + Ruido

Usando los mismos datos hipotéticos para las variables explicativas que antes, obtuve nuevos valores de ingresos usando la ecuación (2) y (sólo en beneficio de la variedad) términos de ruido extraídos de una distribución con desvío estándar de 2,000. Luego estimé una regresión con el nuevo conjunto de datos, incluyendo la variable Género • Escuela como una variable explicativa adicional. Los resultados figuran en la tabla 4, donde la variable "Interacción" es simplemente Género • Escuela.

Tabla 4 – Ilustración de Multicolinealidad

Variable	"Valor verdadero"	Valor estimado	Error estándar	t-estadístico	Prob (2 colas)
Constante	5,000	3500,2	2130,6	1,643	0,108
Escuela	1,000	962,3	144,0	6,679	0,000
Aptitud	50,0	61,2	12,3	4,966	0,000
Experiencia	300,0	288,3	36,7	7,861	0,000
Género	-2000,0	-4243,7	2916,5	-1,455	0,153
Interacción	-200,0	14,8	198,2	0,075	0,941
R^2 = 0,909					

Obsérvese que, a diferencia de la tabla 1, el coeficiente para la *dummy* de género ahora es más alto que el valor real en más del doble. El coeficiente para el término de interacción, por el contrario, tiene el signo incorrecto y es próximo a cero. Las otras estimaciones de los parámetros no están demasiado erradas.

Los pobres resultados de los coeficientes de Género e Interacción resultan casi con certeza de un grave problema de multicolinealidad. Téngase en cuenta que siempre que Género = 0, Interacción = 0 también. Género es positivo sólo cuando Interacción es positivo. Esperaríamos una alta correlación entre los dos y, de hecho, el coeficiente de correlación es 0.96. En estas circunstancias, no es de extrañar que la regresión no pueda separar los efectos de las dos variables con ninguna precisión. El coeficiente estimado de Interacción es insignificante en cualquier test plausible y el coeficiente de Género también tiene un gran error estándar que da lugar a un estadístico *t* bastante pobre a pesar del alto valor absoluto del coeficiente estimado.

Sin embargo, pese a la considerable incertidumbre de los coeficientes estimados, es plausible que el problema de multicolinealidad no sea tan desastroso para alguien que esté interesado en identificar el alcance de la discriminación por género. La razón es que los efectos *conjuntos* estimados de Género e Interacción pueden no estar muy alejados: uno es inflado y el otro es subestimado, los errores se cancelan en gran medida entre sí y, para la cuestión legal, el efecto conjunto estimado puede ser todo lo que se necesita. La advertencia es que la multicolinealidad reduce los estadísticos t de ambas variables y, por lo tanto, podría llevar al investigador a rechazar la hipótesis de que la discriminación esté presente en absoluto. Para tratar los efectos de la multicolinealidad aquí, por lo tanto, el investigador podría simplemente dejar a un lado los estadísticos *t* bajos, o bien omitir una de las dos variables y reconocer que la estimación del coeficiente de la variable incluida estará sesgado e incluirá el efecto de la variable omitida.[33]

En muchos casos, sin embargo, el investigador no quedará satisfecho con una estimación del efecto conjunto de ambas variables, y deberá separarlas. Aquí, la multicolinealidad puede volverse muy problemática. No hay una solución simple y aceptable para

[33] Es importante recordar que este enfoque también plantea el problema del sesgo de variable omitida para las otras variables.

todos los casos, aunque varias opciones merecen ser consideradas, más allá del alcance de esta conferencia.[34]

6 *Una última nota sobre el Derecho: Análisis de regresión y la Carga de la Prueba*

Una cuestión clave que debe enfrentarse cada vez que se introduce un estudio de regresión en un pleito es la cuestión de cuánto peso hay que darle. Espero que las ilustraciones de esta conferencia brinden alguna base para el optimismo de que tales estudios pueden ser útiles, al tiempo que sugieren una base considerable para la precaución en su uso.

Vuelvo ahora a un tema diferido anteriormente en la discusión de los tests de hipótesis: la relación entre *el test de significación estadística* y la *carga de la prueba*. Supongamos, por ejemplo, que para establecer la *responsabilidad* por discriminación salarial sobre la base del género según el Título VII, un demandante simplemente debe demostrar por la preponderancia de la evidencia, que las mujeres empleadas por el acusado sufren alguna medida de discriminación.[35] Con referencia a nuestra primera ilustración, podríamos decir que la demostración requerida de responsabilidad es que, por preponderancia de la evidencia, el coeficiente de la *dummy* de género sea negativo.

Lamentablemente, no existe una relación simple entre esta carga de prueba y el test de significación estadística. En un extremo, si imaginamos que la estimación del parámetro en el estudio de regresión es la única información que tenemos sobre la presencia o ausencia de discriminación, se podría argumentar que la responsabilidad se establece por una preponderancia de la evidencia si el coeficiente estimado para la *dummy* de género es negativo independientemente de su significación estadística o error estándar. La justificación sería que la estimación negativa, sin perjuicio de la incertidumbre, sea insesgada y sea la mejor evidencia que tenemos.

Pero esto es demasiado simplista. Pocas veces se da que la estimación de una regresión sea la única información disponible, y cuando los errores estándar son altos, la estimación puede ser la información disponible menos confiable. Además, el análisis de regresión puede estar sujeto a manipulaciones considerables. No es obvio, precisamente, qué variables deban incluirse en un modelo o qué proxies usar para variables incluidas que no pueden medirse con precisión. Hay espacio considerable para la experimentación, y esta experimentación puede convertirse en *minería de datos*, por medio de la cual un investigador intenta numerosas especificaciones de regresión hasta que aparezca el resultado deseado. Un abogado, naturalmente, puede tener tendencia a presentar sólo las estimaciones que respaldan la posición de su cliente. Por lo tanto, si el mejor resultado que un abogado puede presentar contiene errores estándar elevados y poca significación estadística, a menudo es plausible suponer que numerosos resultados aún menos impresionantes quedan ocultos y, posiblemente, están protegidos de revelación por la *doctrina de protección del producto del trabajo de un abogado*.[36]

---

[34] Ver P. Kennedy, supra nota 17, p. 146-56.

[35] Véase, por ejemplo, *Texas Department of Community Affairs v. Burdine, 450 U.S. 248 (1981)*.

[36] No voy a hacer una digresión sobre las reglas de revelación aquí. En la práctica, los datos en bruto pueden ser detectables, por ejemplo, si bien el análisis no revelado de los datos por parte del experto puede no serlo.

Por estas razones, quienes usan el análisis de regresión en litigios tienden a reportar resultados que satisfacen los tests de significación convencionales, a menudo un nivel de significatividad del 5 por ciento, y suponer que los resultados menos significativos no son demasiado interesantes.[37] Antes de que la mayoría de los expertos se sientan cómodos afirmando que la discriminación por género ha quedado establecida por un estudio como el de nuestra ilustración, por lo tanto, es probable que requieran que el coeficiente estimado de la *dummy* de género sea negativo y estadísticamente significativo. Aún así, anticiparían un enérgico interrogatorio basado en una serie de cuestiones, como las sugeridas por la discusión anterior.

Todavía surgen problemas más delicados cuando se necesita una estimación exacta de parámetros con algún fin, como por ejemplo, para calcular *daños*. El hecho de que el parámetro sea "estadísticamente significativo" simplemente significa que mediante pruebas convencionales, se puede rechazar la hipótesis de que su verdadero valor sea cero. Pero seguramente habrá muchas otras hipótesis sobre el valor del parámetro que no se puedan rechazar, y de hecho la probabilidad de que la regresión produzca una estimación perfectamente precisa de cualquier parámetro es despreciable. La única guía que se puede proporcionar desde el punto de vista estadístico es obvia - los parámetros con errores estándar proporcionalmente bajos tienen menos probabilidades de estar lejos del blanco que los otros.

Por lo tanto, en definitiva los estadísticos en sí no indican qué peso cabe dar a un estudio de regresión, o si es razonable usar un parámetro estimado particular con algún fin legal u otro. Estas evaluaciones son encomendadas inevitablemente a los jueces o a los miembros del jurado, cuyas sentencias al respecto, si están bien informadas, son probablemente tan buenas como las de cualquier otra persona.

[37] Véase la discusión en Franklin Fisher, *Statisticians, Econometricians and Adversary Proceedings*, 81 J. Am. Stat. Assn. 277 (1986).