COMPLEMENTOS DE ECONOMETRÍA[1]

1. Introducción

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

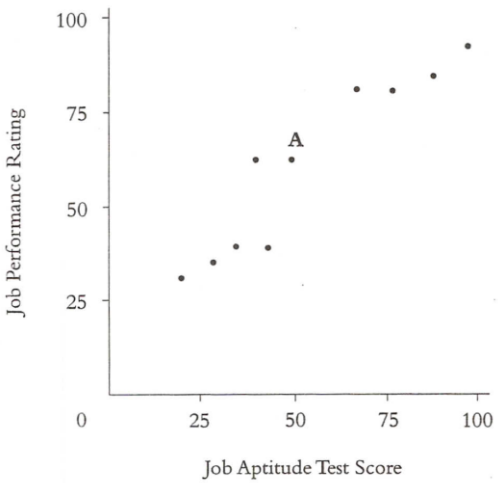


Figura 1 – Diagrama de dispersión

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

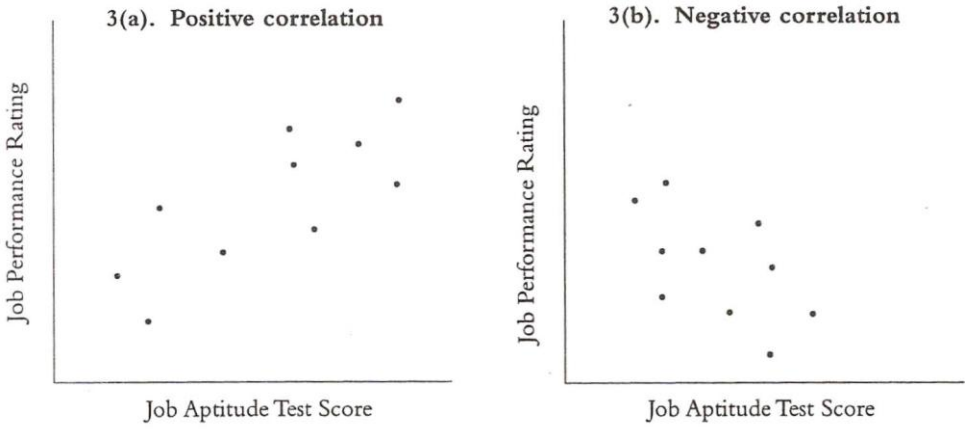
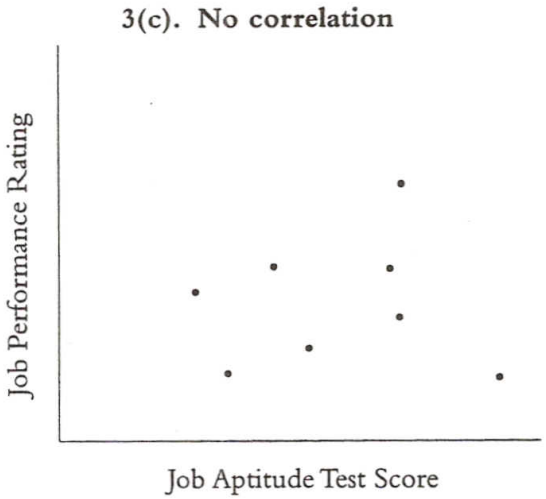


Figura 2

[1] Ver Daniel L. Rubinfeld, Reference Guide on Multiple Regression, en el Reference Manual on Scientific Evidence, 2nd ed., Federal Judicial Center (2000), pp. 204-221. http://www.au.af.mil/au/awc/awcgate/fjc/multi_regression_ref.pdf



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

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

[1] $Y= a + b X$

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

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

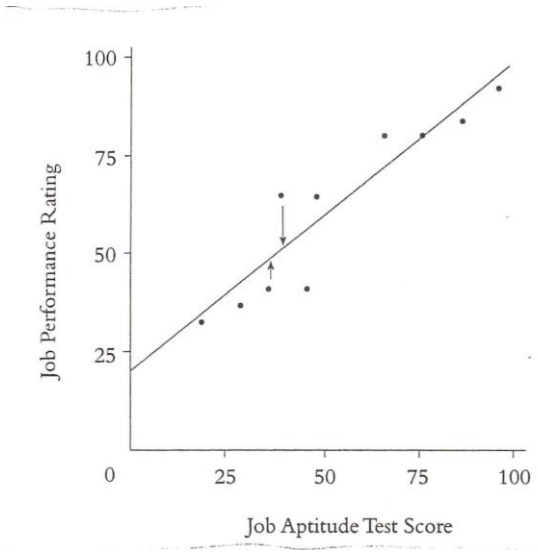


Figura 3. Línea de Regresión

Es típico que la línea de regresión sea estimada usando el método estándar de *mínimos cuadrados ordinarios (MCO)*,[2] donde los valores de "a" y "b" se calculan minimizando la suma de los desvíos al cuadrado de los puntos respecto a la línea de regresión. Así, los desvíos positivos y negativos de igual tamaño son computados de manera similar, mientras que los desvíos amplios cuentan más que los pequeños. En la Figura 3 las líneas de desvío son *verticales porque la ecuación predice la evaluación del rendimiento laboral a partir de los puntajes de las pruebas de aptitud, no los puntajes de las pruebas de aptitud a partir de las evaluaciones del rendimiento laboral.*

Variables importantes que podrían influir sobre la variable dependiente en forma sistemática, cuyos datos puedan ser obtenidos, deberían ser incluidas explícitamente en un modelo estadístico. Las influencias restantes, que pueden ser pequeñas tomadas individualmente, pero que podrían resultar sustanciales en el agregado, se incluyen dentro de un término adicional de

[2] http://es.wikipedia.org/wiki/M%C3%ADnimos_cuadrados

error.[3] La regresión múltiple es un procedimiento que permite separar los efectos sistemáticos (asociados con las variables explicativas) de los efectos aleatorios (asociados con el término de error) y también ofrece un método para evaluar el éxito del proceso llevado a cabo.

2. El Modelo de Regresión Lineal

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

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

En esta forma, Y representa a la variable dependiente (p.ej. el salario de un empleado), y $X_1 \ldots X_k$ a las variables explicativas (p. ej., la experiencia de cada empleado y su sexo, codificado como 1 o 0, resp.). El término de error, ε, representa la influencia colectiva *no observable* de todas las variables omitidas. En una representación lineal, cada uno de los términos adicionados implica parámetros desconocidos, $\beta_0, \beta_1, \ldots, \beta_k$,[4] que son estimados "ajustando" la ecuación a los datos usando mínimos cuadrados.

La mayoría de los estadísticos utilizan la técnica de mínimos cuadrados por su sencillez y sus propiedades deseables. Como resultante, también es utilizada en cuestiones legales.[5]

Ejemplo Un experto desea analizar los salarios de hombres y mujeres en una gran editorial a fin de descubrir si las diferencias de salarios entre los empleados con experiencia similar son evidencia de discriminación.[6] Para empezar con el caso más simple, Y, el salario medido en dólares anuales, es la variable dependiente que debe ser explicada, y X_1 es la variable explicativa – el número de años de experiencia del empleado. El modelo de regresión sería escrito así:

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

En esta ecuación, β_0 y β_1 son los parámetros que deben ser estimados con los datos, y ε es el término del error aleatorio. ¿Cuál es la interpretación de β_0? Es el salario medio de todos los empleados que carecen de experiencia. ¿Y la interpretación de β_1? Mide el efecto promedio que un año de experiencia adicional tiene sobre el salario medio de los empleados.

Una vez que los parámetros de una ecuación de regresión, como la [3], han sido estimados, pueden calcularse los valores "ajustados". Si en la ecuación [3] denotamos a los parámetros de regresión estimados, o *coeficientes de regresión*, como $b_0, b_1, \ldots b_k$, los valores ajustados de Y, que denotaremos como <Y>, vendrán dados por la siguiente ecuación:

[5] $\langle Y \rangle = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$

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

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

[5] En http://www.youtube.com/watch?v=ocGEhiLwDVc hay un documental muy didáctico sobre la técnica de mínimos cuadrados (en inglés).

[6] Los resultados de regresión del ejemplo están basados en datos de 1,715 hombres y mujeres, que fueron usados por el defensor en un caso de discriminación sexual en contra del *New York Times* resuelto en 1978.

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

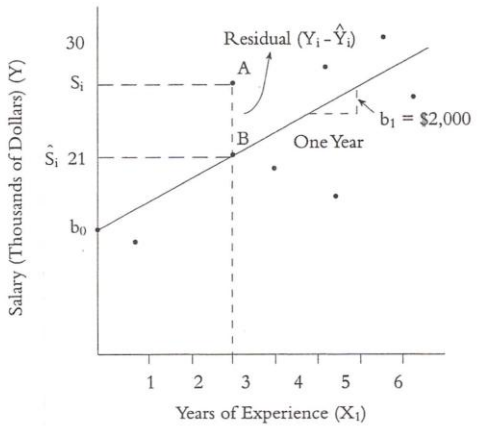


Figura 4. Bondad del Ajuste

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

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

[7] $\langle Y_i \rangle = b_0 + b_{1i}\ X_{1i}$ (punto B).

La ordenada al origen de la línea recta es el valor medio de la variable dependiente cuando la o las variables independientes son iguales a 0; esta ordenada al origen b_0 aparece en el eje vertical de la Figura 4. En forma similar, la pendiente de la línea mide el cambio (promedio) de la variable dependiente asociado al incremento de 1 unidad de la variable explicativa. También está representada la pendiente b_1. Según la ecuación [6], la ordenada al origen igual a $15,000 indica que los empleados inexpertos ganan $15,000 por año. El parámetro de la pendiente implica que cada año de experiencia añade $2,000 al salario de un empleado "promedio".

Ahora supongan que la variable salario está relacionada con el sexo del empleado. La variable relevante indicativa, que a menudo es llamada una variable *dummy*, es ahora X_2, igual a 1 si el empleado es del sexo masculino y 0 si es del sexo femenino. Supongan que la regresión del salario con respecto a X_2 produce el siguiente resultado: $\langle Y \rangle = \$30{,}449 + \$10{,}979\ X_2$. El coeficiente $10,979 mide la diferencia entre el salario medio de los hombres y el salario medio de las mujeres.[7]

3. Residuos de Regresión

Para todo conjunto de datos puntuales, el residuo de regresión es la diferencia entre el valor observado y el valor ajustado de la variable dependiente. Supongan, por ejemplo, que estudiamos el caso de un individuo con 3 años de experiencia y un salario de $27,000. Según la línea de regresión de la Figura 4, el salario medio de un individuo con esa experiencia se ubica en $21,000. Luego, estamos en presencia de un residuo positivo, igual a $6,000. En términos generales, el residuo *e* asociado con un dato puntual como el punto A de la Figura 4, viene dado por $e = Y_i - \langle Y_i \rangle$. Cada punto de la figura tiene un residuo, que es el error cometido por el método de regresión mínimo-cuadrático con ese individuo.

4. No linealidades

Los modelos no lineales toman en cuenta la posibilidad de que la magnitud del efecto de una variable explicativa sobre la variable dependiente cambie a medida que cambia el nivel de la

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

variable explicativa. Hay un modelo útil que produce este efecto, el modelo de interacción entre las variables. Por ejemplo, supongan que:

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

En esta ecuación, S es el salario anual, SEXO es igual a 1 para las mujeres y a 0 para los hombres, EXP representa los años de experiencia laboral, y ε es un término de error. El coeficiente β_2 mide la diferencia del salario medio (para todos los niveles de experiencia) entre hombres y mujeres que no tienen experiencia. El coeficiente β_3 mide el efecto de la experiencia sobre el salario de los hombres (cuando SEXO=0), y el coeficiente β_4 mide la diferencia en el efecto de la experiencia sobre el salario de hombres y mujeres. Se desprende, por ejemplo, que el efecto de un año de experiencia sobre el salario de los hombres es β_3, mientras que el efecto comparativo para las mujeres es $\beta_3 + \beta_4$.[8]

5. Interpretación de los Resultados de una Regresión

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

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

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

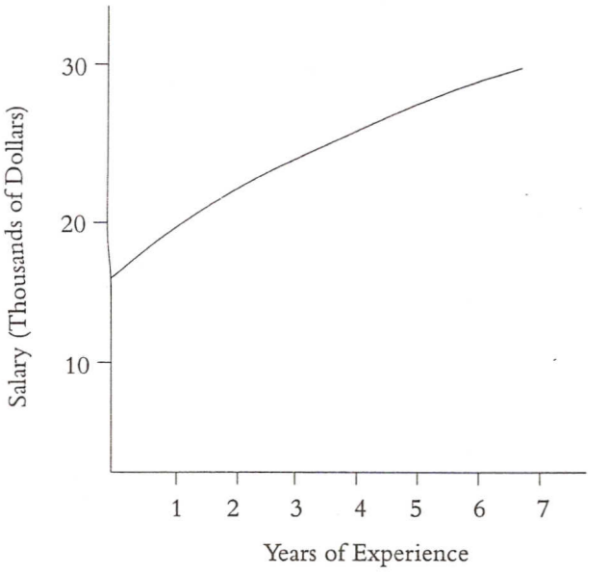


Figura 6. Pendiente de Regresión

Consideren ahora la interpretación de las dos variables de experiencia, X_1 y X_3. El signo positivo de X_1 muestra que el salario crece con la experiencia. El signo negativo de X_3 indica que la tasa de crecimiento del salario disminuye con la experiencia. Para calcular el efecto combinado de X_1 y X_3 podemos hacer algunos cálculos: por ejemplo, veamos cómo cambia el salario medio de las mujeres (X_2=0) a medida que cambia el nivel de experiencia. A medida

[8] La estimación de una ecuación en la cual hay términos de interacción para todas las variables explicativas, como en la [8], es esencialmente lo mismo que estimar dos regresiones por separado, una para los hombres y otra para las mujeres.

que la experiencia crece desde 0 a 1 año, el salario medio crece en $2,251, desde $14,085 a $16,336. Sin embargo, las mujeres con 2 años de experiencia ganan sólo $2,179 más que las mujeres con 1 año de experiencia, y las mujeres con 3 años de experiencia sólo ganan $2,127 más que las que tienen 2 años de experiencia. Además, las que tienen 7 años de experiencia ganan $28,582 por año, que sólo representa $1,855 más que los $ 26,727 ganados por las mujeres con 6 o más años de experiencia.[9] La Figura 6 ilustra estos resultados; la línea de regresión representada corresponde a los salarios de las mujeres; la línea correspondiente a los hombres sería paralela y más elevada en $1,675.

6. Resultados de Regresión por MCO

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

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

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

Bajo supuestos adecuados, los estimadores mínimo-cuadráticos son las “mejores” determinaciones de los verdaderos parámetros subyacentes.[10] De hecho, los mínimos cuadrados tienen diversas propiedades deseables. Primero, los estimadores mínimo-cuadráticos son

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

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

insesgados. Esto significa, intuitivamente, que si la regresión fuera calculada una y otra vez con muestras diferentes, la media de los diversos estimadores de cada coeficiente obtenidos sería el verdadero parámetro. Segundo, los estimadores mínimo-cuadráticos son *consistentes*; si la muestra fuera muy grande, los estimadores estarían próximos a los verdaderos parámetros. Tercero, los estimadores mínimo-cuadráticos son *eficientes*, en el sentido de que los estimadores tienen la menor varianza de todos los posibles estimadores (lineales) insesgados.[11]

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

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

[10] $b \pm 1.96.$ (SE de b).[12]

El experto puede contrastar la hipótesis de que el parámetro en realidad es 0 (llamada la hipótesis nula) mirando el estadístico-*t*, definido como:

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

Si este estadístico-*t* resulta de magnitud inferior a 1.96, el intervalo de confianza al 95% alrededor de *b* debe incluir 0.[13] Como esto significa que el experto no puede rechazar la hipótesis de que β=0, el estimador – cualquiera resulte su valor – se dice que no es estadísticamente significativo. Recíprocamente, si el estadístico-*t* es mayor que 1.96 en valor absoluto, el experto concluye que es improbable que el verdadero valor de β sea 0 y dice que este estimador es estadísticamente significativo (intuitivamente, *b* está "demasiado lejos" de 0 como para que sea consistente con β=0). En tal caso, el experto rechaza la hipótesis de que β=0 sea verdadera y dice que el estimador es estadísticamente significativo. Si la hipótesis nula β=0 es verdadera, usar un intervalo de confianza al 95% implicará que el experto rechace erróneamente la hipótesis nula 5% de las veces. Por consiguiente, decimos que los resultados son significativos al 5%.[14]

---

[11] http://en.wikibooks.org/wiki/Econometric_Theory/Assumptions_of_Classical_Linear_Regression_Model

[12] Ya sabemos que los intervalos de confianza son comúnmente utilizados en los análisis estadísticos, porque el experto nunca puede estar seguro de que el estimador del parámetro sea igual al verdadero parámetro de la población.

[13] Este estadístico-*t* es aplicable a muestras de cualquier tamaño. A medida que la muestra se hace más grande, la distribución subyacente, que es la fuente del estadístico-*t* (la *distribución t de Student*) se va aproximando a la distribución normal. http://en.wikipedia.org/wiki/Student's_t-distribution

[14] Un estadístico-*t* de magnitud igual o mayor a 2.57 está asociado a un nivel de confianza del 99%, o del 1%, que incluye una banda igual a 2.57 desvíos estándar a ambos lados de los coeficientes estimados.

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

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

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

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

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

[13] $1,675 ± $1,435 x 1.96 = $1,675 ± $2,813.

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

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

La información proporcionada por los resultados de regresión contiene no sólo los estimadores puntuales de los parámetros y sus errores estándar o estadísticos-*t*, sino además otra información que nos dice cuán buena es la aproximación de la línea de regresión a los datos. Hay un estadístico, llamado el error estándar de regresión (SER) que es un estimador del tamaño promedio de los residuos de regresión.[15] Un SER=0 significaría que *todos los puntos de los datos yacen exactamente sobre la línea de regresión* – lo cual es algo prácticamente imposible. A otras cosas iguales, a mayor SER, peor será el ajuste de los datos del modelo.

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

El estadístico R^2 mide el porcentaje de variación de la variable dependiente que es tenido en cuenta por todas las variables explicativas.[16] Luego, R^2 facilita una medida global de bondad del ajuste. Su valor oscila entre 0 y 1. Un $R^2=0$ significa que las variables explicativas no explican nada de la variación de la variable dependiente.; si $R^2=1$, las variables explicativas explican toda la variación. El R^2 de la ecuación [12] es .56, lo cual implica que las tres variables explicativas dan cuenta del 56% de la variación de los salarios.

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

[16] Hay que tener en cuenta que R^2 y SER dan aproximadamente la misma información, ya que R^2 es más o menos igual a 1 – SER^2/Varianza de Y. La variación es computada como el cuadrado de la diferencia entre cada Y y el Y medio, sumado a lo largo de todas las observaciones.

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

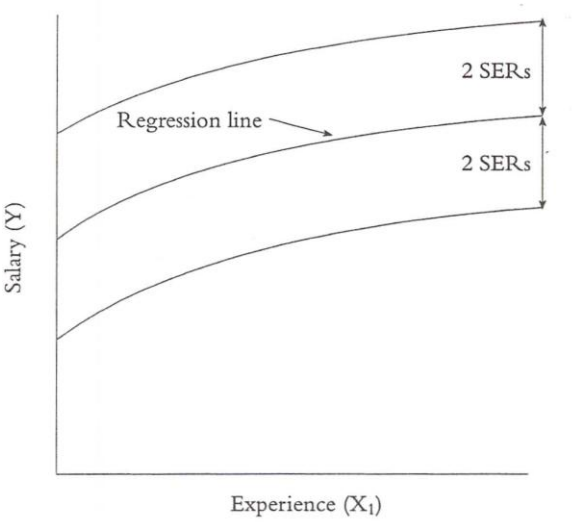


Figura 7. Error Estándar de Regresión

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

La línea de regresión mínimo-cuadrática puede ser sensible a los puntos extremos. Esto se puede apreciar en la Figura 8. Supóngase que inicialmente hay tres puntos (A, B, y C), que vinculan la información de la variable X_1 con la variable Y. La línea 1 representa la mejor regresión entre estos puntos. El punto D es un valor atípico porque se encuentra muy alejado de la línea de regresión que ajusta a los puntos restantes. Si se re-estima la línea de regresión mínimo-cuadrática incluyendo ahora el punto D, se obtiene la Línea 2. Esta figura muestra que D es un dato influyente, ya que tiene un efecto dominante tanto sobre la pendiente como sobre la ordenada al origen de la línea de mínimos cuadrados. Como en mínimos cuadrados se trata de minimizar la suma de los desvíos al cuadrado, la sensibilidad de la línea a estos puntos individuales puede ser a veces sustancial.[17]

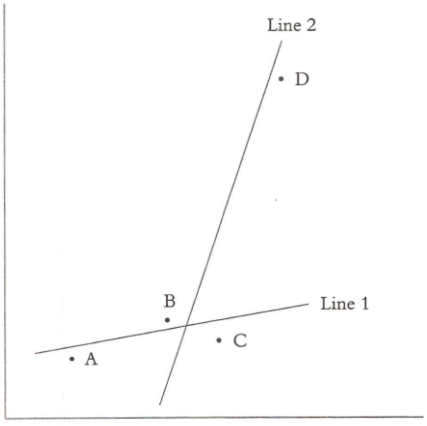


Figura 8. Línea de Regresión Mínimo-Cuadrática

Lo que torna más difícil el problema de los datos influyentes es que el efecto de un valor atípico puede no ser apreciado cómodamente si el desvío es medido a partir de la línea de regresión final. El motivo es que la influencia del punto D sobre la línea 2 es tan sustancial que *su desvío de la línea de regresión no es necesariamente mayor que el desvío de los puntos restantes de la línea*

[17] No siempre es indeseable la insensibilidad, por ejemplo cuando es mucho más importante predecir un punto como D cuando se produce un gran cambio, en comparación con medir los cambios pequeños.

de regresión.[18] Aunque no son tan populares como los mínimos cuadrados, hay otras técnicas de estimación alternativas menos sensibles a los valores atípicos, como la regresión robusta.[19]

7. Lectura del Output de Computadora de Regresión Múltiple

Como aplicación, vamos a analizar los resultados de Ernesto Gaba y Lucio Reca de estimación de la demanda de carne vacuna en Argentina entre 1950 y 1972,[20] que construyeron un índice de precios de un conjunto de alimentos que componen una canasta suficientemente amplia de posibles sustitutos de la carne vacuna e incorporaron el posible efecto de la veda sobre el consumo mediante una variable binaria que adopta el valor 1 en los años en que rigió algún tipo de veda y 10 en los restantes. Para otra variable explicativa – el ingreso por habitante – usaron el ingreso medio de la población a partir de las series de producto bruto nacional y de población. Finalmente, el precio de la carne vacuna fue introducido en el modelo de dos formas alternativas: 1) deflactado por el índice de precios implícitos en el producto; 2) como precio relativo a los sustitutos. Reca y Gaba optaron por usar variables expresadas en logaritmos (lo que confiere a las derivadas parciales de la función demanda de carne el sentido de elasticidades), donde:

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

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

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

[18] La importancia de un valor atípico también depende de su ubicación en el espacio de datos. Es probable que los valores atípicos asociados con valores relativamente extremos de las variables explicativas sean muy influyentes. Ver, p.ej., Fisher v. Vassar College, 70 F.3d 1420, 1436 (2d Cir. 1995) (el tribunal requirió evaluar el "servicio en la comunidad académica", porque el concepto era demasiado amorfo y no constituía un factor significativo para revisar la continuidad de la cátedra), rev'd on other grounds, 114 F.3d 1322 (2d Cir. 1997) (en banc). http://openjurist.org/70/f3d/1420/fisher-v-vassar-college

[19] http://en.wikipedia.org/wiki/Robust_regression

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

La veda, representada en el modelo por la variable binaria, muestra una elevada significación estadística y una considerable estabilidad en todas las regresiones. La primera veda al consumo interno fue en 1952, y si bien fue suspendida 3 años después, de hecho cesó de funcionar aproximadamente al año y medio de ser aplicada. La segunda veda (1964 y 1965) introdujo la restricción en la veda de carne vacuna 2 días por semana, y la última, a partir de marzo de 1971, si bien con interrupciones y contramarchas, se caracterizó por fijar la duración del período de veda en una semana, hasta su derogación en abril de 1973. Los estímulos para la faena clandestina de ganado tal vez incidan en una sobreestimación del coeficiente de la variable binaria, sin embargo. Es decir, que las cifras tomadas para confeccionar el análisis de regresión podrían ser inferiores a las reales y, por consiguiente, la variable binaria recoge además del efecto propio de la veda la subestimación proveniente del consumo no registrado. En cuanto a la magnitud del coeficiente, el promedio simple de la diferencia de consumo con veda y sin veda en los años 1952, 1964 y 1971 arroja una reducción de consumo por habitante y año de 5,7 kg de carne. En otras ecuaciones, la reducción del consumo imputable a la veda arroja 9,5 kg por habitante y por año. Lo cual implica una caída del consumo ligeramente superior a 10%. Para que se logre un efecto similar dejando operar al sistema de precios sin interferencias, dada la elasticidad propia de la demanda de carnes, hubiera sido necesario un aumento del precio de la carne de 35%.

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

[15] $\log (v) = 1{,}245 + 0{,}159\ \text{veda} - 0{,}369 \log (pv) + 0{,}048 \log (ps) + 0{,}38 \log (y)$

(1,3) (6,4) ** (8,9) ** (0,07) (3,8) **

$R^2_{aj} = 0{,}93$

SER = 0,033

DW = 2,03

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

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

conllevó nuevos plazos de pago y condiciones de compra que se extendieron a toda la cadena de comercialización, llegando inclusive al productor que vio reducida su rentabilidad. En la actualidad se estima que entre 35 y 40% de la venta de carnes se concentra en este canal y sólo en Capital Federal representa 60% de la comercialización de carnes vacunas. Hay supermercados que tienen sus propias plantas de faena, lo que les brinda un mayor control del negocio hacia adelante y hacia atrás en la cadena y también elementos clave como la seguridad en la cadena de frío, calidad continua (contratos con los productores en forma directa), mejor distribución de los cortes por zona de consumo, venta de cortes con marca propia y diferente presentación, etc. La comercialización tradicional de carne en el mercado interno se basaba en el traslado de la media res salida de la planta de faena y transportada hacia cada boca de expendio en camiones refrigerados. Esta modalidad a principios de los 1990s representaba 95% y el resto se basaba en la distribución de cortes, pero esta modalidad de venta mostró un fuerte cambio ya que la relación pasó de 95 % a 75 %. Ambos mercados constituyen mundos muy diferentes, en el mercado de la media res las categorías más utilizadas son vaquillona, ternero y novillito colocadas directamente en el local de expendio para su desposte, siendo la modalidad utilizada por los frigoríficos consumeros. De esta forma se incrementaron los beneficios obtenidos en los mercados de alto valor. Estos frigoríficos proveen de carne a supermercados e hipermercados, algunos a través del ingreso a una nómina de proveedores continuos con especificaciones en cuanto a calibre de los cortes; los plazos más comunes oscilan entre 30 y 45 días. Los operadores del mercado interno son los matarifes carniceros.

8. Proyecciones

Volvamos a nuestro ejemplo salarial. Una proyección es una predicción sobre los valores que alcanzará una variable dependiente usando información sobre las variables independientes. Frecuentemente lo que se hace es una proyección ex ante: en tal caso, los valores de la variable dependiente son predichos más allá de la muestra (es decir, más allá del período en que el modelo fue estimado). Empero, las proyecciones ex post son utilizadas a menudo en el análisis de daños.[21] Una proyección ex post se caracteriza porque en el horizonte de proyección todas las variables dependientes y explicativas son conocidas; las proyecciones ex post pueden ser chequeadas comparando con los datos existentes y facilitar un método directo de evaluación.

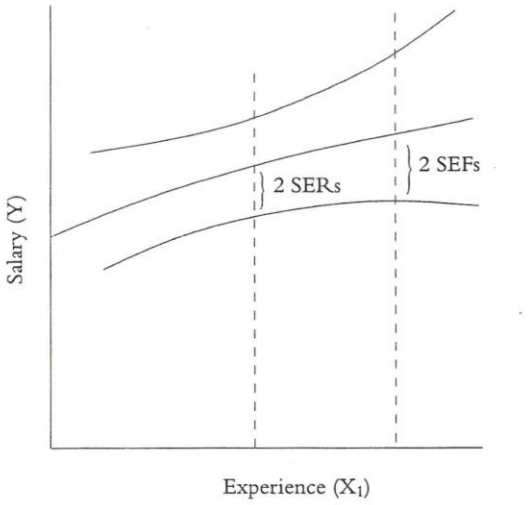


Figura 9. Error Estándar de Proyección

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

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

Para predecir el salario de un hombre con 2 años de

[21] En los casos que implican daños, surge la pregunta frecuente de cómo hubiera sido el mundo si cierto evento no hubiera sucedido. Por ejemplo, en un caso de fijación de precios anti-monopolística, el experto puede preguntarse cuál hubiera sido el precio de un producto si cierto evento asociado con la fijación de precios no hubiera ocurrido. Si los precios hubieran sido más bajos, la evidencia sugiere un impacto. Si el experto puede predecir cuán bajos hubieran podido llegar a ser, los datos pueden ayudarlo a desarrollar un estimador numérico de los daños.

experiencia, el experto calcula:

[17] ‹Y (2) › = $14,085 + ($2,323x2) + ($1,675) – ($36x2) = $20,262

El grado de precisión de las proyecciones ex ante y ex post puede calcularse si la especificación del modelo es correcta y los errores están distribuidos normalmente y son independientes. Al estadístico se lo conoce como error estándar de proyección (SEF, por *standard error of forecast*). El SEF mide el desvío estándar del error de proyección cometido dentro de una muestra de la que se conocen con certeza las variables explicativas.[22] El SEF puede usarse para determinar cuán precisa es una proyección. En [17] el SEF asociado a la proyección es aproximadamente $5,000. Si se utiliza una muestra amplia, la probabilidad es, groseramente, de 95%, de que el salario predicho esté comprendido entre 1.96 errores estándar del valor proyectado. En tal caso, el intervalo apropiado al 95% va de $10,822 a $30,422. Como el modelo estimado no explica efectivamente los salarios, el SEF es amplio, como lo es el intervalo al 95%. Un modelo más completo con variables explicativas adicionales resultará en un SEF más reducido y un intervalo al 95% más pequeño de predicción.

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

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

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