Apéndice B

¿Qué es una regresión lineal?

José Miguel Benavente

I. INTRODUCCIÓN

En varios capítulos de este libro se ocupan regresiones lineales y se afirma que el coeficiente de regresión indica cuánto varía, *en promedio*, una variable cuando otra variable también varía. Por ejemplo, considérese el Gráfico 1, tomado de la página 492 del capítulo 17 de Andrea Tokman, que muestra la densidad residencial (eje vertical) y la distancia entre la comuna de Santiago y cada una de las seis comunas de ingresos altos ubicadas hacia el oriente del Gran Santiago (eje horizontal). Tokman le ajustó a estos puntos una línea de regresión exponencial

$$\ln(\text{densidad residencial}) = D + \beta \times (\text{distancia al centro}), \qquad (1)$$

donde ln es la abreviatura de logaritmo natural[1]. Obtuvo que

$$\ln(\text{densidad residencial}) = 5{,}44 - 0{,}079 \times (\text{distancia al centro}),$$

la línea de regresión que aparece en el Gráfico 1 y que es resumida en el Cuadro 1[2]. Así, concluyó que, en promedio, la densidad residencial cae 7,9 por ciento por cada kilómetro que uno se aleja del centro y que la relación es "estadísticamente significativa". Por contraste, cuando le ajustó la misma curva de regresión a las restantes 28 comunas del Gran Santiago, cuyos hogares ganan ingresos más bajos (véase el Gráfico 2), encontró que

$$\ln(\text{densidad residencial}) = 5{,}15 - 0{,}001 \times (\text{distancia al centro}),$$

Vale decir, en promedio, la densidad cae apenas 0,1 por ciento por cada kilómetro que la comuna se aleja del centro. Además, Tokman concluyó que la relación no era "estadísticamente significativa".

---

[1] En el Recuadro 1, página 84 del capítulo 3, Marcial Echenique explica la función exponencial y el significado del coeficiente β, la así llamada gradiente de la densidad.

[2] Para facilitar la comparación con los gráficos que presenta Tokman, la línea de regresión que muestran los gráficos es (densidad residencial) = $\exp(D) \times \exp(\beta \times \text{distancia al centro})$, donde $\exp(X)$ es la función exponencial, el número e = 2,71828... elevado a X.

Gráfico 1 Densidad y distancia al centro en comunas de ingresos altos

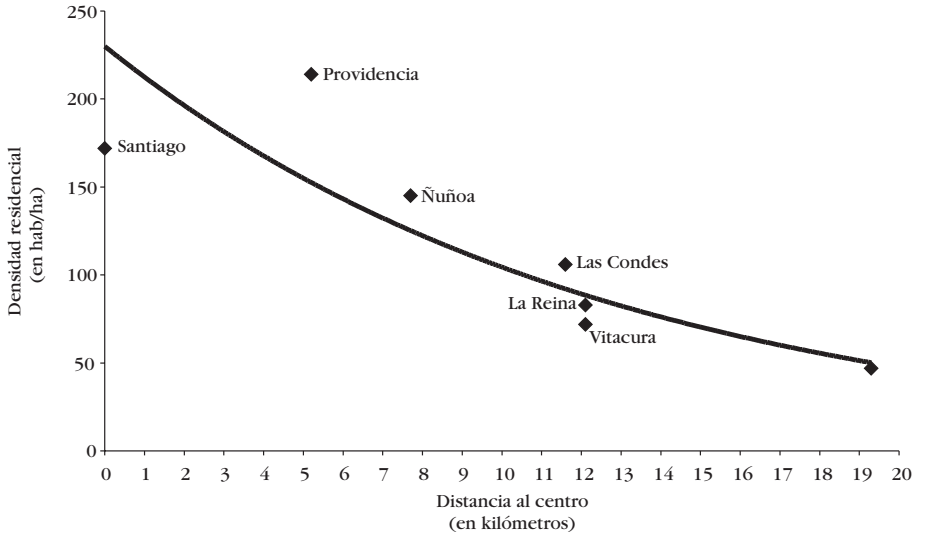
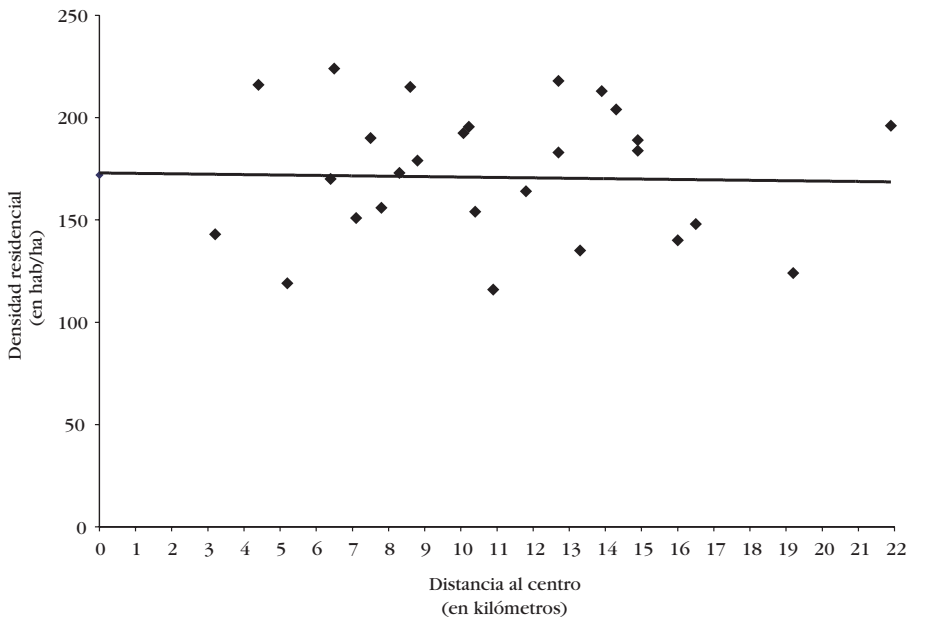


Gráfico 2 Densidad y distancia al centro en el resto de la ciudad



En este apéndice explicaré brevemente qué es una regresión lineal y cuál es el sentido preciso de la afirmación que "indica la variación promedio de una variable cuando otra varía" (sección II). Luego discutiré el concepto de significancia estadística y cómo una regresión nos permite distinguir a las relaciones sistemáticas de aquellas que se deben sólo al azar (sección III).

Cuadro 1 Las regresiones presentadas en los gráficos

	(1) Ingresos altosResto	(2)
Intercepto (D)	5,44 (0,18) [0,00]	5,15 (0,09) [0,00]
Intercepto [exp(D)][1]	230,79	173,14
Densidad residencial (β)	–0,079 (0,016) [0,00]	–0,001 (0,008) [0,87]
n	7	28
Grados de libertad ($n - 2$)	5	26
R^2	0,83	0,00
r_{xy}	–0,91	–0,03

(Desviación estándar entre paréntesis) [Estadístico p entre corchetes]

Nota: (1) En cada caso el intercepto es el resultado de elevar el número e = 2,71828... a D. Así 230,79 = exp(5,44) y 173,14 = exp(5,15).

Antes de seguir, dos advertencias. Primero, la sección II es suficiente para entender por qué el coeficiente de una regresión es una variación promedio y el lector que no esté interesado en la inferencia estadística puede detenerse ahí. Segundo, en lo que sigue examinaré sólo una regresión de dos variables. La extensión a regresión múltiple no cambia la mayoría de lo que diré, pero en todo caso sobrepasa el ámbito de esta nota.

II. LA LÍNEA DE REGRESIÓN COMO UN PROMEDIO

Una simple mirada al Gráfico 1 basta para darse cuenta de que la densidad residencial de una comuna de ingresos altos es, "en general", menor mientras más alejada esté de la comuna de Santiago. Sin embargo, también es claro que esta relación es inexacta. Por ejemplo, en Providencia se vive más densamente que en Santiago. Y la densidad difiere entre Las Condes (106 hab/ha), La Reina (83 hab/ha) y Vitacura (72 hab/ha), a pesar de que la distancia al centro de las tres comunas es 12 km. ¿Cómo incluir en un solo número la variación promedio y la fortaleza de la asociación entre las dos variables?

La respuesta es la línea de regresión. Una regresión se ocupa para estudiar la relación entre una variable dependiente y una o más variables explicativas. En este caso la variable dependiente es el logaritmo natural de la densidad residencial y la variable explicativa es la distancia desde la comuna al centro. La línea de regresión permite estimar o predecir el promedio de la variable dependiente para valores fijos de las variable explicativa. Por lo

mismo, también permite calcular cuánto varía en promedio la variable dependiente para variaciones dadas de la variable independiente. Ése es el sentido de la afirmación, repetida en varios capítulos de este libro, de que "la densidad cae en promedio 7,9 por ciento por cada kilómetro que la comuna se aleja del centro".

¿Cómo se obtienen los coeficientes de regresión, en este caso D y β? Si bien existen varios métodos, lejos el más popular es el de *mínimos cuadrados ordinarios* o MCO. Éste elige los coeficientes D y β para minimizar la suma de las diferencias al cuadrado entre el valor observado de la variable dependiente y el valor predicho por la línea de regresión. Vale decir, si y es el logaritmo natural del valor observado de la densidad residencial en una comuna distante x kilómetros del centro, y $\hat{y} \equiv D + \beta \times x$ es el valor predicho por la línea de regresión, se le ajusta la línea de regresión que minimiza

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left[y_i - (D + \beta \times x_i)\right]^2,$$

donde n es el número de observaciones. A la diferencia $\hat{u}_i \equiv y_i - \beta \times \hat{y}_i$ se le suele llamar el *error* entre el valor observado de la variable y y el valor predicho por la línea de regresión. Así, el método de los mínimos cuadrados elige los coeficientes D y β de forma que la suma de estos errores elevados al cuadrado sea lo más baja posible.

Se puede demostrar que el resultado de esta minimización es

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2}$$

donde $\bar{x}$ es el promedio de la variable explicativa y $\bar{y}$ es el promedio de la variable dependiente; al mismo tiempo, $\hat{D} = \bar{y} + \beta \times \bar{x}$.

Una transformación algebraica de la fórmula (2) permite apreciar desde un ángulo distinto en qué sentido una línea de regresión resume un promedio. En efecto, se define al coeficiente de correlación entre x e y como

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

donde s_x es la desviación estándar de x y s_y la de y [3]. Así, (2) se puede reescribir como

$$\hat{\beta} = r_{xy} \cdot \frac{s_y}{s_x}. \tag{3}$$

[3] Por construcción $1 \leq r_{xy} \leq 1$. Cuando la correlación es perfecta y la relación entre x e y es exactamente lineal, el coeficiente de correlación es ya sea 1 ó –1. A medida que el coeficiente se aleja de los extremos la relación lineal se atenúa y desaparece cuando el coeficiente es 0.

La expresión (3) indica que, en promedio, si la variable explicativa x cambia en una desviación estándar, la variable dependiente y cambiará r_{xy} veces por su desviación estándar, s_y. Así, la magnitud del coeficiente de regresión también se puel cociente de sus desviaciones estándar; y, por el otro lado, la fortaleza de la asociación lineal entre ellas. Una relación cercana a lineal aumenta la magnitud del coeficiente de regresión, mientras que una relación muy alejada de lineal la reduce hasta que desaparece cuando $r_{xy} = 0$.

De lo anterior también se desprende una forma alternativa y sencilla de verificar cuán bueno es el coeficiente $\hat{\beta}$ para medir la forma en que x afecta a y. En efecto, si la suma de los cuadrados de los errores, $\sum_{i=1}^{n} \hat{u}_i^2$, es cercana a cero, el ajuste es muy bueno y se puede afirmar que el coeficiente $\hat{\beta}$ explica relativamente bien la relación entre x e y. Por el contrario, si esta suma es grande, el ajuste es malo y el coeficiente $\hat{\beta}$ explica poco.

Por supuesto, qué tan "grande" o "pequeña" sea esta suma depende de la escala de la variable dependiente y. Por eso, es más conveniente evaluar cuánto de la volatilidad o varianza observada en la variable y es explicada por el modelo lineal. A partir de esta observación se construye un indicador muy usado como grado de ajuste, el así denominado R^2, el cual es igual a

$$1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \qquad (4)$$

Éste indica qué fracción de la variabilidad de y, la variable dependiente, es explicada por la regresión. Si gran parte de la varianza de y es explicada por la multiplicación de x por β, entonces el ajuste es relativamente bueno, de lo contrario el modelo no presenta un buen ajuste. Obviamente, $R^2 = 1$ si $\sum_{i=1}^{n} \hat{u}_i^2 = 0$, y se puede demostrar que es igual a cero si $\hat{\beta} = 0$.

Ahora podemos comparar la relación entre densidad residencial y distancia al centro, reportadas en los gráficos y el Cuadro 1, a la luz de la expresiones (3) y (4). Como ya se vio, esta relación es bastante fuerte en las comunas de ingresos altos, pues, en promedio, la densidad residencial cae 7,9 por ciento por cada kilómetro que la comuna se aleja del centro. Esto no debiera sorprender porque, tal como se aprecia en el Cuadro 1, la correlación entre x e y es –0,91. Por el contrario, la relación entre la densidad residencial y la distancia al centro en el resto de las comunas de Santiago es tenue, pues por cada kilómetro que la comuna se aleja del centro la densidad cae apenas 0,1 por ciento. Nuevamente, esto no debiera sorprender: la correlación en este caso apenas es –0,03. Alternativamente, la magnitud de la relación se puede apreciar por los R^2 respectivos. Se aprecia en el Cuadro 1 que el R^2 de la regresión de las comunas de ingresos altos es igual 0,83, mientras que el del resto de las comunas de Santiago es casi 0.

III. LA LÍNEA DE REGRESIÓN COMO MEDIO PARA LA INFERENCIA ESTADÍSTICA

Es claro que, en promedio, la densidad residencial de una comuna de ingresos altos es menor mientras más lejos está del centro. Por otro lado, la regresión indica que, en promedio, la densidad residencial de una comuna de ingresos más bajos es levemente menor si está más alejada del centro, pero algo cae. ¿Qué tan robusta es la relación en uno y otro caso? A continuación mostraré cómo una regresión lineal también se puede utilizar para inferir estadísticamente si el promedio estimado realmente es el resultado de una relación sistemática o, por el contrario, simplemente efecto del azar.

El núcleo del problema es inferir qué tan probable es que el promedio de y varíe si lo hace x. Si los datos muestran que, para un valor dado de x los y observados tienden a ser similares; y éstos varían cuando lo hace x, entonces se puede inferir que x afecta sistemáticamente a la media de y. Por el contrario, si los datos muestran que para un mismo valor de x los valores de y son muy distintos y su media poco cambia cuando x lo hace, entonces quiere decir que ambas variables tienen poco o nada que ver una con otra.

Esto se puede apreciar comparando los gráficos 1 y 2. Como ya se vio, es claro que la densidad residencial disminuye sistemáticamente a medida que una comuna de ingresos altos se aleja del centro. Hay cierta variación entre Las Condes, La Reina y Vitacura, pero cualquiera es menos densa que Ñuñoa o Providencia y más densa que Lo Barnechea.

Por el contrario, a pesar de que el coeficiente de regresión en el Gráfico 2 es negativo, es posible encontrar comunas con densidad residencial más alta o más baja que los 172 hab/ha del centro a casi cualquier distancia. Más aún, el intercepto estimado, 173,14 hab/ha, es muy parecido al promedio simple de la densidad residencial de las 28 comunas (172,98 hab/ha) y sorprendentemente parecido a la densidad residencial en el centro (173 hab/ha). Pareciera que no existe relación sistemática entre densidad residencial y distancia al centro en el resto de las comunas de Santiago.

La impresión que se obtiene de los gráficos se puede confirmar siguiendo un procedimiento formal. Todo parte del modelo estadístico de regresión lineal, según el cual la densidad residencial observada para cada comuna i es

$$y_i = D + \beta \times x_i + \varepsilon_i .$$

$D + \beta \times x_i$ es la densidad residencial promedio cuando la distancia de la comuna al centro es $x = x_i$ kilómetros. Por otro lado, la desviación del promedio, ε, se supone independiente y distribuida normalmente con media 0 y desviación estándar σ_ε. Un poco de álgebra basada en las reglas de promedios de variables aleatorias muestra que los estimadores de mínimos cuadrados ordinarios, $\hat{D}$ y $\hat{\beta}$, son insesgados (vale decir, su esperanza es D y β). Más aún, $\hat{D}$ y $\hat{\beta}$ se distribuyen normalmente con desviaciones estándar $\sigma_{\hat{D}}$ y $\sigma_{\hat{\beta}}$, las que se pueden estimar a partir de las observaciones de x y de y[4]. Así, los coeficientes $\hat{D}$ y $\hat{\beta}$ estimados se

[4] Una forma general del teorema central del límite implica que, si la muestra es suficientemente grande, las distribuciones de $\hat{D}$ y $\hat{\beta}$ son aproximadamente normales aún si los ε_i no lo son.

pueden entender como el resultado de una muestra aleatoria, y se pueden usar para testear si la variable explicativa *x* afecta en forma sistemática a la variable dependiente *y*. ¿Cómo?

Consideremos primero la regresión de las comunas de ingresos altos, la que arrojó $\hat{\beta} = -0{,}079$. Supongamos ahora que el verdadero coeficiente $\hat{\beta}$ es igual a cero. Si así fuera, ¿qué tan probable sería que una muestra aleatoria cualquiera dé por resultado $\hat{\beta} = -0{,}079$? De manera similar, supongamos que el verdadero coeficiente β es igual a cero en el caso del resto de las comunas de Santiago. ¿Qué tan probable será que una muestra aleatoria cualquiera dé por resultado $\hat{\beta} = -0{,}001$?

Teoremas estándar de la estadística indican que si es verdadera la así llamada *hipótesis nula* que $\beta = 0$, entonces $t = \hat{\beta} / \sigma_{\hat{\beta}}$ es una variable aleatoria que se distribuye normal con media 0 y varianza 1. Además, se puede demostrar que la probabilidad de que *t* sea mayor que 1,96 o menor que –1,96 es 0,05 ó 5 por ciento. Así, si la razón $\hat{\beta} / \sigma_{\hat{\beta}}$ cae más allá de uno de estos dos valores críticos se dice que "el coeficiente β es significativamente distinto de cero con un nivel de confianza de 5 por ciento". Por el contrario, si la razón cae entre –1,96 y 1,96 la hipótesis nula de que $\beta = 0$ se acepta "con un nivel de confianza de 95 por ciento".

Por supuesto, en la práctica nunca conocemos $\sigma_{\hat{\beta}}$. Por eso, para computar el estadístico *t* se ocupa una estimación de la desviación estándar de $\hat{\beta}$ obtenida del mismo análisis de regresión. La desviación estándar estimada del estimador $\hat{\beta}$, llamémosla $s_{\hat{\beta}}$, es igual a

$$\frac{s}{\sqrt{\sum (x_i - \bar{x})^2}},$$

con $s = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}}$. Éste es el estadístico reportado entre paréntesis en el Cuadro 1, y es igual a 0,016 en el caso de las comunas de ingresos altos y 0,008 en las comunas del resto de Santiago.

Se puede demostrar que $\hat{\beta} / s_{\hat{\beta}}$ se distribuye de acuerdo con la distribución *t* de Student con $n - 2$ grados de libertad. En el caso de la regresión de las comunas de ingresos altos, los grados de libertad son 5, $\hat{\beta} / s_{\hat{\beta}} = 4{,}94$ y los límites del intervalo crítico son –2,571 y 2,571. Por lo tanto, se rechaza la hipótesis nula; $\hat{\beta}$ es significativamente distinto de cero al nivel de confianza de 5 por ciento. Por otro lado, en la regresión de las comunas del resto de Santiago los grados de libertad son 26, $\hat{\beta} / s_{\hat{\beta}} = 0{,}125$ y los límites del intervalo crítico son –2,056 y 2,056. $\hat{\beta} / s_{\hat{\beta}}$ cae claramente dentro de este intervalo y, por lo tanto, no se puede rechazar la hipótesis nula que $\beta = 0$. Se concluye, por tanto, que no existe una relación sistemática entre la densidad residencial y la distancia entre la comuna y el centro.

De manera similar, se puede ser más o menos estricto con la hipótesis nula. El estadístico *p*, reportado entre corchetes en el Cuadro 1, indica el máximo nivel de confianza tal que se rechaza la hipótesis nula de que $\beta = 0$. Se puede apreciar que este estadístico marca 0,00 en el caso de las comunas de ingresos altos y 0,87 cuando se trata del resto de las comunas de Santiago. Vale decir, la probabilidad de que $\hat{\beta} = -0{,}079$ sea resultado del azar es casi 0, mientras que es 0,87 u 87 por ciento en el caso de las comunas del resto de Santiago. ■