21. Las relaciones entre variables: parte 3

La inferencia en la regresión

Como en todos los análisis estadísticos, hay dos maneras de pensar en los resultados.

1) La primera es considerar que los resultados son meramente un resumen descriptivo de los datos. Por tanto, un diagrama de tallos y hojas es una descripción de la distribución de una variable, una media es un valor de resumen para el centro de la distribución y una recta es una descripción simple de la relación entre observaciones sobre dos variables.

2) La segunda es pensar en los datos que tenemos como una muestra aleatoria de una población más amplia y en este caso usamos las observaciones para extraer algunas conclusiones sobre la población. Al estimar una media muestral, por tanto, obtenemos un intervalo de confianza en el que cae la verdadera media poblacional o contrastamos una hipótesis específica sobre la media poblacional. En el análisis de regresión tenemos la misma situación. Si nuestras observaciones emparejadas son una muestra aleatoria extraída de una población mayor, entonces podemos utilizar los resultados para realizar ciertas inferencias de las relaciones entre las dos variables en la población.

Nuestro interés principal es saber si el coeficiente de la pendiente de regresión resulta significativo o no, lo que es una prueba de la significación de la relación lineal entre y y x.

En este módulo sobre relaciones entre variables aprenderéis:

- Qué población suponemos al realizar un análisis de regresión.
- Qué es el error estándar de la pendiente.
- Cómo se lleva a cabo un contraste de hipótesis sobre la pendiente.

El uso descriptivo de la regresión

En los ejemplos de ajustar rectas que hemos dado hasta ahora, la recta ha servido de descripción de la relación entre las dos variables, o de resumen de la tendencia de los puntos. Nos centramos mayoritariamente en la pendiente de la recta que resume el incremento (o reducción) de la variable de respuesta y por cada cambio de unidad en la variable explicativa x. En el capítulo consi-

deramos los datos muestreados a partir de una población y, por tanto, los puntos decisivos de la significación estadística del modelo de regresión.

La población a partir de la que se toma una muestra

Hasta ahora en nuestra inferencia estadística hemos considerado la media μ de una distribución normal, por ejemplo, y hemos estudiado la distribución de la media de una muestra obtenida a partir de esta distribución. En el análisis de regresión hallamos una situación similar, pero más general. Para cada valor x de la variable explicativa podemos pensar en una distribución de respuestas posibles y.

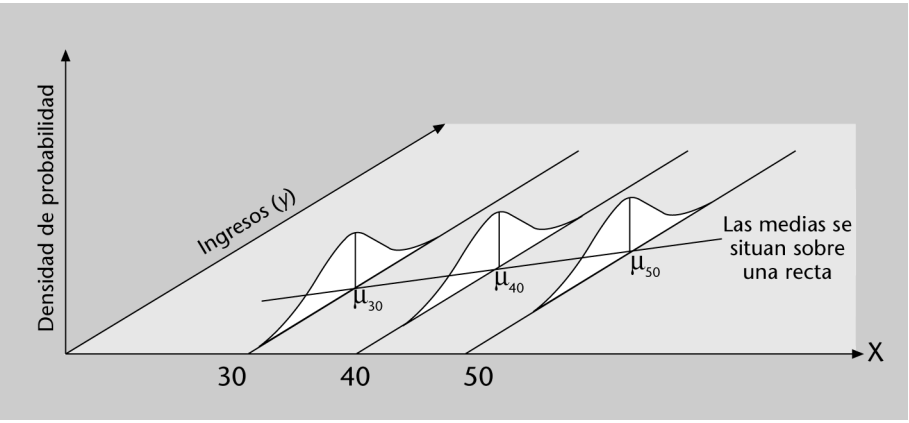
Por ejemplo, si y es ingresos y x es edad, podemos considerar las edades una por una, digamos que 30; entonces observamos la distribución de todos los ingresos de las personas de 30 años. Esta distribución tiene una media que podríamos representar por μ_{30}. Después podríamos observar los ingresos de todas las personas de 31 años, su media poblacional sería μ_{31}. Si lo hiciésemos con todas las edades de nuestra población, tendríamos un conjunto de distribuciones y un conjunto de medias.

Hemos tratado de ilustrar esta idea en la figura 21.1. Imaginemos que los ejes de las ordenadas y las abscisas definen una superficie plana (por ejemplo, el suelo de una habitación) y que el eje vertical muestra la densidad de probabilidad. Para cada valor de x hay una distribución de los valores de y. Hemos mostrado tres de estas distribuciones –para las edades de 30, 40 y 50–, pero, de hecho, hay un conjunto continuo de distribuciones como éstas que se mueven a través del recorrido de x.

La hipótesis principal en el análisis de regresión es que las medias de las distribuciones que corresponenden a cada valor de x se sitúan en una recta, tal como muestra la figura 21.1. Podemos expresar este hecho matemáticamente de la manera siguiente:

$$\mu_x = \alpha + \beta x$$

Figura 21.1.



La distribución en torno a la media se considera una variación aleatoria. Esta es la recta que nosotros tratamos de estimar al llevar a cabo una regresión. Se precisan dos suposiciones más, similares a unas que necesitábamos anteriormente, para realizar inferencias estadísticas. Es necesario que supongamos que todas las distribuciones normales que tenemos para unos valores dados de x tienen la misma desviación estándar. Teníamos que formular las mismas suposiciones cuando contrastábamos las diferencias entre dos grupos en el capítulo 18. Finalmente, tenemos que suponer que nuestras observaciones de y para cada x dada resultan independientes –se trata de una suposición que necesitamos siempre, y queda asegurada al ser nuestra muestra aleatoria.

Por tanto, para resumir, las cosas son como antes excepto por el hecho de que la media depende del valor de la variable explicativa x. Todos los cálculos que realizaremos en cuanto a la media serán respecto de esta media cambiante.

Estimación de la desviación estándar común

Si cada distribución normal para un valor dado de x tiene la misma desviación estándar σ, podemos estimar σ observando todas las desviaciones (residuos) de los puntos de la muestra a partir de la recta de regresión. Supongamos que tenemos estimaciones de las pendientes y el punto de corte de la recta a partir de nuestros datos (x_i, y_i), $i = 1, \ldots, n$, y que llamamos estas estimaciones $\hat{\alpha}$ y $\hat{\beta}$ (antes las llamábamos b y m). Así, la recta que se ajusta mejor a los datos es:

$$y = \hat{\alpha} + \hat{\beta} x$$

La diferencia entre cada observación y_i y el valor predicho de ésta $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ sobre la recta es:

$$y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i \qquad (21.2)$$

(éstos vuelven a ser los residuos). Si queremos calcular la desviación estándar de las y_i, este hecho requiere las desviaciones que presentan a partir de la media $\bar{y}$. Pero aquí queremos estimar la desviación estándar en torno a la recta (la media cambiante); por tanto, sumamos los cuadrados de (21.2). La fórmula que estima σ en este caso es:

$$s = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2} \qquad (21.3)$$

Observad otra pequeña diferencia: aquí dividimos la suma de las diferencias al cuadrado por $n - 2$, no por $n - 1$. Como antes, decimos que los grados de libertad son $n - 2$. Este hecho se debe a que la fórmula 21.3 para s contiene dos parámetros estimados a partir de los datos.

Como ilustración, veamos una vez más el precio de la acción bursátil del capítulo 20. Se estimaba que el modelo era $y = 2.492 + 49{,}52x$, por lo que obtenemos nuestros valores de predicción del precio de la acción para las semanas 1 a la 8 sustituyendo 1, 2, ..., 8 por x en esta ecuación para obtener las cifras de la segunda columna de más abajo:

2.550	2.541,5	8,5	71,9
2.590	2.591,0	−1,0	1,1
2.640	2.640,6	−0,6	0,3
2.670	2.690,1	−20,1	403,2
2.750	2.739,6	10,4	108,2
2.800	2.789,1	10,9	118,4
2.820	2.838,6	−18,6	347,5
2.900	2.888,2	11,8	140,2
			1.190,8

La tercera columna contiene los residuos y la cuarta columna, los residuos al cuadrado (observad que cualquier discrepancia entre la tercera columna y la cuarta se debe a los errores de redondear –p. ej. el primer residuo no es exactamente 8,5, sino un poco menos, y el cuadrado del valor exacto es 71,9, que también se corrige en un decimal). Sumando los residuos al cuadrado, dividiendo después por $n - 2 = 6$ y tomando la raíz cuadrada obtenemos la desviación estándar estimada como $\sqrt{1190{,}8/6} = 14{,}1$.

El error estándar de la pendiente

En el análisis de regresión raramente nos interesa el punto de corte con el eje vertical. Más bien nos interesa la pendiente, y si tenemos una muestra aleatoria que usamos para estimar la pendiente, nos interesa la distribución muestral de nuestra estimación. Si conocemos la distribución de la estimación y su error estándar, podemos construir intervalos de confianza y llevar a cabo contrastes de hipótesis como hacíamos antes. Ahora os mostramos la fórmula para el error estándar de la pendiente:

$$s_{\hat{\beta}} = \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad (21.4)$$

donde se obtiene s por medio de (21.3). Una vez más, efectuaremos una comparación con lo que hemos hecho anteriormente cuando observábamos el error estándar de una media muestral. Si s es la estimación de la desviación estándar, entonces el error estándar de la media es $s/\sqrt{n}$. El numerador de (21.4) también es una estimación de una desviación estándar, pero respecto de la recta de regresión, no con respecto a una media determinada. El denominador de (21.4) también aumenta a medida que el tamaño muestral aumenta, pero lo hace en mayor o menor medida según si el valor x se halla o no lejos de la media. Este hecho muestra cómo se puede reducir el error estándar de $\hat{\beta}$ si los valores x se dispersan mucho o, dándole la vuelta al ejemplo, si las observaciones que tienen el valor x cerca de la media reducen muy poco el error estándar.

El intervalo de confianza para la pendiente

Ahora que ya conocemos el error estándar de la pendiente, podemos construir intervalos de confianza y contrastar hipótesis. Como antes, usaremos la distribución t, en este caso con $n - 2$ grados de libertad. Usar la distribución t una vez más depende de la suposición de que la distribución de la variable y para una x dada resulta normal.

Nota

La razón por la que aquí aparecen $n - 2$ grados de libertad no es obvia. Cuando antes estimábamos la desviación estándar de la distribución normal, hemos "perdido" un grado de libertad, porque hemos tenido que estimar la media μ. En el caso que ahora nos ocupa, en que la media se describe con una recta con dos parámetros, estimando la recta perdemos dos grados de libertad.

Como ilustración, calculamos ahora el error estándar en el precio de la acción bursátil del ejemplo anterior. Ya hemos calculado s como 14,1. Necesitamos la suma de las desviaciones al cuadrado de los valores x respecto de su media:

$$(-3,5)^2 + (-2,5)^2 + \ldots + 2,5^2 + 3,5^2 = 42$$

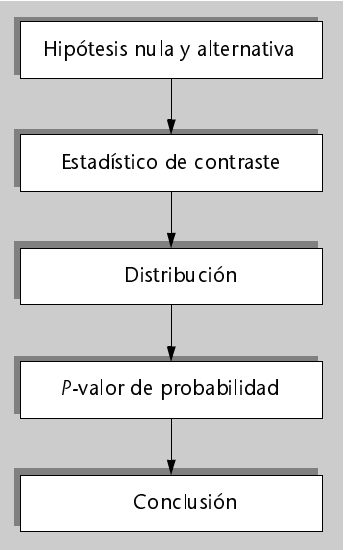
por tanto, el error estándar es $14,1/\sqrt{42} = 2,18$.

Para calcular un intervalo de confianza del 95%, tomamos nuestra estimación de la pendiente, que era 49,52, como punto medio, y calculamos el margen de error usando el error estándar y el valor crítico apropiado de la distribución t, con $n - 2 = 6$ grados de libertad. Utilizando las tablas observamos que el valor –2,4469 corta el 2,5% del extremo de la distribución t (con 6 grados de libertad), es decir, el 95% de la distribución se sitúa entre ±2,4469. Por tanto, el margen de error es 2,4469 veces el error estándar. Así obtenemos el intervalo de confianza de (44,186-54,854) para la pendiente.

El contraste de hipótesis sobre la pendiente

Si la pendiente es cero, y es una constante y no se da una relación lineal entre y y x. Así, la hipótesis nula natural que se contrastará es H_0: $\beta = 0$; en otras pa-

labras, la hipótesis es que en la población no existe ninguna relación entre la media y x, siendo la pendiente que se obtiene debida a la variación aleatoria. La hipótesis alternativa puede resultar unilateral o bilateral según el problema que se trata. Considerad una vez más el ejemplo del alza en el precio de la acción bursátil. A pesar de que resulta obvio que el alza será altamente significativa, realicemos la prueba convencional. Seguiremos los pasos desglosados en el capítulo 16.



1) Las hipótesis nula y alternativa son: H_0: $\beta = 0$; H_1: $\beta > 0$

2) El estadístico de contraste es la pendiente estimada $\hat{\beta}$ dividida por el error estándar de esta pendiente: 49,52/2,18 = 22,7.

3) La distribución usada es la t con 6 grados de libertad y sólo nos interesa la probabilidad del extremo de un lado. Se obtiene el valor de probabilidad, extremadamente pequeño, del orden de 2×10^{-7}.

4) La regresión resulta altamente significativa, y el precio de la acción bursátil muestra un incremento significativamente positivo.

La varianza explicada

Ahora daberíamos hablar de un resultado decisivo. Esta cifra es la que llamamos **varianza explicada por la recta de regresión**, representada por R^2. El valor de R^2 siempre se sitúa entre 0 y 1. Existen muchas maneras de definir este resultado, pero la manera más fácil es pensar que se trata de una medida de la proximidad de la recta al conjunto de puntos: lo llamamos **bondad de ajuste** de la recta a los puntos. En este ejemplo del precio de la acción bursátil, los puntos se sitúan muy cerca de la recta; por lo tanto, la recta consiste en un resumen de los puntos muy bueno –la recta da cuenta de la mayor parte de la varianza en el precio de la acción. Por lo tanto, el valor R^2 se sitúa cerca de 1. Si una recta de regresión consiste en un resumen pobre de los puntos, entonces R^2 será mucho más bajo. El valor $1 - R^2$ cuantifica cuánta varianza no se explica mediante la recta de regresión, efecto que denominamos **varianza inexplicada** (es decir, la varianza de los residuos). Cuando minimizamos el ajuste de la recta a los puntos, minimizamos $1 - R^2$ también, o equivalentemente maximizamos la bondad de ajuste R^2.

Actividad

21.1. Considerad los dos indicadores económicos siguientes para 12 países europeos en el año 1990, el producto interior bruto por cabeza de la población (PIB/C) y el consumo privado per cápita (CP/C):

	PIB/C	CP/C
Bélgica	102,0	104,9
Dinamarca	134,4	117,1

	PIB/C	CP/C
Alemania	128,1	126,0
Grecia	37,7	40,5
España	67,1	68,7
Francia	112,4	110,1
Irlanda	64,0	60,1
Italia	105,8	106,0
Luxemburgo	119,5	110,7
Holanda	99,6	96,7
Portugal	32,6	34,8
Reino Unido	95,3	9,7

a) Ejecutad la regresión de CP/C sobre PIB/C.

b) Interpretad los coeficientes de regresión y comprobad si existe una relación lineal significativa entre estas variables.

c) ¿Cuál es la bondad de ajuste? ¿Qué proporción de la varianza queda por explicar a partir de la recta de regresión?

Glosario

bondad del ajuste
Representada por R^2, proporción de la varianza de la variable respuesta que la recta de regresión "explica". Cuando R^2 se sitúa cerca de 1, la recta es un resumen muy bueno de la relación entre las variables y y x.

error estándar de la pendiente
Representado por $\hat{\beta}$, es

$$s_{\hat{\beta}} = \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

donde s se da a continuación).

estimación de la desviación estándar
Operación que se obtiene de la siguiente manera:

$$s = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}$$

donde $\hat{\alpha}$ y $\hat{\beta}$ son las estimaciones por mínimos cuadrados de los coeficientes de la recta (el punto de corte y la pendiente).

inferencia estadística y análisis de regresión
Para que la inferencia estadística resulte válida en el análisis de regresión necesitamos suponer que para cada valor x las observaciones de y son una muestra aleatoria de una distribución normal con la misma desviación estándar σ y una media aritmética μ_x que es una función lineal de x: $\mu_x = \alpha + \beta x$.

distribución de la pendiente
Bajo la hipótesis nula de que la verdadera pendiente es cero, la estimación de la pendiente $\hat{\beta}$ dividida por su error estándar s presenta una distribución t con $n - 2$ grados de libertad.