

Bondad de ajuste del modelo de regresión

Alfonso Novales

Departamento de Economía Cuantitativa

Universidad Complutense

Septiembre 2008

Coeficiente de correlación en la regresión lineal

- La raíz cuadrada del coeficiente de determinación resulta ser igual al coeficiente de correlación lineal de Pearson, que es otra razón por la cual éste está entre -1 y 1.

- Ninguno de los dos serían muy útiles si la relación entre X e Y ***no fuese lineal***
 - Dos variables pueden tener una relación estrecha, pero no lineal, y presentar un coeficiente de correlación lineal muy reducido
 - y un coeficiente de determinación muy reducido al estimar una regresión lineal

- Cuando dos variables son estadísticamente independientes, ninguna relación estimada entre ellas, lineal o no lineal presentará un buen ajuste
 - En particular, independencia \Rightarrow ausencia de correlación lineal
 - Pero el recíproco no es cierto

¿Es bueno el ajuste del modelo? (1)

- Error Estándar de la regresión (EER)
 - Estimación de σ_u^2 : Raíz cuadrada del cociente entre la Suma de Cuadrados de Residuos (SCR) y el número de grados de libertad de la estimación: $n-2$
 - Puede compararse con la desviación típica de Y, que refleja toda la fluctuación en Y. Así tendremos el porcentaje de fluctuación no explicado por el modelo.
 - Pero no permite comparar regresiones con distinta variable dependiente

- Coeficiente de determinación (R2) : $1 - SCR/ST$
 - Suma Total (ST) : Fluctuación de Y alrededor de su media muestral : n veces la varianza muestral de Y
 - Suma Explicada (SE) : Fluctuación de valores ajustados de Y alrededor de su media

¿Es bueno el ajuste del modelo? (2)

- La desviación entre un dato y su media = (desviación entre el valor ajustado y la media) + residuo:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + \hat{u}_i$$

⇒ Salvo que estimásemos una regresión sin término independiente:

ST = Suma Explicada + SCR , todos no negativos

⇒ $R^2 = SE/ST$,

⇒ R^2 entre 0 y 1

⇒ puede interpretarse como el porcentaje de la fluctuación en Y que queda explicada por el modelo, es decir, por X

Algunas consideraciones a tener en cuenta al interpretar el R2 (1)

- ❑ Un R2 puede ser bajo si el número de observaciones es reducido y hay algunos residuos de elevado tamaño
- ❑ Por otra parte, un número reducido de grados de libertad puede generar un R2 elevado, sin que la relación entre X e Y sea estrecha
- ❑ El R2 suele ser reducido en muestras de sección cruzada amplias (tamaños muestrales de 1000 o superiores)
- ❑ Y también en modelos estimados utilizando variaciones de series temporales
- ❑ Con datos temporales, si X e Y presentan tendencia, el R2 puede ser artificialmente elevado (regresión espuria)

Algunas consideraciones a tener en cuenta al interpretar el R2 (2)

- Con datos temporales, si X e Y presentan tendencia, el R2 puede ser artificialmente elevado (lo que se conoce como regresión espuria)
 - ✓ El R2 depende de la transformación de variables utilizada
 - ✓ No es el mismo al explicar Y que en la regresión de ln(Y)
 - ✓ O en la regresión que explica Y que en la regresión que explica las variaciones temporales de Y
 - ✓ O al estimar un modelo de consumo en datos temporales que en datos transversales

- Y del número de variables explicativas, aumentando siempre al añadir variables al modelo, por lo que suele utilizarse el R2 corregido, que se obtiene:

$$1 - \bar{R}^2 = (1 - R^2) \frac{N - 1}{N - k}$$

- ✓ que es siempre inferior al R2 estándar
- ✓ y no está garantizado su incremento al añadir variables explicativas al modelo

R2 y dependencia estadística

- ❑ Que el coeficiente de determinación sea bajo no indica que la variable dependiente y las variables explicativas sean estadísticamente independientes
- ❑ Un R2 reducido sólo indica la debilidad de la relación lineal entre Y y X
- ❑ El R2 no mide el grado de relación estadística entre X e Y, sino únicamente la calidad de la aproximación lineal a dicha relación. Un $R^2=1$ entre dos variables X e Y, implica que cada una de ellas es función lineal exacta de la otra.
- ❑ Si ambas variables tienen una relación estrecha, pero no lineal, pueden tener un R2 reducido
- ❑ Luego ausencia de correlación (R2 bajo) no indica independencia
- ❑ Si X e Y son independientes, no tendrán relación significativa de ningún tipo, ni lineal, ni cuadrática, exponencial, ...

El R2 y la constante del modelo

- Salvo escasas excepciones en que estimamos un modelo teórico según el cual Y y X guardan una relación de proporcionalidad, en general debemos incluir una constante o término independiente en el modelo
 - ✓ Si los datos no soportan su presencia, la estimación de su coeficiente será reducida y estadísticamente no significativa, sin contaminar la estimación del resto de los parámetros del modelo

- En los casos en que no se incluye constante, el R2 habitual no estaría entre 0 y 1, y podría ser negativo
 - ✓ Por lo que no debería compararse con el de otro modelo que incluya término independiente
 - ✓ Suele entonces utilizarse el R2 **no centrado**, definido como el cociente entre la suma de cuadrados de los valores ajustados y la suma de cuadrados de los valores observados de Y , en ambos casos sin descontar la media muestral
 - ✓ Que generalmente toma un valor superior al R2 habitual

Comparando estimadores mediante el R2

- No debe utilizarse para comparar la estimación de mínimos cuadrados (MC) con la que se obtenga por otros procedimientos (variables instrumentales, etc...) , pues el estimador MC está específicamente diseñado para generar el menor R2 posible
 - ✓ Y ello a pesar de que un estimador con un menor R2 puede tener mejores propiedades estadísticas

- De hecho, no es evidente cómo debe interpretarse el R2 cuando el estimador utilizado no es el de MC

- En esos casos, una posibilidad es utilizar el cuadrado del coeficiente de correlación entre valores observados y ajustados de Y
 - ✓ Que puede interpretarse como el grado de relación entre las variaciones en ambas variables
 - ✓ Aunque seguiría midiendo únicamente la calidad de la aproximación lineal a la verdadera relación entre ambas variables