

# La bondad de ajuste del modelo de regresión lineal simple

Alfonso Novales  
Departamento de Economía Cuantitativa  
Universidad Complutense

Septiembre 2008

## Contents

<b>1</b>	<b>Medidas de bondad de ajuste del modelo de regresión</b>	<b>1</b>
1.1	Error Estándar de la Regresión (EER) . . . . .	3
1.2	El coeficiente de determinación . . . . .	4
1.3	Correlación en el modelo de regresión lineal . . . . .	6

## 1 Medidas de bondad de ajuste del modelo de regresión

Hasta aquí, hemos propuesto un criterio, de entre los muchos posibles, para obtener estimadores de los coeficientes del modelo de regresión lineal simple: minimizar la suma de los cuadrados de los residuos, y hemos obtenido las expresiones analíticas de los estimadores resultantes, así como de sus varianzas y su covarianza. Cada uno de estos estimadores es una función de las observaciones muestrales de ambas variables,  $X$  e  $Y$ , y son, por tanto, variables aleatorias; por eso hemos calculado sus esperanzas matemáticas y varianzas. Si alguno de ellos fuese función únicamente de las observaciones de la variable  $X$  tendría naturaleza determinista, y su valor no cambiaría si en vez de utilizar en la estimación del modelo la muestra de que disponemos, pudiésemos utilizar otra muestra diferente de igual tamaño.

Sin embargo, éste no es el caso: ambos estimadores dependen también de las observaciones de la variable  $Y$ , por lo que tienen naturaleza estocástica, es decir, su valor numérico sería distinto con muestras diferentes. Variando la muestra, obtendríamos distintos valores de 0 y 1, todos los cuales nos describirían el histograma de frecuencias correspondiente a su distribución de probabilidad. En los párrafos anteriores hemos demostrado que la esperanza matemática de cada uno de estos estimadores es el verdadero valor, que es desconocido, del

parámetro que pretende estimar, y hemos deducido las expresiones analíticas de las varianzas de cada una de sus distribuciones de probabilidad.

El procedimiento MCO que hemos utilizado garantiza que la recta de regresión obtenida es la que proporciona la menor Suma de Cuadrados de Residuos que es posible obtener trazando rectas a través de la nube de puntos. Sin embargo, en unas ocasiones tal mejor ajuste puede ser excelente, en otras, el mejor ajuste puede no ser muy bueno. Necesitamos, en cualquier caso, disponer de criterios que puedan resumir en un indicador el grado de ajuste de la regresión MCO a la nube de puntos de que partimos.

Recordemos que:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Si la perturbación aleatoria sigue una distribución de probabilidad Normal, entonces  $y_i = \beta_0 + \beta_1 x_i + u_i$  también sigue una distribución Normal, pues es igual a una constante, más una variable con distribución Normal. Además:

$$\begin{aligned} E(y_i) &= E(\beta_0 + \beta_1 x_i + u_i) = \beta_0 + \beta_1 x_i \\ \text{Var}(y_i) &= \text{Var}(\beta_0 + \beta_1 x_i + u_i) = \text{Var}(u_i) = \sigma_u^2 \end{aligned}$$

de modo que, de acuerdo con el modelo, todas las observaciones de la variable endógena tienen la misma varianza, pero diferente esperanza matemática, pues ésta depende del valor numérico de la variable  $X$ , que varía a lo largo de la muestra.

Puede probarse que el residuo correspondiente a cada observación es una combinación lineal de todos los términos de error del modelo y, por tanto, si la perturbación aleatoria del modelo es Normal, el residuo también tiene distribución Normal. Su esperanza matemática es:

$$\begin{aligned} E(\hat{u}_i) &= E(y_i - \hat{y}_i) = E(y_i) - E(\hat{\beta}_0 + \hat{\beta}_1 x_i) = E(\beta_0 + \beta_1 x_i + u_i) - E(\hat{\beta}_0) - E(\hat{\beta}_1 x_i) = \\ &= \beta_0 + \beta_1 x_i + E(u_i) - \beta_0 - \beta_1 x_i = 0 \end{aligned}$$

Teniendo en cuenta que, entre  $\beta_0, \beta_1, \hat{\beta}_0, \hat{\beta}_1$  y  $x_i$ , sólo  $\hat{\beta}_0$  y  $\hat{\beta}_1$  son aleatorios, puede obtenerse la siguiente expresión para la varianza de cada residuo:

$$\begin{aligned} \text{Var}(\hat{u}_i) &= \text{Var}(y_i - \hat{y}_i) = \text{Var}[(\beta_0 + \beta_1 x_i + u_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = \\ &= \frac{\sigma_u^2 \sum_{j=1}^n (x_j - x_i)^2}{n \sum_{j=1}^n (x_j - \bar{x})^2} \end{aligned}$$

Al tener esperanza cero, la varianza del residuo es un adecuado indicador de su tamaño. Podemos ver que la varianza es tanto mayor (lo cual no es deseable), cuanto mayor es  $\sigma_u^2$ , pero es menor cuanto mayor sea el tamaño muestral. También es menor cuanto mayor es la varianza muestral de la variable

explicativa, lo cual es, por tanto, un aspecto deseable: un apreciable grado de fluctuación en  $X$  no es negativo, sino positivo. Por último, nótese que la observación  $x_i$  correspondiente al residuo  $i$  aparece en el numerador. Cuanto más se separe ésta de la media de todas las  $x_i$ , mayor será la varianza del residuo correspondiente a dicha observación muestral.

## 1.1 Error Estándar de la Regresión (EER)

No sólo es cierto que la esperanza matemática de la distribución de probabilidad de cada uno de los residuos MCO es igual a cero. También se cumple que su media muestral es igual a cero, puesto que la suma de todos ellos lo es, como vimos en las *ecuaciones normales*. Esta es una peculiaridad del método de estimación MCO, que otro procedimiento de estimación no tiene. Si, considerados a lo largo de toda la muestra, los residuos tienen media cero, entonces su desviación típica muestral será un indicador del tamaño promedio de cada uno de ellos. Esto es importante, porque si la recta estimada se ajusta bien a la nube de puntos, entonces los residuos deberían ser pequeños en algún sentido. Utilizar la desviación típica muestral de los residuos parece un criterio razonable de ajuste. Además, sabemos que si utilizamos  $n - 2$  en el denominador, su cuadrado es un estimador insesgado de  $\sigma_u^2$ . La ausencia de sesgo en este estimador puede demostrarse sin necesidad de obtener previamente los residuos de la regresión, tomando esperanzas en la expresión:

$$\begin{aligned}\hat{\sigma}_u^2 &= \frac{SCR}{n-2} = \sum_{i=1}^n \frac{\hat{u}_i^2}{n-2} = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \\ &= \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i y_i = \frac{1}{n-2} \left( \sum_{i=1}^n y_i^2 - \hat{\beta}_0 \sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i y_i \right)\end{aligned}$$

Su raíz cuadrada, la desviación típica estimada, recibe el nombre de error estándar de la regresión EER:

$$EER = \sqrt{\frac{\sum_{i=1}^n \hat{u}_i^2}{n-2}} = \sqrt{\hat{\sigma}_u^2} = \hat{\sigma}_u$$

Es claro que minimizar la varianza residual equivale a minimizar el error estándar de la regresión, EER. Sin embargo, recordemos que la desviación típica tiene, respecto a la varianza, la ventaja de estar medida en las mismas unidades que la variable a la que se refiere, el residuo, que tiene, a su vez, las mismas unidades que la variable endógena  $y_i$ . Para valorar si el ajuste obtenido por la recta MCO a la nube muestral de puntos es bueno, es conveniente utilizar el valor numérico del EER en relación con alguna referencia, y la media muestral de la variable endógena es un buen indicador. Ello nos permite presentar el porcentaje que de la media de  $y_i$  representa el EER, pudiendo decir, por ejemplo: el modelo estimado es bueno, puesto que el EER es tan sólo un 4% de la media de la variable endógena o, por el contrario: "el ajuste obtenido no es muy bueno,

porque el tamaño medio de los residuos, indicado por el EER, es de un 65% de la media de  $Y$ ".

## 1.2 El coeficiente de determinación

El interés del EER como indicador del grado de ajuste de un modelo de regresión disminuye cuando queremos comparar la bondad del ajuste de dos modelos que tienen una *variable dependiente diferente*. En tal caso, no es en absoluto cierto que el modelo con menor EER sea el modelo con mejor ajuste; de hecho, no podremos afirmar nada al respecto, salvo que establezcamos alguna medida relativa de grado de ajuste, que es lo que hacemos en esta sección. A diferencia del EER, el *coeficiente de determinación* que ahora definimos, denotado por  $R^2$ , es un indicador sin unidades, que no es preciso ni tiene sentido poner en relación con ninguna de las variables del modelo.

En primer lugar, escribamos para cada observación  $i$  :

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

que muestra que la distancia entre una observación  $y_i$  y su media  $\bar{y}$  puede escribirse como la distancia entre su valor ajustado  $\hat{y}_i$  y dicha media, más el residuo correspondiente. La distancia a la media del valor ajustado puede ser mayor o menor que la de la observación  $y_i$ , por lo que el residuo puede ser negativo o positivo. La regresión estimada por MCO proporciona el valor numérico de  $\hat{y}_i - \bar{y}$ , que es una aproximación a la distancia  $y_i - \bar{y}$ . El resto es la parte no explicada, o residuo. Como hemos mencionado, la explicación puede exceder o no de  $y_i - \bar{y}$ . La igualdad anterior muestra cómo *la desviación total respecto a la media puede escribirse como la suma de la desviación explicada y el residuo*.

Si elevamos al cuadrado ambos miembros, tenemos:

$$(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + \hat{u}_i^2 + 2(\hat{y}_i - \bar{y})\hat{u}_i$$

y sumando a lo largo de toda la muestra:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{u}_i^2 + 2 \sum_{i=1}^n (\hat{y}_i - \bar{y})\hat{u}_i \quad (1)$$

Pero:

$$\begin{aligned} \sum_{i=1}^n \hat{u}_i (\hat{y}_i - \bar{y}) &= \sum_{i=1}^n \hat{u}_i \hat{y}_i - \bar{y} \sum_{i=1}^n \hat{u}_i = \sum_{i=1}^n \hat{u}_i \hat{y}_i = \sum_{i=1}^n \hat{u}_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \\ &= \hat{\beta}_0 \sum_{i=1}^n \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^n \hat{u}_i x_i = \hat{\beta}_0(0) + \hat{\beta}_1(0) = 0 \end{aligned}$$

donde hemos utilizado repetidamente el hecho de que la suma de los residuos MCO es igual a cero, así como que la suma de sus productos por  $x_i$  también es igual a cero. Ambas condiciones provienen de las ecuaciones normales.

Finalmente, substituyendo en (1), llegamos a:

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

es decir:

$$nS_y^2 = nS_{\hat{y}}^2 + nS_{\hat{u}}^2$$

que expresa cómo la *variación muestral* total en la variable  $Y$ , que es  $n$  veces su varianza, puede descomponerse como la suma explicada por la regresión estimada,  $nS_{\hat{y}}^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ , el primero de los sumandos del miembro derecho, más la suma no explicada, que es la suma de los cuadrados de los residuos. Si dividimos la suma explicada por la variación total en  $Y$ , tenemos la definición de *coeficiente de determinación*:

$$R^2 = \frac{nS_{\hat{y}}^2}{nS_y^2} = \frac{S_y^2 - S_{\hat{u}}^2}{S_y^2} = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

de modo que:

$$R^2 = 1 - \frac{\text{Variación no explicada en } Y}{\text{Variación total en } Y} = \frac{\text{Variación explicada en } Y}{\text{Variación total en } Y}$$

**Proposition 1** *El coeficiente de determinación de todo modelo de regresión toma siempre valores numéricos entre 0 y 1.*

**Proof.** El miembro derecho de la ecuación es el cociente de dos términos positivos, luego es positivo. Además, hemos visto que el numerador es uno de los dos componentes del denominador, luego su valor numérico es inferior al de éste. En consecuencia, el cociente, que es positivo, es inferior a la unidad. ■

El coeficiente de determinación, a veces denominado *R - cuadrado*, nos indica el porcentaje de la variación total en la variable  $Y$  que la regresión estimada es capaz de explicar. La idea es que si la regresión tiene un ajuste suficientemente bueno, será debido a que la variable  $X$  explica buena parte de la variación que  $Y$  experimenta a lo largo de la muestra, los residuos serán generalmente pequeños, la variación explicada en  $Y$  será un porcentaje elevado de su variación muestral total, y el coeficiente de determinación será próximo a la unidad. Lo contrario ocurrirá cuando el ajuste de la recta MCO a la nube de puntos no sea suficientemente bueno, en cuyo caso el coeficiente de determinación será próximo a cero.

Así pues, un coeficiente de determinación próximo a 1 significa que las estimaciones obtenidas para los coeficientes del modelo de regresión hacen a éste capaz de explicar un elevado porcentaje de las variaciones que experimenta la variable endógena. El modelo proporciona en tal caso un buen ajuste a los datos, por lo que puede utilizarse con confianza para efectuar evaluaciones e

inferencias acerca de la cuestión conceptual que lo motivó inicialmente. En el extremo contrario, un coeficiente de determinación próximo a cero significa que las estimaciones obtenidas apenas explican las variaciones que experimenta la variable endógena, por lo que el modelo no puede utilizarse con una gran fiabilidad.

Hay que tener bastante cuidado, sin embargo, con la interpretación del coeficiente de determinación de una regresión. En ocasiones, si la muestra consta de pocas observaciones, quizá uno o dos residuos elevados pueden generar un coeficiente de determinación reducido y, por ello, conducir a creer que la regresión estimada es mala, cuando excepto por dichas observaciones, el ajuste puede ser excelente. Por otra parte, si la muestra consta de muy pocas observaciones, y ningún residuo es especialmente alto, se tendrá un coeficiente de determinación muy elevado, sin que deba interpretarse como un excelente ajuste, sino más bien como un indicador de escasa información muestral.

Otro caso delicado se refiere al uso del coeficiente de determinación con muestras de series temporales que muestran una tendencia similar. En tales casos, el coeficiente de determinación se aproxima a la unidad, aunque la relación entre ambas variables, excepción hecha de sus tendencias, pueda ser pobre. Esto viene indicado por dos ejercicios relacionados: a veces, basta estimar y extraer una tendencia determinista de dos series temporales  $X$  e  $Y$  para que un coeficiente de determinación en torno a 0,90 antes de la extracción de tendencias, se reduzca a 0,3 ó 0,4. El otro ejercicio, casi reverso del anterior, puede efectuarse tomando dos variables con poca relación, y añadiéndoles una tendencia, es decir, el producto de un determinado coeficiente, como  $\beta = 0,27$ , ó  $\beta = 3,45$ , por una variable de tendencia, que toma valores 1,2,3,... . Pues bien, si el coeficiente de determinación antes de añadir la tendencia estaba en torno a 0,20, por ejemplo, podría pasar a ser de 0,80 tras añadir la misma tendencia a ambas variables. Estos ejercicios son importantes, porque no querríamos decir en ninguno de los dos casos que las dos variables están muy relacionadas y que, en consecuencia, el modelo de regresión estimado es bueno, sólo porque el coeficiente de determinación sea elevado debido a la presencia de la tendencia común a ambas variables. Este aspecto, de suma importancia, es conocido como el problema de regresión espúria, y es estudiado en detalle más adelante.

Todo esto hace que, entre otras cosas, se exija un coeficiente de determinación superior en regresiones estimadas con datos de series temporales que con datos de sección cruzada. En todo caso, es imprescindible acompañar toda estimación de un modelo de regresión, con los estadísticos que permitan evaluar la bondad del ajuste entre modelo y datos. Estos incluirán el coeficiente de determinación  $R^2$ , el EER, así como estadísticos que examinaremos en las próximas secciones.

### 1.3 Correlación en el modelo de regresión lineal

Correlación es el grado de dependencia que existe entre variables. Cuando se trata de sólo dos variables, existe una medida, el coeficiente de correlación, introducido por K.Pearson:

$$\rho_{xy} = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)}\sqrt{\text{Var}(y)}}$$

cuya justificación estamos ahora en condiciones de comprender. Vamos a demostrar que el coeficiente de correlación de Pearson mide el grado de *dependencia lineal* que existe entre dos variables,  $X$  e  $Y$ .

Para ello, partimos del coeficiente de determinación de una regresión lineal simple, y extraemos su raíz cuadrada, denotando por  $r_{xy}$  al estadístico que así se obtiene:

$$r_{xy} = \sqrt{R^2} = \sqrt{1 - \frac{S_{\hat{u}}^2}{S_y^2}}$$

Ahora bien, puesto que:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

tenemos:

$$\begin{aligned} S_{\hat{u}}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n \left[ y_i - \left( \bar{y} + \frac{S_{xy}}{S_x^2} (x_i - \bar{x}) \right) \right]^2 = \\ &= \frac{1}{n} \sum_{i=1}^n \left[ (y_i - \bar{y})^2 + \frac{(S_{xy})^2}{(S_x^2)^2} (x_i - \bar{x})^2 - 2 \frac{S_{xy}}{S_x^2} (x_i - \bar{x}) (y_i - \bar{y}) \right] = \\ &= S_y^2 + \frac{(S_{xy})^2}{(S_x^2)^2} S_x^2 - 2 \frac{(S_{xy})^2}{S_x^2} = S_y^2 - \frac{(S_{xy})^2}{S_x^2} \end{aligned}$$

y, en consecuencia:

$$r_{xy} = \sqrt{R^2} = \sqrt{1 - \frac{S_{\hat{u}}^2}{S_y^2}} = \sqrt{1 - \frac{S_y^2 - \frac{(S_{xy})^2}{S_x^2}}{S_y^2}} = \sqrt{\frac{(S_{xy})^2}{S_x^2 S_y^2}} = \frac{S_{xy}}{S_x S_y} = \rho_{xy}$$

obteniendo, precisamente, *el coeficiente de correlación lineal*. Es decir, por haber demostrado que el coeficiente de correlación de Pearson no es sino la raíz cuadrada del coeficiente de determinación en un modelo de regresión lineal, podemos afirmar que el coeficiente de correlación de Pearson mide el grado de relación entre dos variables,  $X$  e  $Y$ , supuesto que la relación entre ambas sea de tipo lineal. Por tanto, su interpretación sólo está realmente justificada en la medida que la regresión óptima entre ambas variables, es decir, la esperanza condicional de  $Y$  dado  $X$ , sea lineal, y no en otro caso.

Asimismo, puesto que ya hemos probado que el coeficiente de determinación está comprendido entre 0 y 1, podemos obtener ahora como corolario que *el coeficiente de correlación de Pearson está siempre comprendido entre -1 y +1*, resultado bien conocido de cursos de Estadística.

Es importante destacar que si la verdadera relación entre dos variables no es lineal, y utilizamos el coeficiente de correlación de Pearson como un indicador del grado en que ambas están relacionadas, podemos cometer todo tipo de errores. En tal situación, habría que tratar de identificar qué forma funcional adopta el mejor modelo de relación entre ambas variables con el objeto de proceder a su estimación y posterior evaluación de los correspondientes residuos. No es difícil encontrar ejemplos de *relación no lineal exacta* entre dos variables a pesar de que ambas presentan un coeficiente de correlación igual a cero.

Como sabemos, si dos variables son independientes, entonces su covarianza es igual a cero. Pero el coeficiente de Pearson es el cociente entre ésta y el producto de las desviaciones típicas de  $X$  e  $Y$ , por lo que, si dos variables son independientes, entonces su coeficiente de correlación lineal es igual a cero. Ello no puede sorprendernos en modo alguno: estamos afirmando que si dos variables  $X$  e  $Y$  son independientes, y ajustamos una recta de regresión, es decir, un modelo lineal, a un conjunto de observaciones muestrales de ambas variables, entonces detectaremos un grado de asociación nulo entre ambas.

También podríamos ajustar modelos de otro tipo, con funciones no lineales; aunque no los hemos examinado aquí, existen procedimientos de estimación de tales modelos. Hecho tal ejercicio, volveríamos a detectar una capacidad nula del modelo no lineal, para relacionar  $X$  e  $Y$ , si bien es cierto que deberíamos utilizar algún estadístico adecuado, que relacionase la suma de cuadrados de los residuos con la suma de cuadrados de la variable  $Y$ . En resumen, si dos variables son independientes, no podemos estimar ninguna forma funcional de relación entre ellas que genere capacidad explicativa alguna; en particular, una recta no explicará ninguna asociación.

Por el contrario, si el coeficiente de correlación de Pearson es nulo, sólo podremos afirmar que la relación lineal entre ambas variables no es muy buena, pues no se detecta un grado apreciable de asociación entre ambas, supuesto que la forma funcional de tal hipotética relación sea lineal. Sin embargo, ello no excluye la posibilidad de que otra forma funcional, no lineal, reflejase un grado de asociación notable entre ambas variables que, en tal caso, serían dependientes. Por tanto, ausencia de correlación lineal entre dos variables, o incorrelación, que es lo que mide el coeficiente de correlación de Pearson, no implica en modo alguno su independencia.

Ahora que conocemos la estrecha relación entre coeficiente de correlación de Pearson y coeficiente de determinación, podemos apreciar que el primero nos proporciona una información acerca de la relación entre las variables que el coeficiente de determinación no consigue transmitirnos. Ello se debe a que el coeficiente de determinación es el cuadrado del coeficiente de correlación, por lo que pierde la información concerniente a su signo; ésta es relevante, excepto en algunas situaciones en que es perfectamente conocido a priori, dada la naturaleza de las variables  $X$  e  $Y$ . Por ejemplo, si estimamos una regresión de la cantidad vendida de un producto en un mercado con cierto poder de monopolio, sobre su precio, sabemos a priori que ésta será una relación de signo negativo: un coeficiente  $\beta_1$  negativo implicará que variaciones positivas, es decir, aumentos en el precio del producto, se transmiten en variaciones negativas, es

decir, descensos, en la cantidad vendida, y viceversa. En este ejemplo, nos interesará tan sólo tratar de estimar el grado en que el precio explica la cantidad vendida: si lo hace en gran medida o si, por el contrario, la capacidad explicativa no es muy elevada y debemos encontrar otros factores explicativos (quizá precios de otros productos con cierto grado de sustitución del nuestro, la renta de las familias, etc.) que añadir al modelo de regresión.

Cuando no contamos con esta información, queremos estimar no sólo la capacidad que  $X$  tiene para explicar las variaciones que experimenta  $Y$ , sino también el signo de su relación. Para ello, observemos que el signo del coeficiente de correlación es el mismo que el de la covarianza, de modo que si ésta es positiva, la relación entre ambas variables es positiva o creciente, siendo negativa o decreciente en el caso alternativo. Por otra parte, los valores numéricos absolutos del coeficiente de correlación de Pearson evolucionan muy en relación con los que toma el coeficiente de determinación: si uno es cero, lo es el otro, mientras que si el valor absoluto del coeficiente de correlación es uno, también es igual a uno el coeficiente de determinación. Además, puesto que el coeficiente de determinación sólo toma valores numéricos entre 0 y 1, necesariamente el coeficiente de correlación toma valores numéricos entre -1 y +1.

Así, decimos que cuando el coeficiente de correlación lineal es próximo a +1, la relación entre ambas variables es estrecha y directa, o de signo positivo, es decir, cuando una aumenta, también lo hace la otra, y también tienden a disminuir simultáneamente. Cuando una de las variables está por encima de su media, la otra variable tiende a estar también por encima de su media, y cuando una está por debajo, también tiende a estarlo la otra. Si fuese exactamente igual a +1, lo que es prácticamente imposible cuando se trabaja con datos reales, diríamos que la relación entre ambas variables es *perfecta, y positiva o directa*. Cuando el coeficiente de correlación es próximo a -1, entonces la relación es muy estrecha, pero inversa, o de signo negativo, es decir, cuando una variable aumenta la otra tiende a disminuir, y viceversa. Cuando una variable está por encima de su media, la otra variable tiende a estar por debajo de su media. Si fuese exactamente igual a -1, diríamos que la relación entre las variables es *perfecta y negativa, o inversa*. Cuando el coeficiente de correlación es próximo a cero, también lo es el coeficiente de determinación, por lo que decimos que la relación lineal entre las variables  $X$  e  $Y$  es prácticamente inexistente.

No debe olvidarse, sin embargo que, a diferencia del coeficiente de determinación, el coeficiente de correlación no es estrictamente cuantitativo: si tenemos dos modelos de regresión para una misma variable dependiente, con coeficientes de correlación de .35 y .70, no podemos decir que el segundo tiene un ajuste doblemente mejor que el primero, si bien podemos afirmar que muestra un ajuste claramente mejor. Tales afirmaciones acerca de comparaciones estrictamente cuantitativas sólo pueden hacerse para el coeficiente de determinación, por su significado como porcentaje de la variación en la variable dependiente que el modelo es capaz de explicar. Si los anteriores valores numéricos correspondiesen a los coeficientes de determinación de ambos modelos, entonces sí que podríamos afirmar que el segundo muestra un ajuste doblemente superior al primero.

En definitiva, los análisis de correlación y de regresión proporcionan respuestas similares acerca de la evolución conjunta de dos variables (o más de 2 variables, en el caso de la regresión múltiple). El análisis de correlación, basado estrictamente en el cálculo del coeficiente de correlación de Pearson, facilita el grado y signo de la asociación, pero no proporciona una idea acerca de la forma funcional de dicha relación, ni tampoco su dirección. Esta, que sí se obtiene con el análisis de regresión, es una ventaja del mismo, pero está condicionada a que se satisfagan las hipótesis del modelo de regresión lineal, que condicionan la validez del método MCO para la estimación del modelo lineal de regresión: así, si a) la verdadera función de relación entre variables, que el analista desconoce, es realmente lineal, b) no se omiten variables explicativas relevantes, c) el término de error del modelo no tiene media significativa, d) ni sus valores para distintas observaciones están correlacionados entre sí, e) si su varianza es la misma para todas las observaciones, y f) si no existe una relación causal de  $Y$  hacia  $X$ , entonces el análisis de regresión mediante la estimación MCO está plenamente justificada y será conveniente utilizarlo, por cuanto que nos proporciona más información que el mero análisis de correlación.

Además, el uso del estimador MCO en el modelo de regresión lineal simple está justificado por sus propiedades de eficiencia: es el estimador lineal de mínima varianza y si, además de las condiciones anteriores, las perturbaciones tienen distribución Normal, entonces es eficiente, pues su varianza alcanza la cota de Cramer-Rao.

Por el contrario, si tenemos razones para creer que una o más de tales hipótesis dejan de cumplirse en un grado apreciable, podemos perder confianza en los resultados que el análisis de regresión pueda facilitarnos, prefiriendo efectuar un análisis de correlación, cuya validez no descansa sobre tantas hipótesis, si bien precisa del supuesto acerca de que la verdadera función de relación entre  $X$  e  $Y$  sea lineal.