



UNIVERSIDADE
DE VIGO

Departamento de Estatística
e Investigación Operativa

Minería de Datos, 2007-2008

Notas sobre R^2 en la Regresión Lineal sin término independiente

1. El problema de regresión lineal consiste en resolver:

$$\min_{\beta} \{L(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\}$$

cuya solución es:

$$\mathbf{0} = \nabla_{\beta} L = 2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 2\mathbf{X}^T\mathbf{y} - 2\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = \mathbf{X}^T\mathbf{e} \quad (1.1)$$

$$\Leftrightarrow \mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\beta \quad (1.2)$$

$$\Leftrightarrow \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

2. Si hay término constante entonces $\mathbf{X} = [\mathbf{1} \ \mathbf{X}']$. Como la ecuación (1.2) es $\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\hat{\mathbf{y}}$ entonces, en ese caso, la primera ecuación del sistema (1.2) es:

$$\begin{aligned} \mathbf{1}^T\mathbf{y} &= \mathbf{1}^T\hat{\mathbf{y}} \\ \bar{y} &= \bar{\hat{y}} \end{aligned}$$

pero, si no hay término constante la primera columna de $\mathbf{X}$ ya no es el vector $\mathbf{1}$, con lo que las medias no tienen por qué ser iguales.

3. La suma de cuadrados de $\mathbf{Y}$ es $nS_Y^2 = \mathbf{y}^T\mathbf{y} - n\bar{y}^2$ y, en caso de existir término constante, la de $\hat{\mathbf{Y}}$ es:

$$\begin{aligned} nS_{\hat{Y}}^2 &= \hat{\mathbf{y}}^T\hat{\mathbf{y}} - n\bar{\hat{y}}^2 \\ &= [\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}]^T [\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] - n\bar{y}^2 \\ &= \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - n\bar{y}^2 \\ &= \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - n\bar{y}^2 \\ &= \beta^T\mathbf{X}^T\mathbf{y} - n\bar{y}^2 \end{aligned}$$

Ambos términos constituyen la fórmula de R^2 :

$$R^2 = \frac{nS_{\hat{Y}}^2}{nS_Y^2} = \frac{\beta^T\mathbf{X}^T\mathbf{y} - n\bar{y}^2}{\mathbf{y}^T\mathbf{y} - n\bar{y}^2} \quad (1.3)$$

4. Se verifica que:

$$\begin{aligned} nS_Y^2 &= \mathbf{y}^T\mathbf{y} - n\bar{y}^2 = \mathbf{y}^T\mathbf{y} - \frac{n}{n^2}(\mathbf{y}^T\mathbf{1})(\mathbf{1}^T\mathbf{y}) = \mathbf{y}^T(\mathbf{y} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{y}) \\ &= \mathbf{y}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{y} \\ &= (\mathbf{y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta})^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(\mathbf{y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta}) \\ &= (\mathbf{e} + \hat{\mathbf{y}})^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(\mathbf{e} + \hat{\mathbf{y}}) \\ &= \mathbf{e}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{e} + \hat{\mathbf{y}}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\hat{\mathbf{y}} + 2\mathbf{e}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\hat{\mathbf{y}} \\ &= nS_e^2 + nS_{\hat{Y}}^2 + 2(\mathbf{e}^T\hat{\mathbf{y}} - \frac{1}{n}\mathbf{e}^T\mathbf{1}\mathbf{1}^T\hat{\mathbf{y}}) \\ &= nS_e^2 + nS_{\hat{Y}}^2 \end{aligned} \quad (1.4)$$

pues las estimaciones son ortogonales a los errores:

$$\begin{aligned} \mathbf{e}^T \hat{\mathbf{y}} - \frac{1}{n} \mathbf{e}^T \mathbf{1} \mathbf{1}^T \hat{\mathbf{y}} &= \mathbf{e}^T \hat{\mathbf{y}} - \frac{1}{n} n \bar{e} \cdot n \bar{y} = \mathbf{e}^T \hat{\mathbf{y}} - n \bar{e} \bar{y} = \mathbf{e}^T \hat{\mathbf{y}} \\ &= (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{0} \end{aligned}$$

ya que por la condición (1.1) de óptimo: $\mathbf{0} = \mathbf{X}^T(\mathbf{y} - \mathbf{X} \boldsymbol{\beta}) = \mathbf{X}^T \mathbf{e} \Rightarrow \mathbf{0} = \boldsymbol{\beta}^T \mathbf{X}^T(\mathbf{y} - \mathbf{X} \boldsymbol{\beta}) = \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{e} = \hat{\mathbf{y}}^T \mathbf{e}$.

5. Por tanto, una fórmula equivalente a (1.3) es:

$$R^2 = \frac{n S_Y^2}{n S_Y^2} = \frac{n S_Y^2 - n S_e^2}{n S_Y^2} = 1 - \frac{n S_e^2}{n S_Y^2} \quad (1.5)$$

6. Sin embargo, cuando se impone término independiente nulo:

- (a) No se verifica $\bar{\hat{y}} \neq \bar{y}$ con lo que el numerador de R^2 en (1.3) ya no es la varianza de las estimaciones sino un valor arbitrario.
- (b) La fórmula (1.5) no es equivalente a la anterior porque en este caso no tiene por qué verificarse que $\bar{e} = 0$ por cuya razón no tiene por qué cumplirse la descomposición (1.4) a pesar de que $\hat{\mathbf{y}} \perp \mathbf{e}$.

Nótese que tampoco es válida la fórmula:

$$R^2 = 1 - \frac{\sum_i e_i^2}{n S_Y^2}$$

pues el numerador de la segunda expresión no es S_e^2 al ser $\bar{e} \neq 0$.

- 7. Al final, lo que hace Excel en la función de regresión del complemento *Análisis de Datos* es no utilizar las medias en las sumas de cuadrados y realizar el ANOVA con unas sumas de cuadrados no centradas¹, pero ello tiende a inflar el valor de R^2 cuando la escala de la variable es grande. Sin embargo, con la línea de tendencia lineal que puede obtenerse en el gráfico de dispersión se calcula el R^2 con las fórmulas tradicionales, lo que es erróneo.

¹Probablemente, Excel utilice en estos casos un test ANOVA basado en la distribución F no centrada. Dicha distribución se obtiene como cociente de χ^2 no centradas, distribuciones que siguen las sumas de cuadrados no centrados.