Koeficient determinace

základní definice jak pro centrované, tak pro necentrované veličiny

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \text{ kde}$$

SSR je regresní součet čtverců $SSR = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}}\right)^2$

SST je celkový součet čtverců $SST = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$

SSE je reziduální součet čtverců $SSE = \sum_{i=1}^{n} \left(e_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}\right)^2$

přičemž $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$; $\bar{\hat{y}} = \frac{\left(\sum_{i=1}^{n} \hat{y}_i\right)}{n}$ a dále platí $\bar{y} = \bar{\hat{y}}$

Poznámka: u SSE není třeba odečítat nulový průměr, neboť víme, že $\sum_{i=1}^{n} e_i = 0$

Dále víme, že platí $y = \beta + \varepsilon$, resp. $\hat{y} = Xb = X\left(X'X\right)^{-1} X'y$ tj. $\hat{y} = M \cdot y$, kde $M = X\left(X'X\right)^{-1} X'$

Dále uvažujme centrované proměnné y^* a $x_1^*,....,x_k^*$, které jsou vyjádřeny jako odchylky od svých průměrů $y_i^* = y_i - \bar{y}$; $x_{ij}^* = x_{ij} - \bar{x}_j \;\; j = 1,...,k$:

Vyjádříme-li SSE výrazem $\left(y^* - X^* b\right)' \left(y^* - X^* b\right) = e'e$, pak platí

$$R^2 = 1 - \frac{SSE}{SST} = \frac{y^{*\prime} y^* - \left(y^* - X^* b\right)' \left(y^* - X^* b\right)}{y^{*\prime} y^{\prime}} =$$

$$= \frac{y^{*\prime} y^* - y^{*\prime} y^* + b'X^{*\prime} y^*}{y^{*\prime} y^*} = \frac{b'X^{*\prime} y^*}{y^{*\prime} y^*} = \frac{\left(X^* b\right)' y^*}{y^{*\prime} y^*}$$

protože (a) $e'e = \left(y - Xb\right)' \left(y - Xb\right) = y'y - b'X'y - y'Xb + b'X'Xb$

a dále (b)

$$e'X = \left(y - Xb\right)' X = y'X - b'X'X = y'X - \left[\left(X'X\right)^{-1} X'y\right]' X'X = y'X - y'X\left(X'X\right)^{-1} X'X = 0$$

Poznámka : vztahy (a) a (b) platí jak pro centrované, tak pro necentrované veličiny ..

Ověření vztahu SST=SSE+SSR obecně pro necentrované veličiny:

SST = $\sum_1^T (y_i - \bar{y})^2 = \sum_1^T y_i^2 - 2\sum_1^T y_i \cdot \bar{y} + T \cdot \bar{y}^2 = \sum_1^T y_i^2 - T\bar{y}^2$

SSE = $\sum_1^T e_i{}^2 = \sum_1^T (y_i - \hat{y}_i)^2 = \sum_1^T y_i^2 - 2\sum_1^T y_i \hat{y}_i + \sum_1^T \hat{y}_i^2$

SSR = $\sum_1^T (\hat{y}_i - \bar{\hat{y}})^2 = \sum_1^T \hat{y}_i^2 - 2\sum_1^T \hat{y}_i \bar{\hat{y}} + T \cdot \bar{\hat{y}}^2$

SSE + SSR =

$= \sum_1^T (y_i - \hat{y}_i)^2 + \sum_1^T \left(\hat{y}_i - \bar{\hat{y}}^2\right) = \sum_1^T y_i^2 - 2\sum_1^T y_i \hat{y}_i + \sum_1^T \hat{y}_i^2 + \sum_1^T \hat{y}_i^2 - 2\sum_1^T \hat{y}_i \bar{\hat{y}} + T \cdot \bar{\hat{y}}^2 =$

$= \sum_1^T y_i^2 - T \cdot \bar{\hat{y}}^2$ (= SST), neboť

$-2\sum_1^T y_i \hat{y}_i + \sum_1^T \hat{y}_i^2 + \sum_1^T \hat{y}_i^2 = -2\sum_1^T y_i \hat{y}_i + 2\sum_1^T \hat{y}_i^2 = 0$, protože

$\sum_1^T y_i \hat{y}_i = \sum_1^T (\hat{y}_i + e_i) \cdot \hat{y}_i = \sum_1^T \hat{y}_i^2 + \sum_1^T e_i \hat{y}_i = \sum_1^T \hat{y}_i^2$ a dále

$$\bar{y} = \frac{\sum y_i}{T} = \frac{\sum (\hat{y}_i + e_i)}{T} = \frac{\sum \hat{y}_i}{T} + \frac{\sum e_i}{T} = \bar{\hat{y}}$$

Korigovaný (rektifikovaný) koeficient determinace:

$$\widetilde{R}^2 = R^2 - \frac{k-1}{T-k}.(1-R^2) = 1-(1-R^2).\frac{T-1}{T-k}$$

se užívá pro porovnání výstižnosti dvou (nebo více) **specifikací regresních rovnic se stejnou vysvětlovanou proměnnou a různým počtem vysvětlujících proměnných (a nejvhodněji tehdy, jsou-li aspoň některé z vysvětlujících proměnných shodné).**

Motiv pro vyvození $\widetilde{R}^2$:

Vyjděme z vyjádření

$$(1-R^2).(T-1) = (1-\widetilde{R}^2).(T-k) \ ,$$

které představuje jistý „kompenzační“ efekt mezi „nevystiženými variabilitami“ závisle proměnné při různých specifikacích vysvětlujících proměnných (při přechodu od 1 vysvětlující proměnné ke k proměnným). Dělením $T-k$ máme:

$$(1-R^2).\frac{T-1}{T-k} = 1-\widetilde{R}^2 \quad \text{neboli}$$

$$\widetilde{R}^2 = 1-(1-R^2).\frac{T-1}{T-k} \quad \text{, což je „definiční tvar 2 “.}$$

Tvrzení: **Platí $\widetilde{R}^2 \le R^2$ s rovností toliko pro $R^2 = 1$ nebo pro $k = 1$.**

Ověření:

Okamžitě je vidět, že od R^2 odečítáme součin dvou nezáporných členů, z nichž první je nulový jen tehdy, když $k = 1$, a druhý tehdy, když $R^2 = 1$.

Vztah obou koeficientů:

Pro velká T se oba tyto koeficienty determinace liší jen velmi málo, Pokud je ale počet stupňů volnosti $T-k$ malý, bude $\widetilde{R}^2$ o hodně menší než R^2 a může nabýt i záporné hodnoty. (pak ho pokládáme za nulový, shodně s případem, kdy je kladný, ale hodnotou velmi malý) .

Koeficient determinace R^2 nemůže hodnotou nikdy klesnout s přidáním další vysvětlující proměnné do regresní rovnice. Je tomu tak zřejmě proto, že přidání jakékoli (i nevhodné) **vysvětlující proměnné nemůže snížit stupeň korelace mezi závisle proměnnou a nejlépe ji vystihující lineární kombinaci z vysvětlujících proměnných.**

Korigovaný koeficient determinace R^2 může s přidáním další vysvětlující proměnné do regresní rovnice klesnout. Bude tomu tak tehdy, když přidané vysvětlující proměnné způsobí menší pokles hodnoty $1-R^2$, než kolik je třeba ke kompenzaci rostoucího podílu $(T-1)/(T-k)$.