# More on goodness-of-fit and selection of regressors.

- General remarks on R-squared
  - A high R-squared does not imply that there is a causal interpretation
  - A low R-squared does not preclude precise estimation of partial effects

- Adjusted R-squared
  - What is the ordinary R-squared supposed to measure?

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{(SSR/n)}{(SST/n)} \quad \text{is an estimate for} \quad 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

Population R-squared

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

$R^2 \uparrow$ as $k \uparrow$

number of estimator $X_i$.

$$R^2 = 1 - \frac{E\left(\frac{SSR}{n-1}\right) \neq \sigma_u^2}{E\left(\frac{SST}{n-1}\right) = \sigma_y^2}$$

<u>adjusted R-squared</u>: $\bar{R}^2 = 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}}$

$$= 1 - \frac{SSR}{SST} \cdot \frac{n-1}{n-k-1}$$

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

增加自变量不一定会令 $\bar{R}^2$ 增加

if and only if $|t_{new}| > 1$, $\bar{R}^2$ 增加

Set $t_{x_j} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$ ($\hat{\beta}_j \uparrow \to$ significant, $se(\hat{\beta}_j) \uparrow \to$ bad.), if $|t_{x_j}| > 1$, $\bar{R}^2 \uparrow$ as $x_j$ is added to the model.

if $|t_{x_j}| < 1$, $\bar{R}^2 \downarrow$ as $x_j$ is added.

Nested models: $ur$: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_5 x_5 + u$.

$H_0$: $\beta_4 = \beta_5 = 0$

$r$: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$.

$$F = \frac{(R^2_{ur} - R^2_r)/2}{(1 - R^2_{ur})/(n-5-1)}.$$

$\underbrace{\text{Non-nested}}$ Model $y = \beta_0 + \beta_1 \log x_1 + u$.

if neither model is a special case of the other $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u$.

$$rdintens = \beta_0 + \beta_1 \log(sales) + u \leftarrow R^2 = .061, \bar{R}^2 = .030$$

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u \leftarrow R^2 = .148, \bar{R}^2 = .090$$

use $\bar{R}^2$

- A comparison between the R-squared of both models would be unfair to the first model because the first model contains fewer parameters
- In the given example, even after adjusting for the difference in degrees of freedom, the quadratic model is preferred

Can't use $R^2$ or $\bar{R}^2$ to compare models with different definition of the dependent model. $\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\ \log y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u. \end{cases}$