R^2 increase

Proposition 3.1.4

Anastasiia Gordienko, Patrick Pöpperling

Research Module - Econometrics and Statistics

November 4, 2019

Content

1. Reminder: What is R^2 and why could it be problematic?
2. Proofs
   2.1 R^2 increase
   2.2 Adjusted R^2
3. Conclusion

What is R^2?

- statistical measure that *represents the proportion of variance* for a dependent variable *that can be explained* by an independent variable
- used to express how well a model fits the observed data

What is R^2?

$$R^2 = \frac{\overbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}^{\text{Explained Variation}}}{\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Total Variation}}} = 1 - \frac{\overbrace{\sum_{i=1}^{n}\hat{\epsilon}_i^2}^{\text{Unexplained Variation}}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- $R^2 \in [0, 1]$

What is R^2?

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

- If Explained variation = Total Variation $\Rightarrow R^2 = 1$
  - $R^2 = 1$ indicates that the model explains *all the variablility*
  - $R^2 = 0$ indicates that the model explains *none of the variability*

$\Rightarrow$ The higher R^2, the better the model fits the observed data.

Why could R^2 become problematic?

- R^2 only describes how well the model fits the observations, it does neither validate nor reject it
- R^2 increase: Additional regressors always increase R^2, independent of their relevance

$$\text{assume we get } R_1^2 \text{ from } y = X_1 b_{11} + \hat{\epsilon}_1$$
$$\text{and we get } R_2^2 \text{ from } y = X_1 b_{11} + X_2 b_{22} + \hat{\epsilon}_2$$

$$\text{then } R_2^2 \geq R_1^2 \text{ always holds}$$

Proof: R^2 increase

- Consider the sum of squared residuals

$$S(b_{21}^*, b_{22}^*) = (y - X_1 b_{21}^* + X_2 b_{22}^*)'(y - X_1 b_{21}^* + X_2 b_{22}^*)$$

- This sum is minimized by OLS estimators b_{21} and b_{22}:

$$\hat{\epsilon}_2{}'\hat{\epsilon}_2 = S(b_{21}, b_{22}) \leq S(b_{11}, 0) = \hat{\epsilon}_1{}'\hat{\epsilon}_1$$

- This implies that

$$1 - \frac{\sum_{i=1}^n \hat{\epsilon}_2{}'\hat{\epsilon}_2}{\sum_{i=1}^n (y_i - \bar{y})^2} \geq 1 - \frac{\sum_{i=1}^n \hat{\epsilon}_1{}'\hat{\epsilon}_1}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$\Leftrightarrow R_2^2 \geq R_1^2 \quad \square$$

R^2 adjusted

- Introduce an adjusted $\bar{R}^2$ to deal with this problem.

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-K}\sum_{i=1}^{n}\hat{\epsilon}_i^{\,2}}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{n-1}{n-K}(1 - R^2)$$

- $\bar{R}^2$ is better than R^2 if $\bar{R}^2 \leq R^2$

Proof: $\bar{R}^2 \leq R^2$

$$
\begin{aligned}
\bar{R}^2 &= 1 - \frac{n-1}{n-K}(1-R^2) \\
&= 1 - \frac{n-1}{n-K} + \frac{n-1}{n-K}R^2 \underbrace{- \frac{K-1}{n-K}R^2 + \frac{K-1}{n-K}R^2}_{=0} \\
&= 1 - \frac{n-1}{n-K} + R^2 + \frac{K-1}{n-K}R^2 \\
&= -\frac{K-1}{n-K} + R^2 + \frac{K-1}{n-K}R^2 \\
&= R^2 - \frac{K-1}{n-K}(1-R^2) \leq R^2 \quad \square
\end{aligned}
$$

Conclusion

- R^2 shows how well a model fits the observed data
- but: R^2 increases with the number of regressors eventhough they might not be relevant for the model
- therefore we need to adjust it and use $\bar{R}^2$