Ch. 6-3

Monday, February 24, 2025 10:38 AM

ch6-3

- R-square doesn't work.
- Why it doesn't work
- The fix : $\bar{R}^2$

6.5 Adjusted R-squared

We should no longer use R^2 in the multiple regression model. This is because when we add a new variable to the model, R^2 must always increase (or at best stay the same). This means that we could keep adding "junk" variables to the model to arbitrarily inflate the R^2. This is not a good property for a "measure of fit" to have. Instead, we will use "adjusted R-squared", denoted by $\bar{R}^2$.

$y = \hat{y} + e$

$s_y^2 = s_{\hat{y}}^2 + s_e^2$

$$R^2 = \frac{\text{variance we can explain}}{\text{total variance in } y} = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\frac{1}{n-1}\Sigma(\hat{y}_i - \bar{y})^2}{\frac{1}{n-1}\Sigma(y_i - \bar{y})^2} = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = \frac{ESS}{TSS}$$

6.5.1 Why R^2 must increase when a variable is added

To see why R^2 must always increase when a variable is added, we begin by looking again at the formula:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{TSS}$$

and again at the minimization problem that defines the OLS estimators:

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} e_i^2$$

$y = \beta_0 + \beta_1 X_1 + \epsilon$ TRUE

$\beta_2 X_2$ in

$y = b_0 + b_1 X_1 + e$

$b_2 X_2$

$$\uparrow R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\Sigma e_i^2 \downarrow}{TSS}$$

$e_i = y_i - \hat{y}_i$

$b_0 + b_1 X_1 + b_2 X_2$

When we add another X variable, the minimized value of $\sum_{i=1}^{n} e_i^2$ must get smaller! OLS picks the values for the bs so that the sum of squared vertical distances are minimized. If we give OLS another option for minimizing those distances, the distances have to get smaller (or at the worst stay the same). So, adding a variable means RSS decreases, so R^2 increases. The only way that R^2 stays the same is if OLS chooses a value of 0 for the associated slope coefficient, which never happens in practice.

$$b_1 = \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2} \longrightarrow \text{came from } \min \Sigma e_i^2$$

b_0, b_1, b_2

add X_2

As an example, let's try adding a nonsense variable to the house price model: random dice rolls. Using R, 1728 die rolls are simulated (to match the house price sample size of $n = 1728$), are recorded as a variable `Dice`, and added to the regression. Notice the difference in "Multiple R-squared" (R^2) and "Adjusted R-squared" ($\bar{R}^2$) between the two regressions:

```
summary(lm(Price ~ Fireplaces + Living.Area))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.730146   5.007563   2.942  0.00331 **
Fireplaces   8.962118   3.389066   2.644  0.00827 **
Living.Area  0.109312   0.003041  35.951  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68.98 on 1725 degrees of freedom
Multiple R-squared:  0.5095,    Adjusted R-squared:  0.5089
F-statistic: 895.9 on 2 and 1725 DF,  p-value: < 2.2e-16

summary(lm(Price ~ Fireplaces + Living.Area + Dice))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.105583   6.072084   1.994  0.04635 *
Fireplaces   8.829134   3.384696   2.609  0.00087 **
Living.Area  0.109576   0.003062  35.754  < 2e-16 ***
Dice         0.743706   0.972575   0.764  0.44469
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68.99 on 1724 degrees of freedom
Multiple R-squared:  0.5097,    Adjusted R-squared:  0.5088
F-statistic: 597.3 on 3 and 1724 DF,  p-value: < 2.2e-16
```

- R^2 always $\uparrow$ when a variable is added
- Why? $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\Sigma e_i^2}{TSS}$ ← LS is trying to get as small as possible

$$s_y^2 = \frac{\Sigma(y_i - \bar{y})^2}{n-1}$$

$e_i = y_i - \hat{y}_i \rightarrow \hat{y}_i = b_0 + b_1 X_1 + b_2 X_2 + \ldots$

$$\uparrow\downarrow \bar{R}^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\downarrow\Sigma e_i^2 / (n - k\uparrow - 1)}{\Sigma(y_i - \bar{y})^2 / (n-1)}$$

← d.o.f. correction

if you add a variable: $k\uparrow \rightarrow \bar{R}^2\downarrow$

$\Sigma e_i^2 \downarrow \rightarrow \bar{R}^2 \uparrow$

The variable `Dice` has no business being in the regression of house prices, and we fail to reject the null hypothesis that its effect is zero, yet the R^2 increases. The adjusted R-squared ($\bar{R}^2$) decreases, however.

6.5.2 The $\bar{R}^2$ formula

Adjusted R-squared ($\bar{R}^2$) is a measure-of-fit that can either increase or decrease when a new variable is added. $\bar{R}^2$ is a slight alteration of the R^2 formula. It introduces a penalty into R^2 that depends on the number of X variables in the model. (Remember that the number of Xs in the model is denoted by k.)

$$\bar{R}^2 = 1 - \frac{RSS \ / \ (n-k-1)}{TSS \ / \ (n-1)} \tag{6.4}$$

The $\bar{R}^2$ formula is such that when a variable is added to the model, k goes up, which tends to make $\bar{R}^2$ smaller. We know from the previous discussion, however, that whenever a variable is added, RSS must decrease. So, whether or not $\bar{R}^2$ increases or decreases depends on whether the new variable improves the fit of the model enough to beat the penalty incurred by k.

The justification for the $(n-k-1)$ and $(n-1)$ terms is from a degrees-of-freedom correction. How many things do we have to estimate before we can calculate RSS? $k+1$ $\hat{\beta}$s must first be estimated before we can get the OLS residuals, and RSS. If you want to use RSS for something else (such as a measure of fit), we recognize that we don't have n pieces of information left in the sample, we have $(n-k-1)$. A similar argument can be made for the $(n-1)$ term in equation 6.4.