14 Nov - Lec 9

Consider a generic multiple Gaussian linear model

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \qquad \underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 I_n)$$

$\underline{Y}$ vector of response variables

X $n \times p$ matrix of covariates

$\underline{\beta} = [\beta_1 \ \beta_2 \ \dots \ \beta_p]^T$ vector of regression parameters

We have seen the statistical tests to evaluate the model's adequacy:

- test about an individual coefficient

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$$

If I do not reject $\beta_j = 0$ for some j, I can remove that covariate from the model specification and estimate a new one with one less covariate but the same accuracy at predicting y.

- test about a subset of coefficients

$$\begin{cases} H_0: \beta_{p_0+1} = \beta_{p_0+2} = \dots = \beta_p = 0 \\ H_1: \text{at least one is} \neq 0 \end{cases}$$

similarly, if I do not reject H_0, I can remove that subset of covariates without losing accuracy

- test about the overall significance

$$\begin{cases} H_0: \beta_2 = \dots = \beta_p = 0 \\ H_1: \overline{H_0} \end{cases}$$

in this case, the model is useless.

R^2 and R^2_{adj} (adjusted R^2)

We have seen how the coefficient R^2 describes the proportion of variability explained by the model. Hence, we could think of using R^2 to choose between different model specifications.

However, if I use R^2 to compare nested models (i.e. one can be obtained starting from the other through a set of constraints), R^2 is not a valid measure.

consider: (a) $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

(b) $\tilde{y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 x_i + \tilde{\beta}_3 w_i$ | add one covariate

$R^2_{(a)} \leq R^2_{(b)}$ BY CONSTRUCTION!

Recall that $R^2 = \frac{SSR}{SST}$

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ does not depend on the model $\Rightarrow$ $SST_{(a)} = SST_{(b)}$

However, $SSR_{(b)} \geq SSR_{(a)}$

the SSR of model (b) can not be smaller than SSR (a).

If w_i is useful to predict y, $SSR_{(b)} > SSR_{(a)}$

In the worst case (if w_i is really useless), I set $\tilde{\beta}_3 = 0$ and I obtain $SSR_{(b)} = SSR_{(a)}$.

The more variables I include in the model, the larger R^2 will be.

So we can not use it to compare, for example, models (a) and (b) – or, in general, NESTED MODELS.

In general:

MORE COVARIATES $\Rightarrow$ R^2 increases; less interpretable; overfit

FEWER COVARIATES $\Rightarrow$ parsimony; interpretable

of course, we want few covariates, but not too few!

- ADJUSTED R^2 $\qquad R^2_{adj} = 1 - (1 - R^2) \cdot \frac{n-1}{n-p}$

it is "adjusted" for the model dim. p

penalizes models with many covariates.

when I introduce a new covariate:

- R^2 can remain the same or increase

R^2_{adj} can increase, remain the same, or decrease

$\Downarrow$

R^2_{adj} can be < 0!

$\Rightarrow$ To compare nested models we can use R^2_{adj}.