Statistics

Week 10: Regression (Chapter 10 & 11)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Outline

1 Variable selection

Variable/model selection

If we can reject $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$, then it remains to determine which subset of the predictor variables gives the best model. As mentioned, r^2 is no longer a good measure.

A basic approach is to look at the confidence interval for each β_i, and check if it contains 0 (alternatively, compare the p-value to α).

Example: do this for the *US economy* spreadsheet.

"Essentially, all models are wrong, but some are useful."

– *George Box*

Standardized regression coefficients

Another approach is to compare the effects of each predictor variable on y.

Suppose we have a regression $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. To compare x_1 and x_2 in terms of their effects on y, we cannot compare the sizes $\hat{\beta}_i$ directly, since they may be in different units.

One method is to standardize the data:

$$y_i' = \frac{y_i - \bar{y}}{s_y}, \qquad x_{ij}' = \frac{x_{ij} - \bar{x}_j}{s_{x_j}},$$

then perform the multiple regression.

(In simple linear regression, the new regression line is $\hat{y}' = r x'$.)

Exercise

For the spreadsheet *sales2*, show that x_1 has the larger effect.

Adjusted r^2

Given a subset of size p of the predictor variables x_i's, define the adjusted r^2 as

$$r^2_{adj} := 1 - \frac{n-1}{n-1-p}(1-r^2).$$

Then the subset of the x_i's which gives the highest adjusted r^2 can be considered the 'best' model.

This definition is motivated by the observation that a good model should fit the data well using few predictor variables, hence there is a penalty on the number of predictors used.

Exercise

Compute the adjusted r^2 for each of the 3 models for the spreadsheet *sales2*.

More information

The *Akaike information criterion* (AIC) is also commonly used for model selection; it measures the quality of each model relative to the others.

The total number of subsets grows quickly with k, so it is impractical to test for all subsets. *Stepwise regression* (textbook Section 11.7) uses a heuristic for finding a good subset quickly.

AIC and stepwise regression are implemented in R.