Theme: The least degrees of squares

TEACHER:
PROF. G. SHADMANOVA

- The method of least squares is often used to generate estimators and other statistics in regression analysis.

Consider a simple example drawn from physics. A spring should obey Hooke's lawwhich states that the extension of a spring yis proportional to the force, F, applied to it.

$$y = f(F, k) = kF$$

constitutes the model, where F is the independent variable.

To estimate the force constant, k, a series of n measurements with different forces will produce a set of data

$$(F_i, y_i),\ i = 1, \ldots, n,$$

where y is a measured spring extension. Each experimental observation will contain some error. If we denote this error we may specify an empirical model for our observations, ε,

There are many methods we might use to estimate the unknown parameter k.

$$y_i = kF_i + \varepsilon_i.$$

- Noting that the n equations in the m variables in our data comprise an overdetermined systemwith one unknown and n equations, we may choose to estimate k using least squares. The sum of squares to be minimized is

$$S = \sum_{i=1}^{n} (y_i - kF_i)^2 .$$

- The least squares estimate of the force constant, k, is given by

$$\hat{k} = \frac{\sum_i F_i y_i}{\sum_i F_i^2} .$$

Here it is assumed that application of the force causes the spring to expand and, having derived the force constant by least squares fitting, the extension can be predicted from Hooke's law.

- In regression analysis the researcher specifies an empirical model. For example, a very common model is the straight line model which is used to test if there is a linear relationship between dependent and independent variable. If a linear relationship is found to exist, the variables are said to be correlated. However, correlation does not prove causation, as both variables may be correlated with other, hidden, variables, or the dependent variable may "reverse" cause the independent variables, or the variables may be otherwise spuriously correlated. For example, suppose there is a correlation between deaths by drowning and the volume of ice cream sales at a particular beach.

- In order to make statistical tests on the results it is necessary to make assumptions about the nature of the experimental errors. A common (but not necessary) assumption is that the errors belong to a normal distribution. The central limit theorem supports the idea that this is a good approximation in many cases.
- In a linear model, if the errors belong to a normal distribution the least squares estimators are also the maximum likelihood estimators

In a least squares calculation with unit weights, or in linear regression, the variance on the jth parameter, denoted usually estimated with $\mathbf{var}(\hat{\beta}_j)$, is

$$\mathbf{var}(\hat{\beta}_j) = \sigma^2([X^T X]^{-1})_{jj} \approx \frac{S}{n-m}$$

where the true error variance σ2 is replaced by an estimate based on the minimised value of the sum of squares objective function S. The denominator, n − m, is the statistical degrees of freedom; see effective degrees of freedomfor generalizations.

Thanks for attention