Computer Class Model Identification and Data Fitting Linear Regression

Pietro Bonizzi & Ralf Peeters

You have seen in class that the method of least squares is about estimating parameters by minimizing the squared discrepancies between observed data, on the one hand, and their expected values on the other. You have seen the method in the context of a regression problem, where the variation in one variable, called the response variable y , can be partly explained by the variation in the other variables, called covariables x_i (Multivariable Linear Regression). For example, variation in exam results are mainly caused by variation in abilities and diligence of the students, or variation in survival times are primarily due to variations in environmental conditions.

Exercise 1 Load the .mat file `dataMIDFLR1` into Matlab. The variable y is supposed to depend on x by means of the following model:

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_t^2 + e_t$$

in which e_t is white noise having a standard normal distribution.

(a) Find the least square estimator of the parameter vector β. Then, plot on the same graph both the original data y and the expected ones, given by $\hat{y} = \mathbf{\Phi}\beta$. What do you notice?

The covariance matrix of an estimator gives an impression of the reliability of the estimation outcomes and it makes it possible to construct a confidence area for the location of the true parameter vector. (The construction of such a confidence area proceeds analogously to the construction of, for instance, a 95% confidence interval on the basis of the estimated variance or standard deviation of a scalar stochastic variable.)

In order to construct a confidence interval for the components of β one needs an estimator of the covariance matrix of $\hat{\beta}_{LS}$. Given $\mathbf{\Phi}$, the covariance matrix of the estimator $\hat{\beta}_{LS}$ is equal to

$$\text{cov}(\hat{\beta}_{LS}) = \sigma^2(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}$$

where σ^2 is the variance of the noise e_t, for which an unbiased estimator is given by:

$$s_{LS}^2 = \frac{2}{N-n} V(\hat{\beta}_{LS}).$$

Here, N is the amount of data available, n is the number of parameters in the model, and $V(\hat{\beta}_{LS})$ is the exploited criterion function. Therefore, the estimate of the variance of a component $\hat{\beta}_j$ is

$$\text{var}(\hat{\beta}_j) = \tau_j^2 s_{LS}^2$$

where τ_j^2 is the jth element on the diagonal of $(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}$. A confidence interval for β_j can then be obtained by taking the least square estimator $\hat{\beta}_j \pm$ a margin:

$$\hat{\beta}_j \pm c\sqrt{\text{var}(\hat{\beta}_j)}$$

where c depends on the chosen confidence level. For a 95% confidence interval, the value $c = 1.96$ is a good approximation when N is large. For smaller values of N, one usually takes a more conservative c using the tables for the Student t-distribution with $N - n$ degrees of freedom.

(b) Calculate the covariance matrix of the estimator β, namely $\text{cov}(\hat{\beta}_{LS})$.

(c) Exploit $\text{cov}(\hat{\beta}_{LS})$ to obtain a 95% confidence interval for all components $\hat{\beta}_j$. Then, test if each $\hat{\beta}_j$ is significantly different from zero at the 5% level, and, hence, if the null hypothesis $H_0 : \ \hat{\beta}_j = 0$ can be rejected. What do you notice? What does it means in terms of the original model assumed to describe the observed data?

(d) Build a new model by retaining only the significant $\hat{\beta}_j$, and estimate the new vector of parameters $\hat{\beta}_{LS}$. Compare the new values obtained for the components of $\hat{\beta}_{LS}$ with the old ones. What do you notice? Why?

Exercise 2 Repeat all steps of Exercise 1, but now using the variables stored in the .mat file `dataMIDFLR1_2`. What do you notice with respect to the previous situation? Why?

Every time you plot a linear regression line, you can plot as well the corresponding confidence region. This can be obtained as follows: given an estimate y_t, the corresponding confidence region can be obtained as:

$$y_t \pm c\sqrt{\text{Var}[y_t]},$$

where

$$\text{Var}[y_t] = \phi_t^T \, \text{cov}(\hat{\beta}_{LS})\phi_t.$$

This computation is repeated at each time point t. You should obtain plots like the one below, where the red line represents the regression model, and the green lines mark the confidence region.

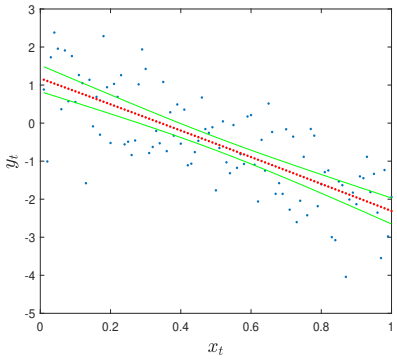

Figure 1: Example of plot including regression line (red) and corresponding confidence region (green lines).