EC475 Problem set 1

Data handling and mechanics of OLS

Antoine Goujard

23/10/09

- Antoine Goujard,a.j.goujard@lse.ac.uk ;
- Office hours: S684, Wednesday 12.30 – 13.30 ;
- Class webpage:
  http://personal.lse.ac.uk/goujard/
  Under Teaching/EC475.
  1. Introduction to Gauss (in progress) ;
  2. Examples of codes (in Gauss and Stata).
- Useful references for STATA programming (nothing is required):
  1. Baum, An Introduction to Stata Programming ;
  2. Cameron and Trivedi, Microeconometrics using STATA.
- Useful references for GAUSS programming:
  1. http://www.aae.wisc.edu/aae637/gausscode.htm ;
  2. Gauss user's guide (Aptech) ;

According to the lecture notes, we set up (for $s \in \{1, ..., S\}$):

$$y_s = \beta_1.x_{s1} + ... + \beta_k.x_{sk} + \epsilon_s$$

Or in matrix notations: $y_s = \boldsymbol{x}'_s.\boldsymbol{\beta} + \epsilon_s$ where $\boldsymbol{x}_s, \boldsymbol{\beta}$ are $k \times 1$ vectors.

So that stacking the S equations: $\boldsymbol{y} = \boldsymbol{X}.\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_S \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_S \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}'_1 \\ \boldsymbol{x}'_2 \\ \vdots \\ \boldsymbol{x}'_S \end{pmatrix}$$

The OLS estimator minimizes the RSS (wrt $\boldsymbol{b}$):

$RSS(\boldsymbol{b}) = \sum_{s=1}^{S}(y_s - \boldsymbol{x}'_s.\boldsymbol{b})^2 = (\boldsymbol{y} - \boldsymbol{X}.\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}.\boldsymbol{b})$

Under A_1 (full column rank), we have $\hat{\beta}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

We can then define: $\hat{\epsilon}_s := y_s - \boldsymbol{x}'_s.\hat{\beta}_{OLS}$ and $\hat{\varepsilon}$.

And we can compute the indicators of this PS:

- The $RSS_{OLS} = RSS = \hat{\varepsilon}'\hat{\varepsilon}$

- If a constant is included, the $R^2 = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{(\boldsymbol{y}-\overline{\boldsymbol{y}})'(\boldsymbol{y}-\overline{\boldsymbol{y}})} := 1 - \frac{RSS}{TSS^*}$ (rk: $\overline{\boldsymbol{y}} := \overline{y}\boldsymbol{\imath}$).

- The estimated regression variance: $s^2_{OLS} = s^2 = RSS \times 1/(S-k)$ (rk: k includes the constant).

Lqdata.dat is a panel of 10000 observations corresponding to individuals (uniqid) over time (yearcur).
The variables from lqdata.dta are:

- uniqid
- yearcur
- choice12
- age
- race
- dispy
- constant

Stata command: ***reg y x*** reports the centered $R^2 = R_c^2$ which corresponds to what we want to compute. This is "stored" in ***e(r2)***.

But: ***reg y x, noconstant*** reports the un-centered $R^2 = R_u^2$ instead of the centered R_c^2. This is "stored" in ***e(r2)***.

This does not corresponds to our definition of the R^2.

$R_u^2 = ESS/TSS$ is based on the decomposition:
$TSS = ESS + RSS$ or $\boldsymbol{y}'.\boldsymbol{y} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \hat{\varepsilon}'\hat{\varepsilon}$

$R_c^2 = 1 - RSS^*/TSS^*$ is based on the decomposition:
$TSS^* = ESS^* + RSS^* = ESS^* + RSS = (\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}})'(\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}}) + \hat{\varepsilon}'\hat{\varepsilon}$
(If the regression includes a constant: $\bar{\hat{\varepsilon}} = 0$ (handout1, p4.)).

The reason to use the centered R_c^2 instead of R_u^2 is that it will be invariant to a re-scaling of $\boldsymbol{y}$ by adding a constant α to each observation while it is not the case for R_u^2.
(If we have a constant in x, $R_u^2(\alpha) = \frac{||P_x(\boldsymbol{y}+\alpha.\imath)||^2}{||\boldsymbol{y}+\alpha.\imath||^2} = \frac{||P_x.\boldsymbol{y}+\alpha.\imath||^2}{||\boldsymbol{y}+\alpha.\imath||^2} \rightarrow_{\alpha\rightarrow\infty} 1$).

Gauss