How to detect an outlier (whether or not it exists)

by Bo E. Honoré[1]

Consider the model

$$y_i = x_i'\beta + \gamma d_i + \varepsilon_i$$

where (x_i, ε_i) is i.i.d. $E[\varepsilon_i | x_i] = 0$ and

$$d_i = \begin{cases} 1 & \text{if} \quad i = 1 \\ 0 & \text{if} \quad i \neq 1 \end{cases}$$

It is tempting to test whether the first observation is an outlier by tesing whether $\gamma = 0$. As we will see, this can be a good idea if you *want* to conclude that it is, but not necessarily if you actually want to know "the truth."

Let $z_i = (x_i, d_i)$. The OLS estimator of (β, γ) minimizes

$$\sum_{i=1}^{n} (y_i - x_i'b - gd_i)$$

It is clear that the solution to this would be

$$\begin{aligned} \widehat{\gamma} &= y_1 - x_1'\widehat{\beta} \\ \widehat{\beta} &= \left(\sum_{i=2}^{n} x_i x_i'\right)^{-1} \sum_{i=2}^{n} x_i y_i \end{aligned}$$

hence $\widehat{\beta}$ is consistent etc.

The heteroskedasticity–consistent standard errors of the OLS estimator of β and γ are the square roots of the diagonal of

$$\begin{aligned} \widehat{V} &= \left(\sum_{i=1}^{n} z_i z_i'\right)^{-1} \left(\sum_{i=1}^{n} e_i^2 z_i z_i'\right) \left(\sum_{i=1}^{n} z_i z_i'\right)^{-1} \\ &= \begin{pmatrix} \sum_{i=1}^n x_i x_i' & \sum_{i=1}^n x_i d_i \\ \sum_{i=1}^n d_i x_i' & \sum_{i=1}^n d_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n e_i^2 x_i x_i' & \sum_{i=1}^n e_i^2 x_i d_i \\ \sum_{i=1}^n e_i^2 d_i x_i' & \sum_{i=1}^n e_i^2 d_i^2 \end{pmatrix} \\ &\qquad \begin{pmatrix} \sum_{i=1}^n x_i x_i' & \sum_{i=1}^n x_i d_i \\ \sum_{i=1}^n d_i x_i' & \sum_{i=1}^n d_i^2 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \sum_{i=1}^n x_i x_i' & x_1 \\ x_1' & 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=2}^n e_i^2 x_i x_i' & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n x_i x_i' & x_1 \\ x_1' & 1 \end{pmatrix}^{-1} \end{aligned}$$

[1] Comments should be emailed to bo@honore.com

because $e_1 = 0$ (e_i is the residual for the i'th observation)

To simplify the notation, let

$$\begin{pmatrix} \sum_{i=1}^{n} x_i x_i' & x_1 \\ x_1' & 1 \end{pmatrix}^{-1} = \begin{pmatrix} a & b \\ b' & c \end{pmatrix}$$

then

$$\begin{aligned} \widehat{V} &= \begin{pmatrix} a & d \\ d' & c \end{pmatrix} \begin{pmatrix} \sum_{i=2}^{n} e_i^2 x_i x_i' & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a & d \\ d' & c \end{pmatrix} \\ &= \begin{pmatrix} a\left(\sum_{i=2}^{n} e_i^2 x_i x_i'\right) a & a\left(\sum_{i=2}^{n} e_i^2 x_i x_i'\right) d \\ d'\left(\sum_{i=2}^{n} e_i^2 x_i x_i'\right) a & d'\left(\sum_{i=2}^{n} e_i^2 x_i x_i'\right) d \end{pmatrix} \end{aligned}$$

The estimated variance of $\widehat{\gamma}$ is $d'\left(\sum_{i=2}^{n} e_i^2 x_i x_i'\right) d$. Clearly the term in the middle is of order n. Moreover

$$d = x_1' \left(\sum_{i=1}^{n} x_i x_i' - x_1 x_1' \right)^{-1}$$

which is of order $\frac{1}{n}$. It therefore follows that the estimated variance of $\widehat{\gamma}$ is of order $\frac{1}{n}$ and since $\widehat{\gamma}$ will converge to $\gamma + \varepsilon_1$ this means that (unless $\gamma + \varepsilon_1$ happens to equal 0) the calculated t–statistic will converge to $\pm\infty$ with probability 1 whether or not $\gamma = 0$.