Intervention dummies and OLS residuals

Walter Sosa Escudero*

July 2020

1 The result

Consider the following linear model

$$y_i = \beta\, x_i + \gamma\, d_i + u_i, \qquad\qquad i = 1, \ldots, n$$

where x_i is a scalar regressor and $d_i = 1[i = s]$; there is no intercept. d_i is a dummy variable that 'singles out' the $s-$th observation. Let $\hat{\gamma}$ be the OLS estimator of γ from regressing y_i on x_i and γ, and let e_i be the OLS residuals of regressing y_i on x_i without including d_i.

Result:

$$\hat{\gamma} = \frac{e_s}{1 - h_s},$$

with $h_s \equiv x_s^2 / \sum_{i=1}^n x_i^2$

Proof: by the Frisch-Waugh-Lovell theorem

$$\hat{\gamma} = \frac{\sum d_i^* e_i}{\sum d_i^{*2}}, \tag{1}$$

where d_i^{*2} are residuals from regressing d_i on x_i; all summations run from $i = 1$ to n. Now $d_i^* = d_i - \hat{\delta} x_i$ with

$$\hat{\delta} = \frac{\sum x_i d_i}{S_x},$$

with $S_x \equiv \sum x_i^2$. Replacing

*Universidad de San Andres and CONICET. Email: wsosa@udesa.edu.ar I thank Martin Rossi for motivating this pedagogical note. This note is part of a series of vignettes and curiosities in statistics and econometrics. The usual disclaimer applies.

$$d_i^* = d_i - \frac{x_s x_i}{S_x}.$$

Now, the numerator of (1) is

$$\begin{aligned}
\sum d_i^{*2} &= \sum \left(d_i - x_s x_i / S_x\right)^2 \\
&= \sum \left(d_i^2 + \left(x_s x_i\right)^2 / S_x - 2 d_i x_s x_i / S_x\right) \\
&= 1 - x_s^2 / S_x
\end{aligned}$$

The denominator of (2) is:

$$\begin{aligned}
\sum d_i^* e_i &= \sum \left(d_i - x_s x_i / S_x\right) e_i \\
&= e_s - x_s \left(\sum x_i e_i\right) S_x \\
&= e_s,
\end{aligned}$$

since $\sum x_i e_i = 0$ by the first order conditions of the OLS problem of regressing y on x. Replacing in (1), the desired result follows.

The generalization to the case with an intercept and K regressors proceeds by a further application of the Frisch/Waugh/Lovell theorem.

2 Motivaton

The result is standard in the literature, usually obtained when studying the effect of a single observation on OLS results; see, for example, Davidson and MacKinnon (2009) for a detailed discussion. Nevertheless, it is interesting to appreciate it from the perspective of an impact evaluation problem. Suppose $i = 1, \ldots, n$ denotes the passage of time and that at period s there is one-time intervention that alters the linear relation between y_i and x_i. Intuitively, the intervention at s produces and outlier, and the main goal of the analysis is to estimate the size of this intervention. A simple model that includes an 'intervention dummy' (like the one discussed previously) identifies the desired effect, and OLS produces unbiased estimates under standard conditions.

A relevant question is whether such interventions can by explored by looking at residuals of regressing y_i on x_i *without* including an intervention dummy. The result suggest that the answer is generally no, and it is related the way OLS weighs errors according to x_i.

It is easy to verify (see Davidson and MacKinnon, 2009) that $\hat{\gamma}$ equals the OLS error of predicting y_s from a regression of y on x when the $s-$th observation is omitted. According to the discussed result, when $h_i = 0$, $\hat{\gamma}$ also coincides with the error of a regression that includes the $s-$th observation, then the effect of the intervention can be alternatively captured by the OLS estimate of the

dummy ($\hat{\gamma}$) or by the residual of the regression with all observations. But when $h_s \neq 0$, $\hat{\gamma}$ differs from the residual of the regression with all points. Intuitively h_s measures how large is x_s with respect to the total variability of x. When x_s is in the center of its variability, the OLS estimator gives zero weight to such observations, hence omitting it has no effect in the regression, so errors of omitting the $s-$th observation (always equal to $\hat{\gamma}$) are equal to those using all data. On the contrary, when x is far from its center, the OLS residuals now give a large weight to such observation, reducing the error of the s-th observation when all data is used. Consequently, full data residuals reflect interventions more appropriately when they take place when the corresponding x_s is closer to the center of the data, and get 'masked' by the regression otherwise.

This material is standard in the outlier/robustness literature, but it is illuminating to observe it from the perspective of the more recent impact evaluation perspective. Rather surprisingly, basic (and even advanced) econometric texts give little space to the study of influential observations, at least compared to most branches of applied statistics. The purpose of this note is to provide an alternative route to appreciate the relevance of outlier analysis in basic econometrics.

3 References

Davidson, R. and MacKinnon, J., 2004, *Econometric Theory and Methods*, Oxford University Press, Oxford.