Peter Hull March 30, 2018

On Residualized Outcome Regressions

Since the seminal work of Frisch and Waugh (1933) and Lovell (1963), researchers have known that the coefficients of a multivariate regression can be obtained by regressing the outcome on a residualized regressor − specifically, the residual from projecting the regressor on all other right–hand side variables. Occasionally researchers will flip this logic, regressing instead a residualized outcome on a set of non-residualized treatments. This is sometimes done to simplify computation of high-dimensional models, especially if the treatments are group indicators. For example, computing classroom-level averages of residualized test score outcomes can be significantly easier and faster than estimating a large-scale teacher value-added model in a single step. Other times researchers may use residualized outcomes to visualize a study's identifying variation, for example by plotting group-by-time trends of outcome residuals in a difference-in-differences design.

This note shows that these sorts of residualized outcome regressions can be difficult to interpret, especially in settings with multiple treatments. Residualizing outcomes is harmless when the partialled-out controls are independent from treatment, as in a randomized control trial or regression discontinuity design; in some cases this may even increase precision (Lee and Lemieux, 2010). In general, however, regressing a residualized outcome on a single treatment variable yields an attenuated estimate of the "true" regression coefficient, while multivariate residualized outcome regressions identify particular linear combinations of the true coefficients. Thus in the above difference-in-differences example the coefficients may mix together multiple different true leads and lags, complicating the interpretation of the residualized outcome plot.

Formally, suppose our population regression of interest is

$$Y_i = \alpha + D_i'\beta + X_i'\gamma + \epsilon_i, \tag{1}$$

where D_i contains a set of J treatment variables and X_i is a vector of auxiliary controls. Typically the "true" treatment parameter β is estimated by ordinary least squares; in matrix notation and by the classic theory of Frisch and Waugh (1933) and Lovell (1963), this can be written

$$\hat{\beta} = (\tilde{D}'\tilde{D})^{-1}\tilde{D}'Y. \tag{2}$$

Here Y and $\tilde{D}$ collect observations of Y_i and $\tilde{D}_i'$, where $\tilde{D}_i$ denotes the residuals from projecting D_i on X_i and a constant. In contrast, a residualized outcome regression estimate can be written

$$\tilde{\beta} = (\underline{D}'\underline{D})^{-1}\underline{D}'\tilde{Y}, \tag{3}$$

where $\underline{D}$ and $\tilde{Y}$ collect observations of de-meaned D_i' and the residuals from regressing Y_i on X_i and a constant (note here D_i is de-meaned to account for the constant in the second step regression).

To link these two estimates let $M_X = I - X(X'X)^{-1}X'$ be the control residual-maker matrix, where I denotes the identity matrix and X collects observations of X_i' and a constant. Since $\tilde{Y} = M_X Y$, $M_X\underline{D} = \tilde{D}$, and M_X is both symmetric and idempotent, we have

$$\begin{aligned}\tilde{\beta} &= (\underline{D}'\underline{D})^{-1}(M_X\underline{D})'Y \\ &= \hat{\Omega}\hat{\beta},\end{aligned} \tag{4}$$

where $\hat{\Omega} = (\underline{D}'\underline{D})^{-1}(\tilde{D}'\tilde{D}) = (\underline{D}'\underline{D})^{-1}(\underline{D}'\tilde{D})$ is a $J \times J$ matrix containing the coefficients from regressing the elements of $\tilde{D}_i$ on $\underline{D}_i$. With $\hat{\beta} \xrightarrow{p} \beta$, we will generally not also have $\tilde{\beta} \xrightarrow{p} \beta$ unless $\hat{\Omega} \xrightarrow{p} I$. Of course the residualized outcome regression is consistent for β when X_i is uncorrelated with D_i, as in a randomized trial, since then $\tilde{D}_i$ and $\underline{D}_i$ coincide asymptotically.

To unpack this result, first suppose that we have a single treatment ($J = 1$). Then

$$\begin{aligned}\hat{\Omega} &= \frac{\sum_i \tilde{D}_i^2}{\sum_i \underline{D}_i^2} \\ &= 1 - \hat{R}^2 \in (0,1), \end{aligned} \tag{5}$$

where $\hat{R}^2$ is the sample R-squared from regressing D_i on X_i. Thus in the single-treatment case the residualized outcome regression gives an attenuated estimate of β while preserving the sign of $\hat{\beta}$. This is, in fact, classic attenuation bias: the residual outcome regression uses a mismeasured regressor D_i in place of the true regressor $\tilde{D}_i$, with uncorrelated measurement error $D_i - \tilde{D}_i$.

Unfortunately, the bias from $\hat{\Omega}$ becomes more complicated when there are multiple maintained treatments. Unless $\hat{\Omega}$ is diagonal, the estimates $\tilde{\beta}_j$ will mix together multiple true coefficient estimates $\hat{\beta}_k$, and may thus not even be of the right sign. To see this most simply, suppose D_i contains a set of mutually-exclusive treatment indicators $D_{i1}, \dots D_{iJ}$ with one control group indicator D_{i0} omitted. Writing the auxiliary regression of D_i on X_i as

$$D_i = \mu + \Gamma X_i + \nu_i, \tag{6}$$

we have the (j,k)th element of the matrix $\hat{\Omega}$ satisfying

$$\begin{aligned}\hat{\Omega}_{jk} &\xrightarrow{p} E[D_{ik} - \mu - \Gamma_k X_i \mid D_{ij} = 1] - E[D_{ik} - \mu - \Gamma_k X_i \mid D_{i0} = 1] \\ &= \mathbf{1}\{k = j\} - \Gamma_k \left(E[X_i \mid D_{ij} = 1] - E[X_i \mid D_{i0} = 1] \right), \end{aligned} \tag{7}$$

where Γ_k denotes the kth row of Γ. Thus with many group treatments the residualized outcome regression estimate of β_j will be contaminated by $\hat{\beta}_k$ for $k \neq j$ unless either the controls are balanced across treatments j and 0 (in which case the term in parentheses is zero) or are uncorrelated with treatment k (in which case $\Gamma_k = 0$). Both conditions are again satisfied when X_i and D_i are independent, as in a randomized trial.

Except in special cases, residualized outcome regressions are therefore unlikely to capture the underlying regression parameters of interest, or even their sign. Difference-in-difference trends in residualized outcomes will in general mix together multiple true leads and lags, while two-step teacher value-added estimation may misattribute one classroom's test score improvements to another. Researchers hoping to reduce the computational burden of large-scale regressions, or to simply illustrate a research design, may wish to avoid this two-step procedure in favor of estimating and plotting conventional regression coefficients. For this the classic theory of Frisch and Waugh (1933) and Lovell (1963), as well as more recent advances in high-dimensional regression estimation (Abowd et al., 2002; Guimaraes and Portugal, 2010; Correia, 2016), may provide alternative simplifying tools.

References

Abowd, J., R. Creecy, and F. Kramarz (2002). "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data," Longitudinal Employer-Household Dynamics Technical Paper.

Correia, S. (2016). "A Feasible Estimator for Linear Models with Multi-way Fixed Effects," Working Paper.

Frisch, R. and F. V. Waugh (1933). "Partial Time Regressions as Compared with Individual Trends," *Econometrica*, 1(4), 387–401.

Guimaraes, P. and P. Portugal (2010). "A Simple Feasible Procedure to Fit Models with High-Dimensional Fixed Effects," *Stata Journal*, 10, 628–649.

Lee, D. S. and T. Lemieux (2010). "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48, 281-355.

Lovell, M. (1963). "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis," *Journal of the American Statistical Association*, 58(304), 993–1010.