Comments on "Quantifying Heterogeneous Returns to Genetic Selection: Evidence from Wisconsin Dairies" by Jared Hutchins, Brent Hueth, and Guilherme Rosa

Paul T. Scott
March 27, 2020

Introduction

Since Henderson [1963], the contribution of a dairy bull to his daughters’ productivity has typically been treated as a random effect, meaning it is assumed to be uncorrelated with any unobserved determinants of the daughter’s productivity. However, it’s possible that farmers select sires that are appropriate matches based on unobserved characteristics of their farms or the dams in question. Indeed, correlation between producer’s input decisions and unobserved productivity has long been a central concern in the econometrics of production functions [Marschak and Andrews, 1944].

Hutchins, Hueth, and Rosa (henceforth, HHR) use instrumental variables to assess whether the measured productivity of sires is affected by selection. The main equations of interest have the form

\[ y_{ijt} = \mu_{ij} z_{ij} + \beta X_{ijt} + \epsilon_{ijt}, \]

where \( i \) indexes farms, \( j \) indexes cows, and \( t \) denotes the month. The dependent variable \( y_{ijt} \) measures cow \( j \)'s production, and the regressor of interest \( z_{ij} \) is the Predicted Transmitting Ability (PTA) of cow \( j \)'s sire. In one version of the regression equation, \( y \) is the cow’s production of butterfat, and \( z \) is the sire’s PTA for butterfat. In the other version, \( y \) and \( z \) are the analogous values for protein production.

Instruments

In the paper’s original draft, the average treatment effect of sire PTA was negative after controlling for selection. While HHR anticipate that selection will create upward bias, it’s not plausible that the true effect would be negative. A sire’s PTA value is based on the observed production of his biological daughters.
A sire with an above-average PTA value will have daughters with above-average production. For the true effect of PTA to be negative, the market would need to be in an equilibrium in which beliefs are systematically wrong about who the best sires are: somehow, relatively unproductive sires end up being paired with highly productive farms and/or the best dams so that sires with undesirable genetic traits end up looking like they have desirable genetic traits.

The authors originally instrumented for the sire’s PTA $z_{ij}$ using the lagged nutrient prices. Careful consideration of this instrument can help explain why it would lead to the implausible finding of negative impacts of the sire’s PTA. Given that farmers might take steps to increase production of a nutrient when its price is high, and given that commodity prices tend to be serially correlated, we might worry that such instruments are not exogenous.

Indeed, by thinking about the properties of the instrumental variables estimator, we can also understand why the use of lagged nutrient prices as instruments would lead to the implausible result of negative treatment effects. Recall that the instrumental variables estimator is consistent for $\frac{\operatorname{Cov}(w, z)}{\operatorname{Cov}(w, y)}$, where $w$ is the instrument, $z$ is the explanatory variable, and $y$ is the dependent variable. We would expect a negative covariance between nutrient prices and sire characteristics ($\operatorname{Cov}(w, z) < 0$) just by the law of demand. Furthermore, we would expect farmers to take steps to increase nutrient production when the nutrients get expensive, meaning $\operatorname{Cov}(w, y) > 0$. Thus, we should expect the IV estimator with nutrient price instruments to deliver a negative coefficient for the effect of sire characteristics.

Fortunately, better instruments are available. A sire’s measured PTA value is based on a finite sample and updated as the production of more daughters is observed. Thus, the observed PTA value becomes a better and better measure of the sire’s true characteristic over time. At a given point in time, a farmer making a decision only has access to the current PTA value. Thus, information revealed about sire $s$ after a farmer chooses him will be correlated with his true PTA value but plausibly uncorrelated with anything influencing farmer’s decision. This means that the change in a sire’s PTA value after a farmer’s choosing that sire can be used as a valid instrument for the sire’s PTA value.\footnote{Note that the econometrician may as well use the most up-to-date PTA value in the regression equation, as this is presumably the most accurate measure of the sire’s true characteristic. The difference between this up-to-date PTA value and the observable PTA value at the time the farmer made the decision is plausibly a valid instrument because there is no way the farmer could have predicted how the PTA value would be updated. It’s also a plausibly relevant instrument because the updates to a sire’s PTA value will naturally be positively correlated with the final value and true value.}

In the latest draft, the use of these instruments leads to the more plausible results of positive effects of sire PTA on daughter productivity, but there is still evidence of selection based on unobserved productivity differences. Ordinary least squares seems to overstate the impact of a sire’s PTA.
Other concerns

The main regression equations include other choice variables as controls: herd size, lactation length, and a measure of milking frequency. These variables may also suffer from the standard simultaneity problem with unobserved productivity, and if they are also correlated with PTA values, then we may still obtain a biased estimate of PTA values. In principle, instruments are needed for all endogenous explanatory variables, not only the explanatory variable of interest. HHR are interested in regression equations for both protein and butterfat production, but they estimate these equations separately. More asymptotically efficient estimates could presumably be obtained through the use of joint estimation such as three-stage least squares.

Looking ahead

The paper presents an econometric approach to dealing with selection in measurement of the impact of sire’s characteristics on daughters’ productivity. However, one might wonder if better measurement of sire’s characteristics would eliminate the selection problem – i.e., we might hope to eliminate an omitted variables problem by measuring the omitted variables. Detailed measurement of genetic characteristics has been a feature of the dairy industry for over a decade now. However, this does not mean that selection problems will necessarily go away. Even when sires' genotypes are perfectly measured, there may be correlation between genetic characteristics and unobserved farm-level determinants of productivity, creating a selection problem in quantifying the impact of genetic characteristics. In other words, better measurement of sire characteristics does not solve the omitted variables problem if the omitted variables are correlated with farm-level characteristics. Genetic characteristics have displayed impressive but imperfect accuracy in predicting daughter productivities [Wiggans et al., 2011], suggesting there is at least some room for approaches like HHR’s to improve the quantification of the causal effect of sire characteristics.

References

