Quantifying Heterogeneous Returns to Genetic Selection:
Evidence from Wisconsin Dairies

Introduction

The productivity of the agriculture sector has been growing for most of the 20th century, and the dairy industry is no exception; specifically, per cow milk yield has been growing nearly 3-4% per year and 50% of this growth is attributed to genetic selection (Thornton, 2010; Wang et al., 2015). Selective breeding of dairy cattle as a vector of productivity growth has been facilitated by two main factors. First, genetic selection has been made possible at a large scale by breeding and herd testing associations that collect massive amounts of data on animal production and lineage. These data are used for isolating the contribution of genetics to cow productivity to rank and price different genetics using a variety of statistical tools. Second, the invention of artificial insemination (AI) has made nearly any of these tested bulls available around the country. Combined, these two sector innovations have made sire evaluation through herd testing a huge source of productivity growth.

The statistical methods used to evaluate sire genetics in the industry, however, have tended to ignore the role of management in the productivity of genetics. In fact, genetics is an input whose productivity depends directly on managerial ability and farm environment. In animal science, this is known as "genotype by environment interaction," (GxE) and in the economics literature on technology adoption as heterogeneous returns to technology. (Suri, 2011; Mundlak et al., 2012) Management ability matters because it affects how a certain genotype expresses in a given physical environment. Feeding environment, climate, and herd level production all have been shown to affect how certain sires perform in different environments (Kearney et al., 2004; Hayes et al., 2003). By ignoring these effects, genetic evaluations of sires do not currently capture the full extent to which management and genetics interact or how productive a given sire would be in different environments. This limits both our knowledge of how productivity has grown in the dairy sector and the usefulness of breeding values to managers who want to know the variance of certain sires across physical environments and, most importantly, management practices.

We investigate heterogeneous returns to genetics in the dairy industry using herd testing data and explore the extent to which management practices can hinder or enhance the productivity of certain sires. We define the cow level production function as a Correlated Random Coefficients (CRC) model of Mundlak (1978) and Heckman and Vytlacil (1998) where there is an average return but also a farm level return that is random. Using the CRC model, we use a control function approach with instrumental variables to identify farm specific returns to genetics and thus the distribution of returns across farms. After estimating farm specific returns to genetics, we use unsupervised machine learning techniques like K-means clustering to divide farms by management practices from herd testing data (Bonhomme and Manresa, 2015).

After explaining the methodology and data, we present preliminary results on the heterogeneity of returns in using sires with high predicted transmitting ability (PTA) in Wisconsin. Using the coarsest degree of heterogeneity possible, herd level dummy variables, we estimate herd specific returns to the amount of PTA in the chosen sire. We find that there is significant heterogeneity in returns from increasing PTAs across both fat and protein at the lactation level.

Theory and Methodology

The seminal model of sire evaluation, the Henderson mixed model, describes a progenys trait y as a function of its sire Z and its environment X:

$$y = X\beta + Zu + e$$

where β is a fixed coefficient vector and u is a random coefficient vector. Each sire has an estimated draw from u which is divided in half to get the PTA for the trait y. The matrix X contains factors that are systematically affecting progeny performance, and usually includes an intercept specific to that herd. The matrix Z is an incidence matrix for sires but the effects u are considered realizations from a random variable. This model assumes that $Cov(u, e) = 0$, which is typically violated when sires are not randomly selected by farmers. Henderson (1975) details the effect of sire selection on observable characteristics that violates $Cov(u, e) = 0$ but depends on knowing how selection is made on the observed characteristics (Robinson, 1991). Unfortunately, this does not help when, as can often be the case, selection is done on unobservables; such endogeneity is partly addressed by specifying relationships between different sires, but still requires detailed relationships be known for all of the sires in the sample.

The economics literature frequently deals with the problem of endogenous technology choice and heterogeneous returns, often by focusing on choice of inputs being correlated to an unobserved productivity shock in the error term. In the cow production function, the selection of sires can be endogenous to the herd's specific return for that sire, similar to how choice of schooling is endogenous to individual returns to schooling (Heckman and Vytlacil, 1998) and choice of seeds is endogenous to farm specific returns to those seeds (Suri, 2011).

Suppose we rewrite the returns to a certain sire as $u = \bar{u} + u_i$ where $\bar{u}$ is average return across all farms and u_i is a farm specific return drawn from a distribution with mean zero. We would then rewrite the Henderson equation for one progeny i as:

$$\begin{aligned} y_i &= X_i\beta + Z_i(\bar{u} + u_i) + e_i \\ y_i &= X_i\beta + Z_i\bar{u} + (Z_i u_i + e_i) \end{aligned}$$

There is a bias in our estimate of $\bar{u}$ when $Cov(Z_i, u_i) \neq 0$, and often u_i is not observed. The above equation is the same formulation as the Correlated Random Coefficient (CRC) model as described by Wooldridge (2003) and Heckman and Vytlacil (1998) for choice of sires; using instrumental variables for both u_i and Z_i, we would recover both the average return to a sire and the distribution across herds under the typical conditions on the instruments (relevance and independence). Unlike the mixed model, getting unbiased estimates in CRC does not rely on knowing the relationships between all sires or progeny in the data, only the incidence of each sire. Once we estimate the distribution of sire effects, unsupervised machine learning techniques such as K-means clustering can be used to partition u_i into management groups dictated by the data to explore heterogeneous sire effects (Bonhomme and Manresa, 2015).

When we treat each sire as its own technology, a straightforward instrument choice is the posted price of each sire; the price of a sire affects choice of that sire but is plausibly exogenous to the performance of indvidial progeny provdided that the price used is the one at the time of selection. Unfortunately, instrumenting for each sire also assumes that every sire is available to every farmer. If we describe bulls not as discrete as in the Z matrix but in terms of PTAs, we have more options for instruments and avoid defining the exact choice set. Using net merit (NM), a linear combination of PTAs with economic weights, as the characteristic describing sires leads to a straightforward instrument choice, which is the economic weights used to calculate NM at the time the genetics were purchased; since such weights are updated every couple of years, these weights should determine the net merit of bulls at the time they were chosen but be plausibly exogenous from the progenys performance nearly three years later. Instruments for the farm specific returns to net merit could be, similar to Suri (2011), a linear projection of the history of sire choices.

Preliminary Results

Here we present preliminary results of an empirical exercise to illustrate the heterogeneity to returns in genetics using PTA's of sires at the time they were selected. By describing sires using PTA's of traits, we treat PTA as an input into the production function. Specifically, we estimate farm specific returns to lbs of PTA for fat and protein.

Dairy Herd Improvement Associations (DHIA) collect monthly, cow level observations on milk yield, somatic cell count, fat, protein, and other breeding and replacement decisions. Such field data is routinely used for genetic evaluations in dairy science research or for calculating PTA's in the industry, as a system of cow and sire IDs helps connect relatives, especially sires to their progeny. We focus on lactation level records of fat and protein yield for cows in Wisconsin between June 2011 and January 2015 with days in milk between 270 and 350 on farms with at least 50 cows, which encompassed around 440,000 cows, 10,000 sires, and 1,300 herds for a total of about 1.2 million records.[1] For the sires in the sample, we recover their PTA numbers from around the time they were chosen (when their progeny was conceived) from historical valuation data from the Council on Dairy Cattle Breeding. Heterogeneity in returns to chosen sire PTA here is captured by interacting PTA with a herd dummy variable, creating around 1,300 different coefficients. The regression equation is thus:

$$y_{it} = \alpha + \beta X_{it} + \sum_{j=1}^{H} \alpha_j \mathbb{1}\{h_{it} = j\} + \gamma_j \mathbb{1}\{h_{it} = j\} \times PTA_i + e_{it}$$

Where y_{it} is the trait, X_{it} is the vector of controls, and PTA_{it} of the sire for the trait. The herd specfici returns to PTA_{it}, γ_j, are calculated for each herd j in H. Controls included calving month, parity number, parity number squared, days in milk, breed, age at calving, and herd level intercept α_j. If returns to PTA are the same across herds, we should expect all γ_j to be roughly the same. If not, some herds have different returns to increasing the amount of PTA for that trait in the sires they select. The standard errors were clustered at

[1] Around half of Wisconsin dairy herds use DHIA services, so the sample only represents herds in Wisconsin that use DHI services.

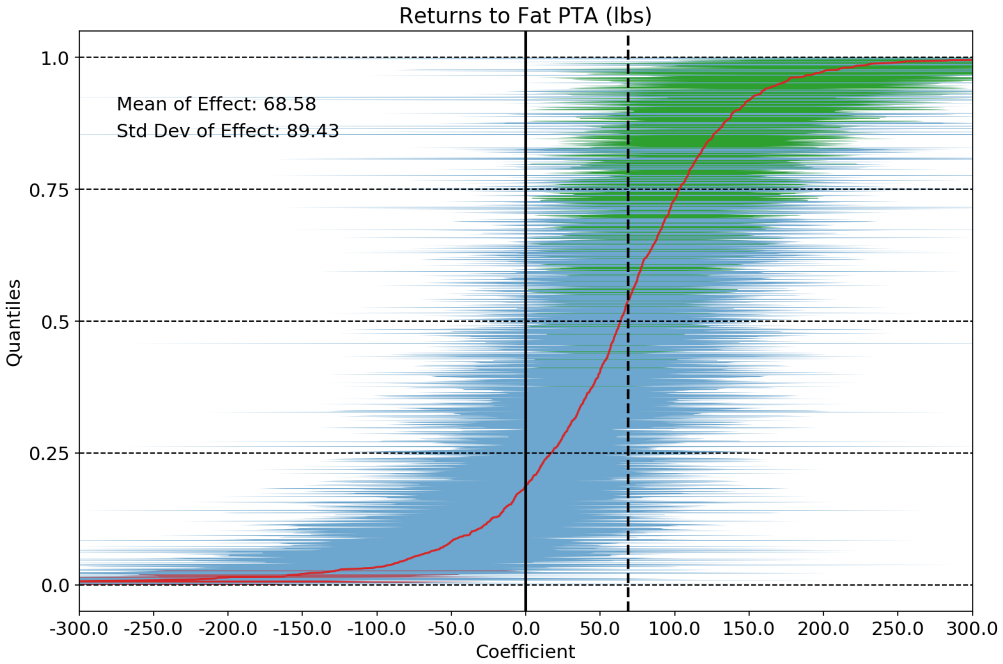
the progeny level and PTA numbers are in units of one-hundred.[2]

Figures 1 shows graphs of the increases in revenue from an increase in 100 PTA. The red lines are the point estimates, the blue lines the upper and lower bounds, and the black dotted line the mean effect, recovered by estimating the effect of PTAs on y without interactions. Areas colored in green indicate significantly different than zero coefficients that are positive and red that are negative.

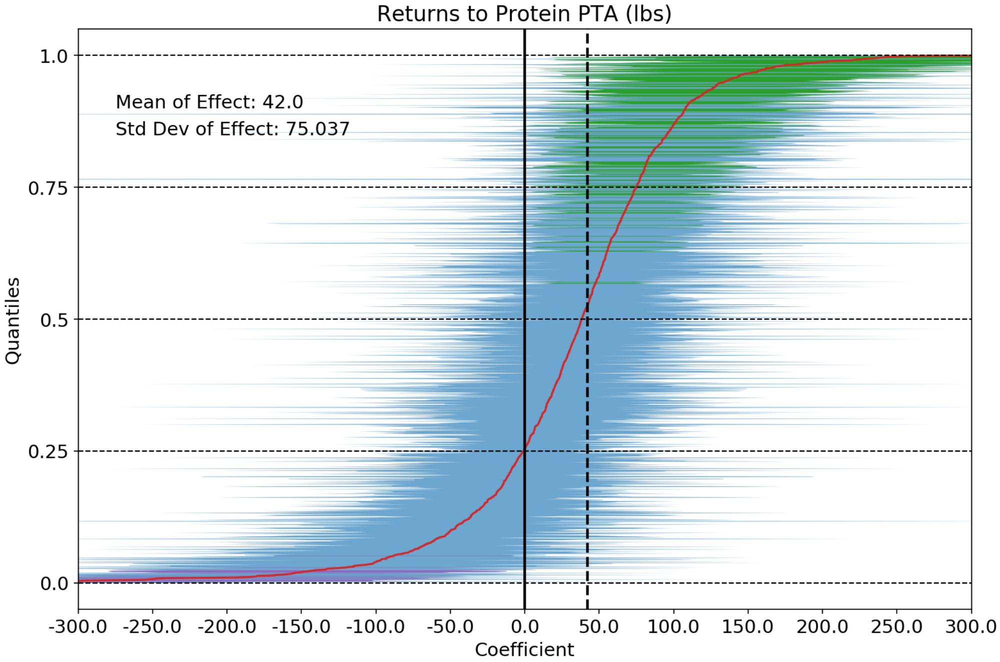
The mean increases in fat and protein were 68 and 42 lbs respectively, meaning purchasing 100 lbs of PTA gives, on average, 68% and 42% return per lactation. Coefficients are noisy since there is usually considerable variance in how a given sire trait will express in progeny. There is a large amount of variation in the return across herds, as the top quartile of herds can get on average 150 lbs of return on fat and 113 lbs of return on protein, more than twice the average. Conversely, the bottom quartile got on average negative returns, though most all the coefficients in that quartile were not statistically different than zero.

While great progress has been made with dairy cow productivity, there is clearly variability in returns for certain genetics across farms that has not been well explored; as such, while there has been an overall upward trend in dairy cow productivity, it is not clear how genetic selection has generated productivity across environments or management practices, making it difficult to understand the interaction between genetics and management. We hope to address this gap in understanding by characterizing heterogeneous returns to genetics across useful dimensions of farm heterogeneity using economic modeling via the CRC and unsupervised machine learning to uncover useful patterns in farm heterogeneity.

Figures 1-2: Returns to PTA (units of 100) in lbs



[2] PTA numbers themselves are in pounds of fat or protein.



References

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*(3), 1147–1184.

Hayes, B. J., M. Carrick, P. Bowman, and M. E. Goddard (2003, November). Genotype-Environment Interaction for Milk Production of Daughters of Australian Dairy Sires from Test-Day Records. *Journal of Dairy Science 86*(11), 3736–3744.

Heckman, J. and E. Vytlacil (1998). Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources*, 974–987.

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics 31*(2), 423–447.

Kearney, J. F., M. M. Schutz, P. J. Boettcher, and K. A. Weigel (2004, February). Genotype Environment Interaction for Grazing Versus Confinement. I. Production Traits*. *Journal of Dairy Science 87*(2), 501–509.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, 69–85.

Mundlak, Y., R. Butzer, and D. F. Larson (2012). Heterogeneous technology and panel data: The case of the agricultural production function. *Journal of Development Economics 99*(1), 139–149.

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science 6*(1), 15–32.

Suri, T. (2011). Selection and comparative advantage in technology adoption. *Econometrica 79*(1), 159–209.

Thornton, P. K. (2010). Livestock production: recent trends, future prospects. *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1554), 2853–2867.

Wang, S. L., P. Heisey, D. Schimmelpfennig, and V. E. Ball (2015). Agricultural productivity growth in the United States: Measurement, trends, and drivers.

Wooldridge, J. M. (2003, May). Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. *Economics Letters 79*(2), 185–191.

Conflict of Interest Statement

The authors declare that there is no conflict of interest.