Comments on "The Effect of College Curriculum on Earnings: Accounting for Non-Ignorable Non-Response Bias" by Daniel Hamermesh and Stephen Donald

Manuel Arellano

Quinta do Lago, May 21, 2005

1. Summary

- This paper looks at the relationship between college major and earnings using a new dataset, which contains, among other virtues, a richer set of ability-related controls than other papers in the literature.
- Earlier studies have found large effects of college major on earnings (e.g. an economics major would earn 48 percent more than a philosophy & theology major, but 21 percent less than a chemical engineering major, according to Black, Sanders, and Taylor, 2003).
- The question this paper helps to answer is whether the observed differentials in returns to college by major are actual monetary premia or spurious reflections of differing abilities of individuals choosing different majors.
- The authors collect earnings survey data from the population of graduates of their own university. By matching this information with University administrative records, they get detailed measures on college careers, and pre-college background and ability.

- Regrettably the response rate was only 25 percent. Thus, although the administrative data are available for all graduates, the earnings data are only available for one in four individuals in the target sample.
- The problem is that respondents may be an endogenously self-selected group. Another selection problem is that earnings are only observed for employed respondents.
- To address these problems, the authors consider an econometric model with two selection rules and exclusion restrictions in the earnings equation. Both parametric and semiparametric versions are considered.
- The determinants of employment excluded from the wage equation are indicators of the presence of young children.
- The determinant of response excluded from the wage equation is an indicator of membership of the University's alumni association.
- The idea of using such a kind of affinity measures for adjusting for non-response bias is another nice contribution of the paper that may have wider applicability.
- The main conclusion is that about half of the earnings differences across majors are accounted for by differences in endowments and post-college activities. The qualitative effect of accounting for non-response on the results turns out to be small.

2. Comments on data and variable construction

- The first issue raised by this kind of data is the extent of its specificity.
  - (a) How specific is the Texas-Austin population of graduates? Some majors in Austin may enjoy a particularly high (low) consideration, leading to abnormally high (low) returns.
  - (b) The distribution of SAT scores for Austin graduates may be more compressed than the distribution from a mix of colleges (given the prevalence of college sorting in the US). This might explain the small effect of SAT results on earnings.
  - (c) Related to this: why not separate math scores from verbal scores?
  - (d) Could graduates from the years 1994-95 and 1999-2000 be affected by the "top 10 percent" rule in Texas (top 10 percent students of their high school classes are guaranteed admission to leading campuses).
- Other aspects that merit discussion are the classification by major and major switches.
  - (a) Why not separate natural sciences between life sciences and math-physics?
  - (b) Why not separate economics (and psychology perhaps) from the other social sciences (Turner and Bowen found them to be very different).
  - (c) Do the authors know who switched majors? I understand transferring majors is not uncommon. Is there are switched major effect?

3. Does the exclusion of a dummy variable help identification?

- The paper emphasizes the role of exclusion restrictions for identification in a selection model, and suggests one restriction in the context of survey non-response.
- A nice aspect of the results is the marked contrast between estimates that rely exclusively on the nonlinearity and those that use exclusion restrictions.
- The point I discuss here is that with an excluded dummy variable (like membership of the alumni association), functional form assumptions play a more fundamental role in securing identification than in the case of an excluded continuous variable.
- The object of interest is

$$E\left(Y \mid X, Z\right) = g_0\left(X\right),$$

but we observe

$$\begin{aligned} E\left(Y \mid X, Z, D = 1\right) &= \mu\left(X, Z\right) = g_0\left(X\right) + \lambda_0\left[p\left(X, Z\right)\right] \\ E\left(D \mid X, Z\right) &= p\left(X, Z\right) \end{aligned}$$

- The question is whether $g_0\left(.\right)$ and $\lambda_0\left(.\right)$ are identified from knowledge of $\mu\left(X, Z\right)$ and $p\left(X, Z\right)$.

- Consider first the case of continuous X and Z. If there is another solution (g^*, λ^*) then

$$g_0(X) - g^*(X) + \lambda_0(p) - \lambda^*(p) = 0$$

Differentiating

$$\frac{\partial(\lambda_0 - \lambda^*)}{\partial p}\frac{\partial p}{\partial Z} = 0$$

$$\frac{\partial(g_0 - g^*)}{\partial X} + \frac{\partial(\lambda_0 - \lambda^*)}{\partial p}\frac{\partial p}{\partial X} = 0$$

Under the assumption that $\partial p/\partial Z \neq 0$, we have

$$\frac{\partial(\lambda_0 - \lambda^*)}{\partial p} = 0, \quad \frac{\partial(g_0 - g^*)}{\partial X} = 0$$

so that $\lambda_0 - \lambda^*$ and $g_0 - g^*$ are constant (i.e. g_0 is identified up to an unknown constant). This is the identification result of Das, Newey, and Vella (2003).

- Suppose X is continuous but Z is a dummy variable. In general $g_0(X)$ is not identified. To see this, consider

$$\mu(X, 1) = g_0(X) + \lambda_0[p(X, 1)]$$
$$\mu(X, 0) = g_0(X) + \lambda_0[p(X, 0)]$$

so that we can identify the difference

$$\Delta\mu(X) = \lambda_0[p(X, 1)] - \lambda_0[p(X, 0)]$$

but this does not suffice to determine λ_0 up to a constant.

- Take as an example a simple logit or probit model of the form
$$p(X, Z) = F(\beta X + \gamma Z),$$
so that, letting $h_0(.) = \lambda_0[F(.)]$,
$$\Delta\mu(X) = h_0(\beta X + \gamma) - h_0(\beta X).$$
Any other solution h^* should satisfy
$$h_0(\beta X + \gamma) - h_0(\beta X) = h^*(\beta X + \gamma) - h^*(\beta X),$$
which holds for a multiplicity of periodic functions.

- If X is also discrete, there is clearly lack of identification. For example, suppose X and Z are dummy variables:
$$\begin{aligned} \mu(1, j) &= g_0(1) + \lambda_0[p(1, j)] \\ \mu(0, j) &= g_0(0) + \lambda_0[p(0, j)] \quad (j = 0, 1) \end{aligned}$$
Since $\lambda_0(.)$ is unknown $g_0(1) - g_0(0)$ is not identified. Only $\lambda_0[p(1, 1)] - \lambda_0[p(1, 0)]$ and $\lambda_0[p(0, 1)] - \lambda_0[p(0, 0)]$ are identified.

- The conclusion is that excluding a binary variable does not guarantee identification by itself, although it may imply tighter bounds, and more credible identification in conjunction with functional form assumptions.

- Perhaps the idea of using affinity measures can be pursued to using donations from former graduates, or length of membership of the alumni association, which would be variables with a larger support.

4. Discussion of results and extensions

- (Observed) heterogeneity in the effects of major
  - (a) Within each major, a substantial fraction of college graduates pursue advanced degrees. Are there interactions with returns to majors? Others have found interactions of this sort with fewer controls.
  - (b) Black et al. found male-female differences in premia. Do these differences remain after controlling for the background measures available in the present dataset?
- Issues relating to the affinity measure and its effects
  - (a) Why not showing effects of affinity on the probability of response, to see how good the instrument is given controls?
  - (b) I missed an analysis of determinants of membership of the alumni association. Presumably membership varies with year of class.
  - (c) It is suspicious that the estimates imposing $\rho = 0$ are so similar to unrestricted estimates. Is there evidence of lack of identification? Estimated $\widehat{\rho}$ is not shown.
  - (d) An interesting conclusion is the small effect of lack of adjustment for non-response bias. Subject to the exclusion restriction, a test of

$$H_0: \quad E(Y \mid X, Z, D = 1) = E(Y \mid X, D = 1)$$

is a test of self-selection, whereas subject to ruling out self-selection it is a test of the validity of the exclusion restriction. Either way it seems a useful diagnostic.

Possible extensions

- Distributional effects of major choice (e.g. effects on the variance of earnings).
  (a) From Table 2 we see large differences in unconditional standard deviations of earnings by major, but how much of this is accounted by controls we do not know.
  (b) The issue is of interest because there could be a mean-variance trade-off in the earning consequences of different major choices.
- Endogeneity of major choice
  (a) No attempt is made of accounting for endogeneity or unobserved heterogeneity in major choice.
  (b) I share the authors' pessimism on the possibilities of achieving identification of this kind using data on graduates from a single university.
  (c) Such an exercise would require multi-college data to be able to exploit variation in proximity measures, tuition fees, etc.

5. Heterogeneity of treatments vs. heterogeneity of gains from treatment

- The results in this paper have implications for the interpretation of the heterogeneity in returns to college found in the literature. The nice aspect of the sample used is its high homogeneity and the wealth of background controls available. Yet the authors find significant differences in returns to different majors.
- Suppose for the sake of the argument that there are two majors with indicators (D_A, D_B) and potential outcomes (Y_0, Y_{1A}, Y_{1B}) corresponding to high-school, college with major A, and college with major B. Observed earnings are
$$Y = Y_0 + (Y_{1A} - Y_0) D_A + (Y_{1B} - Y_0) D_B$$
and the college–high school indicator is
$$D = D_A + D_B.$$
- Suppose we observe Y and D. Under exogeneity $(Y_0, Y_{1A}, Y_{1B}) \perp (D_A, D_B)$, the OLS impact is a linear combination of the two average returns:
$$\beta = E(Y \mid D = 1) - E(Y \mid D = 0) = \pi E(Y_{1A} - Y_0) + (1 - \pi) E(Y_{1B} - Y_0)$$
where $\pi = \Pr(D_A = 1 \mid D = 1)$. Since π is the result of choice, β does not measure any meaningful causal effect.
- Note that also
$$Y = Y_0 + (Y_{1B} - Y_0) D + (Y_{1A} - Y_{1B}) D_A,$$
so that under exogeneity, OLS of Y on D_A from data for graduates gives $E(Y_{1A} - Y_{1B})$.

- Now suppose non-exogeneity but that the IV assumption holds $(Y_0, Y_{1A}, Y_{1B}) \perp Z$, with Z binary.
- Let the potential indicators of major choice be
$$D_A = \begin{cases} D_{1A} & \text{if } Z = 1 \\ D_{0A} & \text{if } Z = 0 \end{cases} \quad D_B = \begin{cases} D_{1B} & \text{if } Z = 1 \\ D_{0B} & \text{if } Z = 0 \end{cases}$$
and assume absence of defiers in both groups.
- It turns out that the IV impact is a linear combination of average treatment effects for two different groups of compliers:
$$\begin{aligned} \frac{Cov(Z,Y)}{Cov(Z,D)} = {} & E(Y_{1A} - Y_0 \mid D_{1A} - D_{0A} = 1)(1-\lambda) \\ & + E(Y_{1B} - Y_0 \mid D_{1B} - D_{0B} = 1)\lambda \end{aligned}$$
where λ is given by the odds of compliers in one college major relative to the other. Specifically, $\lambda = 1/(1+\varphi)$ and $\varphi = \Pr(D_{1A} - D_{0A} = 1) / \Pr(D_{1B} - D_{0B} = 1)$.
- If $Y_{1A} - Y_0$ and $Y_{1B} - Y_0$ are constant, the IV impact is just a weighted combination of the two returns.
- In neither case the IV parameter provides a meaningful causal effect.
- These expressions can be easily generalized to more than two majors.