NBER WORKING PAPER SERIES

THE EFFECT OF COLLEGE CURRICULUM ON EARNINGS: ACCOUNTING FOR NON-IGNORABLE NON-RESPONSE BIAS

Daniel S. Hamermesh
Stephen G. Donald

Working Paper 10809
http://www.nber.org/papers/w10809

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2004

We thank the Andrew W. Mellon Foundation for financial support of this project, and Dean Richard Lariviere for having suggested conducting the survey. Merrick Brown and Mark Pocock offered excellent research assistance, while Davis Phillips provided invaluable aid in identifying and choosing the sample. Most important, Jamie M. Doyle was instrumental in improving, fielding and analyzing the initial results of the questionnaire. Joseph Altonji, Ronald Ehrenberg, Scott Gehlbach, Gary Solon and Jeff Underwood gave very helpful suggestions, and useful comments were received from participants in seminars at several universities and the NBER. The views expressed herein are those of the author(s) and not necessarily those of the National Bureau of Economic Research.

The Effect of College Curriculum on Earnings: Accounting for Non-Ignorable Non-Response Bias
Daniel S. Hamermesh and Stephen G. Donald
NBER Working Paper No. 10809
September 2004
JEL No. J310, I210

ABSTRACT

We link information on the current earnings of college graduates from many cohorts to their high-school records, their detailed college records and their demographics to infer the impact of college major on earnings. We develop an estimator to handle the potential for non-response bias and identify non-response using an affinity measure – the potential respondent's link to the organization conducting the survey. This technique is generally applicable for adjusting for unit non-response. In the model describing earnings, estimated using the identified (for non-response bias) selectivity adjustments, adjusted earnings differentials across college majors are less than half as large as unadjusted differentials and ten percent smaller than those that do not account for selective non-response.

Daniel S. Hamermesh
Department of Economics
University of Texas
Austin, TX 78712-1173
and NBER
hamermes@eco.utexas.edu

Stephen G. Donald
Department of Economics
University of Texas
Austin, TX 78712-1173
donald@eco.utexas.edu

1 Introduction

The impact of undergraduate curriculum on earnings has attracted substantial attention from economists, sociologists and educational researchers over the past two decades. (Wise, 1975; Grogger and Eide, 1995; Black et al, 2003, are just a few of the studies that have considered the issue of college major, and even high school curriculum has garnered some attention—Altonji, 1995; Rose and Betts, 2002). Partly the focus has stemmed from the large gender disparity in choices of major and the disproportionate fraction of women who choose majors with low earnings potential for both sexes (cf. Gerhart, 1990; Datcher Loury, 1997; Turner and Bowen, 1999; Joy, 2003). Despite the extensive examination of the relationship between college curriculum and earnings, however, there is a substantial need for additional research on this topic due to a variety of problems with existing studies.

Among the difficulties in the literature are: 1) Most of the studies examine the effect of major on earnings only a few years after students complete college, thus failing to measure the lifetime-income effects of different majors. This matters to the extent that the slopes of age-earnings profiles differ across majors; 2) Much of the research in this area does not account for background/ability measures that may be correlated with choice of major; 3) Most of the literature lacks information on the actual courses taken by the college student, focusing solely on his/her major field. 4) Some of the literature considers college major without considering performance in college, thus perhaps confusing the impact of major with the possibly correlated performance in college. 5) Definitions of college majors are not standard across universities, so that comparisons may generate errors insofar as the representation in the sample of majors by alma mater is non-random.[1]

All of these are problems with the underlying data—none of the data sets used (typically large national surveys) deals with all these difficulties. The data set we have created here circumvents all these problems. It describes a well-defined set of college majors from one single (albeit very large) university; there is more background information

[1]Perhaps the most useful survey along these criteria is that underlying Bowen and Bok (1998), which deals with all of these difficulties except the third.

on the students than has hitherto been available in studies of this issue; there is detailed information on the courses the students took in college, not just on the broad indicator of their major or their average performance; and the graduates in the sample range from the very recent to those who left the institution over 20 years before the survey was fielded.

The entire literature has an additional deficiency, one that is recognized as pervasive in survey-based research generally: The respondents on whom the studies are based are unlikely to be random individuals from among the college graduates who form their sampling frames. Even with a carefully chosen sampling frame, those people who respond and provide current information on their earnings are likely to be a selected sample of graduates. This issue, of non-ignorable non-response bias, is a specific problem in sample selectivity that seems to be of general importance but that has received relatively little attention in the literature on survey methods (Little and Rubin, 2002). In this study we provide an approach to identifying the selectivity that generates the non-response bias. While the particular method is specific to our particular problem, the basic idea provides guidelines for handling this difficulty in survey work generally.

2 The Underlying Data

We designed a questionnaire to elicit information on graduates' current earnings and occupational attachment, hours of work, postgraduate education (if any), how they financed their college education, and current family status (marital status, and age and number of children). (A copy of the questionnaire is available on request.) The questionnaire was purposely kept short to ease the task of respondents and to allow it to be readily machine-readable.

The sampling frame consists of randomly chosen undergraduates taken from the population of graduates from the years 1979-80, 1984-85, 1989-90, 1994-95 and 1999-2000 of the largest single-campus university in the United States, the University of Texas at Austin. This choice of graduating classes enables us to study a cross section of the current earnings of people covering a range of ages concentrated between 23 and 43, and to

use the other information so that we have background data on their current demographic status and some characteristics of their current employment. The goal was to obtain 1600 members of each class as potential respondents to the survey.[2] This amounted to a sample of slightly over 25 percent of the graduates in those classes.

From the University's administrative records system we obtained information on the entire college careers of everyone in the sampling frame, including the identity of and grade in each course taken at the university, the total number of credits received and grade-point average attained, and college major. Moreover, we had information on their pre-college backgrounds, including their rank in their high school graduating class and their SAT score (or the SAT-equivalent of their ACT score) from their applications to the University. In addition, since we knew the area where they lived while in high school, we were able to construct environmental background measures for each student, including the median family income (in 1990 dollars) in the Census tract (or untracted town) where the student graduated from high school.

The first choice to be made is how to classify graduates by major. This choice pervades all studies in this literature. The goal is presumably to ensure reasonable cell sizes while guaranteeing some degree of homogeneity of the kinds of majors included in each cell. There are over 80 majors in the University, ensuring that the number of respondents in some of them would be minute if we treated every major separately. Moreover, some majors that existed in 1979-80 disappeared, while new ones were created during the two decades. To circumvent both problems we generally define major in terms of the college in which the student's major was situated. A list of the eleven majors is shown in the left column of Table 1, while the second and third columns show the numbers of graduates in the sampling frame and the number of survey respondents.

Using this sampling frame, we mailed the survey to 7970 graduates from the five classes under study.[3] Regrettably, although the University had addresses on all of these

[2]We ensured randomness by using the random number generator in SAS to produce samples of 1600 individuals in each graduating class.

[3]For a variety of reasons a few people slipped out of the sampling frame, leading to the 7970 actually surveyed rather than the desired 8000.

graduates, apparently not all were valid. Thus during the seven months from the date the survey was mailed out (November 20, 2001) until the date the books were closed on the survey (June 30, 2002), 424 envelopes were returned as undeliverable. No doubt too some of the other non-respondents never received the survey and thus never had the choice to select into the sample of respondents. We thus treat all 5,955 non-respondents identically, classifying them as one group and the 2,015 respondents as the other group.

3 Initial Analyses

The means of the responses to the central questions from the survey, on the high-school background variables, and on the variables describing experience at the University, are presented in Table 2 separately by college major and then for the aggregate of respondents. The data on earnings are conditional on a respondent having reported positive earnings (as some respondents were not working at the time of the survey, while a minute number indicated that they were working but did not respond on the question about earnings).

Clearly, there are substantial differences across major in average earnings, with the highest-earning majors (Honors Plan II, and "hard" business) having an average almost three times that of the lowest (Education). Much of the differences across majors must be due to differences in what the students bring to and do at the University. Students in the higher-earning majors generally have higher SAT totals upon entry, and the fractions of students taking upper-division math and science courses and doing well in them are greater too. The differences are also consistent with the results of differential effort in the labor market and male-female differences in earnings. Thus respondents in the higher-earning majors tend to state that they work longer hours than those in lower-earning majors; and except for the Honors Plan II major, the fraction of women in the higher-earning majors is lower. On the other hand, advanced degrees are more prevalent among those graduates who have majored in subjects that eventually generate lower earnings. Similarly, family incomes in the areas where the students attended high school do not

differ across majors; and absolutely unsurprisingly there are no differences across majors in average GPA (again, except for the Honors Plan II major).

We have collected the data on this sample, including the unusually complete background and pre-college information, and the very long follow-up of some the graduates. The standard approach would be to estimate earnings equations controlling for all the high-school background measures, college achievement, additional demographic information, and college major for those respondents who were working at the time of the survey. The results of various versions of this regression model describing the logarithm of earnings are presented in the first three columns of Table 3, in which the excluded major is Education.[4] Even adjusting for all the other variables on which we have information, there are important differences in adjusted earnings across major fields, with most of them being in the directions one would expect. Engineering and the "hard" business disciplines of accounting and finance pay more. These pay differentials are, however, nowhere nearly as large as one might have expected, because many of the important control variables are correlated with both major and earnings (as was suggested by the means in Table 2). A difference of 21 log points between earnings in the "hard" business majors and social science majors is large, but not huge; and the difference between the "soft" business majors and others is even less.[5]

Other than our findings about the adjusted differences in earnings by major, perhaps the most interesting results are those on the variables that we believe are unique to the data set that we have assembled. Even within major, taking more upper-division science or math courses and doing better in them raise eventual earnings. While the effects are

[4]For respondents on whom we did not have information on the additional variables included in the estimates in the second and third columns we added an indicator of that variable being missing. Re-estimating the equations eliminating the observations with item non-responses had only minute effects on these coefficient estimates and the others.

[5]If we separate economics majors from others in the social sciences, we find that their adjusted earnings are halfway between those in the "hard" and "soft" business curricula. Re-estimating the equations using only the classes of 1980 and 1985 to remove people who might still be investing substantially in their skills hardly changes the estimated coefficients on the indicators for college major. Similarly, using the entire sample, allowing for major-specific time trends in adjusted earnings does not add to the equations' explanatory power and has little effect on inferences about adjusted earnings differences across majors.

not highly significant statistically, the t-statistics generally exceed 1.28. A student who takes 15 credits of upper-division science and math courses and obtains a B average in them will earn about 10 percent more than an otherwise identical student in the same major (based on the estimates in Column (3)) who takes no upper-division classes in these areas. There is clearly a return to taking these difficult courses, even after we have adjusted for differences in mathematical ability by using the total SAT score (with only tiny differences in this and the other results if the total SAT score is broken down in math and verbal scores).[6]

Also intriguing is the role of the student's background, which we have proxied by the average income in the area where he/she attended high school. Those students who performed equally well on the SAT and in college, but who came from an area where the average income was one standard deviation above the mean, earned about 7 percent more than those who came from an area where it was one standard deviation below the mean. Whether these long-lasting effects stem from differences in home background, in the quality of schooling in high school, or in access to information networks cannot, of course, be inferred from these results. Nonetheless, that we find long-term impacts of this size in this fairly homogeneous sample on which we have so much other information suggests how powerful differences in students' neighborhood and family backgrounds can be.

Not surprisingly, earnings are higher among graduates in the earlier classes (students further along in their careers), with the pattern of coefficients exhibiting perfectly the usual inverse J-shaped age-earnings profile. Indeed, the annual returns in the first five post-college years are 6 percent, tapering off to 3.5 percent in the next quinquennium, and 1.5 percent in each of the next two quinquennia.[7] The adjusted female-male wage gap is small for single women compared to single men (8 percent). Comparing married

[6]The importance of access to this information should not be underestimated. Estimated earnings differences across majors are substantially higher (e.g., the premium for "hard" business rises to 64 log points, that for engineering to 50 log points) when the information on science and math courses is excluded from the equation in Column (1).

[7]We explored this further by adding a quadratic in age in these equations. The coefficient estimates were unsurprisingly small and statistically insignificant.

women to married men the gender differential is a 25-percent wage disadvantage, even in this sample of graduates of the same institution. Married men earn substantially more than single men, while married women suffer a wage penalty compared to their single counterparts (in data where we cannot observe a woman's labor-force history).

While the means by major implied that graduate degrees had little effect on earnings, the regressions on the micro data show that a doctorate or doctoral-equivalent degree raises earnings by between 13 and 18 percent (depending on the specification), but a Masters degree has no effect on earnings. The implied return to a year spent obtaining a Masters degree is slightly negative, while that to obtaining a doctorate (or equivalent) is not large, perhaps only 4 to 6 percent. Remembering that this category includes Ph.D. degrees and professional (J.D. and M.D.) degrees, the differences between these results and others' findings (going back as far as Ashenfelter and Mooney, 1968) about the returns to obtaining a doctorate are slight.[8]

The remaining control variables generally produce the expected results. Each additional hour of market work raises earnings by almost 3 percent (and, while a quadratic term is statistically significant, its inclusion does not alter the other results). Neither of the high-school achievement variables is highly significant, nor is college GPA. The former may stem from the relative homogeneity of the sample, but the latter is more surprising. While we saw that average GPA does not differ across majors, it is surprising that within a major going from a B to an A average raises annual earnings by only 7 percent.

One of the final variables measured whether having paid part of the cost of college through student loans provides otherwise identical graduates an incentive (conditional on their major and their hours of work), to search for higher-paying jobs or put forth more effort per hour on the job. In this data set it does not; nor does the effect become significant when we fail to condition on hours of work. In this sample self-employed respondents earn substantially more than otherwise identical graduates, but this result stems mainly

[8]Excluding the indicators for advanced degree from our regressions has only minute effects on the estimated impacts of the indicators of college major on earnings.

from their presence in certain high-paying occupations, particularly medicine and law.[9]

There are two potentially serious difficulties with these estimates, neither of which appears to have been addressed in the literature on the impact of college major on earnings. The former, and probably the less important, is the possible correlation of potential earnings with the choice of whether to participate in the labor market. This is the standard self-selection problem that has been addressed repeatedly, initially by Gronau (1974) and Heckman (1976). Although we do pay attention to it, it is unlikely to be very important in these data, since the overwhelming majority of the graduates in this sample are employed.

The potentially more important difficulty with these estimates is that they are based on a sample (respondents) that is self-selected, perhaps non-randomly. To examine this potential problem, consider the probit results shown in Column (4) of Table 3. The dependent variable indicates whether the person responded to the survey. The probit only describes a small portion of the variation in the probability of response, but it is significant overall. More important, there are differences across major in the probability of response: The vector of ten indicators of college major is statistically significant ($\chi^2(10) = 30.37$, with the 1-percent level of significance being 23.21). We believe a similar problem pervades estimates throughout the literature on the impact of college major on earnings. Hence the rest of this study is devoted to devising ways to account for the problem of non-response bias, which appears to be present in the usual least-squares estimates, and to determine how large the impact of this apparently non-ignorable non-response bias is in our particular example. In doing so we lay out an approach to identifying the effects of non-response that may be useful in a wide variety of other problems in the social sciences.

[9]When a large vector of indicators of detailed occupations is included in these regressions, the effect of being self-employed drops substantially, and the vector of coefficients on the occupation indicators is significant as a whole. Including it does not, however, greatly change the relative sizes of the estimated impacts of college major on earnings.

4 A General Approach to Accounting for Non-ignorable Non-response Bias

The previously unaddressed problem that we face is how to account for the possible (and in our sample demonstrated) nonrandomness in the probability of response to our survey. We can capture the problem by the following three-equation model:

$$y_1 = 1(x_1'\delta_1 + \varepsilon_1 > 0) \text{ – responded} \quad (1)$$

$$y_2 = 1(x_2'\delta_2 + \varepsilon_2 > 0) \text{ – employed if responded} \quad (2)$$

$$\log y_3 = x_3'\delta_3 + \varepsilon_3 \text{ if } y_2 = 1 \text{ – earnings} \quad (3)$$

where (1(.)) denotes an indicator equaling one if the event in brackets is true.

Problems of double selectivity have been addressed often in the empirical literature (e.g., Ham, 1982; Tunali, 1986), but little attention has been paid to the crucial question of identification. In most cases either the estimator merely accounts for the cross-correlation in the errors of the pair of equations or identifies the model through nonlinearity in the specification of the selection equations. The crucial problem with this double-selectivity model, as in the entire literature on selectivity, is identification.[10] Selectivity into the labor market among survey respondents is the less serious problem—we can use the standard identifier of presence of young children, which arguably determines labor-force participation but does not affect earnings conditional on participation. Thus the number of children and the age of the youngest child are included in x_2 but not in x_3 in the log-earnings equation (3). The more serious issue is that of finding a variable or set of variables that might identify whether a randomly chosen graduate has decided to respond to the survey—some measure that is included in x_1 but not in x_2 or x_3. The identifier obviously must, moreover, be something that is obtainable for all graduates in the sampling frame, not merely for those who chose to respond.

All of the variables included in the probits in Column (4) of Table 3 are included

[10] In addition to the most well-studied problem, that of selectivity into the labor force, selectivity out of longitudinal samples is one that has been analyzed in a similar framework and with equal or greater difficulties of finding appropriate identifying variables (Falaris and Peters, 1998).

in the earnings equation (3), so none of them solves the problem. We believe that in any randomly chosen sampling frame some of the individuals will be more likely to respond because they have developed some ties to the organization that is requesting their cooperation. We call an indicator of those ties an affinity measure—an indicator of the potential respondent's affinity for the people or group conducting the survey and thus his/her potential interest in responding.[11] This might be a commonality of religion, evidence of participation in some group in common with the organization doing the survey, or something else. In our case the Alumni Office of the University provided us with information on whether or not an individual in the sampling frame is currently a member of the University's alumni association. The variable, TexasEx, is thus an indicator variable that we believe will be correlated with the individual's likelihood of completing the survey.[12]

One might argue that membership in the alumni association affects wages—perhaps membership generates contacts that help the members in their careers. We believe that this argument is weaker than an objection to identification of female labor-force participation through the presence of young children. In that case one might reasonably claim that employers, knowing that a woman has young children, will pay her lower wages because of additional expected absences and a perceived lack of interest in her career. In the end, whether this particular variable is a good identifier is ultimately a philosophical question. There is no doubt, however, that the affinity measure helps to distinguish between respondents and non-respondents: The membership rate among respondents is 25.0 percent, while among non-respondents it is only 13.0 percent.

We believe that this approach to identification is generally applicable—in most social surveys one has some inkling of which members of the sampling frame have a greater or lesser affinity for those people fielding the survey. Moreover, increased response rates should be obtainable by (subtle) indications in introductions to the survey that might

[11] We do not believe this general approach has been suggested before. Attempts have been made (e.g., Copas and Farwell, 1998) to use respondents' expressed willingness to answer questions to extrapolate to the characteristics of nonrespondents.

[12] The alumni association of this institution calls itself the TexasExes.

elicit the respondents' affinities for the researchers who are conducting it (university affiliation in our case).[13] In many cases at least some of the determinants of that affinity can be argued to be independent of the outcomes that the researcher is interested in studying. An affinity measure can thus be viewed as similar to Philipson's (1997) proposal to view survey response probabilities as something amenable to market forces, except that here we have defined the data market as an implicit market in attachment to the University (and to its researchers who conducted the survey).

5 Adjusting For Non-Response Bias

We estimate the model described in equations (1)-(3) under a variety of conditions. The key feature of the model is that we only see earnings if $y_1 = 1$ (responded) and $y_2 = 1$ (working if responded). In addition we only see y_2 if the individudal responded to the survey. We first present the correction term under the general condition that $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ are jointly i.i.d. and then proceed to the particular assumptions used in this paper. The general form of the correction term can be written in the usual way as,

$$E(\log y_3 | x, y_1 = 1, y_2 = 1) = x_3'\delta_3 + E(\varepsilon_3 | x, y_1 = 1, y_2 = 1)$$

$$E(\varepsilon_3 | x, \varepsilon_1 > -x_1'\delta_1, \varepsilon_2 > -x_2'\delta_2) = g(x_1'\delta_1, x_2'\delta_2)$$

where we are assuming that $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ are distributed indepenently of x. In particular,

$$g(x_1'\delta_1, x_2'\delta_2) = \frac{\int_{-\infty}^{\infty}\int_{-x_2'\delta_2}^{\infty}\int_{-x_1'\delta_1}^{\infty} \varepsilon_3 f(\varepsilon_1, \varepsilon_2, \varepsilon_3) d\varepsilon_1 d\varepsilon_2 d\varepsilon_3}{\int_{-\infty}^{\infty}\int_{-x_2'\delta_2}^{\infty}\int_{-x_1'\delta_1}^{\infty} f(\varepsilon_1, \varepsilon_2, \varepsilon_3) d\varepsilon_1 d\varepsilon_2 d\varepsilon_3}$$

Thus in the most general case we have the partially linear model,

$$\begin{aligned} \log y_3 &= x_3'\delta_3 + g(x_1'\delta_1, x_2'\delta_2) + \xi \\ E(\xi | x, y_1 &= 1, y_2 = 1) = 0 \\ V(\xi | x, y_1 &= 1, y_2 = 1) = \sigma_\xi^2(x_i) \end{aligned}$$

[13] While this is generally true, the affinity can occasionally backfire: One potential respondent wrote back a six-page diatribe against the University, accusing it of spoiling his life.

In this model the slope coefficients will be identified provided,

$$E(x_3|x_1'\delta_1, x_2'\delta_2, y_2 = 1) \neq x_3$$

The reasoning for this is exactly analogous to that given in Newey (1988) and Ahn and Powell (1993) where semi-parametric methods are employed to estimate the standard sample selection model. Although assuming a particular distribution for the residuals would seem to circumvent this requirement (in the case of normality at least), identification would be achieved through arbitrary choice of functional form, may not be entirely convincing and in any event may give rise to imprecise estimates if there is little nonlinearity in the correction term.

We can specialize things a bit by assuming,

$$E(\varepsilon_3|\varepsilon_1, \varepsilon_2) = \gamma_1\varepsilon_1 + \gamma_2\varepsilon_2$$

in which case we get,

$$\begin{aligned} E(\varepsilon_3|x, \varepsilon_1 &> -x_1'\delta_1, \varepsilon_2 > -x_2'\delta_2) \\ &= \sum_{j=1}^{2} \gamma_j E(\varepsilon_j|\varepsilon_1 > -x_1'\delta_1, \varepsilon_2 > -x_2'\delta_2) \end{aligned}$$

Assume joint normality of $(\varepsilon_1, \varepsilon_2)$, with variance-covariance matrix Σ normalized to be such that the variances are 1 and the covariance is equal to the correlation ρ_{12}. Letting the correlations between ε_3 and ε_j be ρ_{j3} (for $j = 1, 2$) and the standard deviation of ε_3 be σ_3, as shown in the Appendix (using $\phi(.)$ and $\Phi(.)$ as the standard normal pdf and cdf and $\phi(., .; \Sigma)$ as the joint normal pdf of $(\varepsilon_1, \varepsilon_2)$) we can write,

$$E(\varepsilon_3|x, \varepsilon_1 > -x_1'\delta_1, \varepsilon_2 > -x_2'\delta_2) = \sigma_3\rho_{13}\frac{\phi(c_1)(1 - \Phi(c_2^*))}{\int_{c_1}^{\infty}\int_{c_2}^{\infty}\phi(\varepsilon_1, \varepsilon_2; \Sigma)d\varepsilon_2 d\varepsilon_1} + \sigma_3\rho_{23}\frac{\phi(c_2)(1 - \Phi(c_1^*))}{\int_{c_1}^{\infty}\int_{c_2}^{\infty}\phi(\varepsilon_1, \varepsilon_2; \Sigma)d\varepsilon_2 d\varepsilon_1}$$

where we use the shorthand notation $c_j = -x_j'\delta_j$,

$$c_1^* = \frac{c_1 - \rho_{12}c_2}{\sqrt{1 - \rho_{12}^2}},\ c_2^* = \frac{c_2 - \rho_{12}c_1}{\sqrt{1 - \rho_{12}^2}}$$

In this paper we employ two- (or three-step) estimation procedures analogous to the original suggestion of Heckman (1979) in the context of the simpler sample selection

model. We first estimate δ_1, δ_2 and ρ_{12} using suitable models of response and the decision to work, plug these into the two correction terms and then do OLS estimation of the earnings equation for the subsample of responders who reported earnings. We use three different approaches that are described in the Appendix and describe them in decreasing generality. The most efficient and general procedure is based on estimating δ_1, δ_2 and ρ_{12} jointly using maximum likelihood under the assumption of joint normality of $(\varepsilon_1, \varepsilon_2)$. This involves taking into account the endogenous response in the estimation of the decision to work - a Bivariate Probit model with selection. The likelihood for this is provided in the Appendix. This is referred to as the "ρ free, two-step" approach. A second, less efficient approach allows for ρ_{12} to be free, but estimates δ_1, δ_2 and ρ_{12} (based on the likelihood for y_1 and y_2) using a three-step estimation method. First, δ_1 is estimated by a Probit model of response. Second, δ_2 and ρ_{12} are estimated by a Probit model for y_2 that takes into account the endogenous response and takes the initial estimates of δ_1 as given. Then all of the estimates are used to construct the correction terms for inclusion in the earnings equation, which is estimated in a third step by OLS. This approach is referred to as the "ρ free, three-step" approach. The final method based on these correction terms is obtained under the assumption that $\rho_{12} = 0$. In this case the correction terms simplify to

$$E(\varepsilon_3|x, \varepsilon_1 > -x_1'\delta_1, \varepsilon_2 > -x_2'\delta_2) = \sigma_3\rho_{13}\frac{\phi(c_1)}{\Phi(-c_1)} + \sigma_3\rho_{23}\frac{\phi(c_2)}{\Phi(-c_2)}$$

so that there are two of the usual sample selection corrections as in Heckman (1979) – one accounting for endogenous response and the other accounting for the endogenous decision to work. Additionally, the assumption that $\rho_{12} = 0$ implies that there is no sample selection problem in estimating a model of the decision to work using only responders. Hence we can estimate δ_1 using a Probit model of response and δ_2 using a Probit model on the decision to work in the sample of reponders. This approach is referred to as the "$\rho = 0$" method.

As is usual in selection correction models we face the problem of correcting standard errors for the use of estimates of parameters to estimate the correction terms. Under the assumption that the correction terms have zero coefficients, one can obtain appropriate

standard errors using heteroskedasticity consistent robust standard errors. We use these robust standard errors for the "$\rho = 0$" and "ρ free, three-step" methods. For the "ρ free, two step" which is the most general and efficient, we provide standard errors that correct for the pre-estimation error in the construction of the correction terms. The method for doing this is as described in Newey and McFadden (1994).

Although we have argued that our exclusion restrictions are sufficient to allow identification of the parameters in the earnings equation, one may call into question the normality assumption and the various linearities assumed in the above structure. It is infeasible to perform a fully non-parametric approach to estimation of the earnings equation (using, say, the non-parametric approach in Das, Newey and Vella (2003), hereafter DNV) due to the number of explanatory variables and the small sample size.[14] It is possible to gain some insight into identification as well as robustness by adapting the framework of DNV to our case and then considering a restricted version of their methods (as they indeed do in their empirical example). As noted in their paper, if we were to let $E(\log y_3|x) = \mu(x_3)$, then conditional on responding and reporting positive earnings one can write down the selection-corrected regression function as,

$$\begin{aligned} E(\log y_3|x, y_2 &= 1, y_1 = 1) = \mu(x_3) + \lambda(p_1(x_1), p_2(x_2)) \\ p_1(x_1) &= P(y_1 = 1|x) = P(y_1 = 1|x_1), \\ p_2(x_2) &= P(y_2 = 1|x) = P(y_2 = 1|x_2) \end{aligned}$$

The $p_j(x_j)$ are the propensity scores for response and for working. As shown in DNV, the function $\mu(x_3)$ is identified provided there are variables affecting the two propensity scores that do not enter into $\mu(x_3)$. This result is similar to the reasoning given above for the linear model with a general correction term. Then DNV provide methods that involve non-parametric estimation of the propensity scores and then estimation of the earnings equation using methods for additive regression estimation, with the propensity scores entering the selection correction.

[14] See Vella (1998) for a discussion of propensity scores in estimating models involving one selection equation.

In our case things are complicated by the fact that, although we can identify and estimate $p_1(x_1)$ through the response variable, the propensity score is not as easy to obtain because we do not observe the decision to work for a random sample. We potentially have sample selection bias in the estimation of the decision to work because this is only observed for responders. Thus we cannot generally estimate $p_2(x_2)$ using the decision to work variable for responders. In our case we are only able to estimate $P(y_2 = 1|y_1 = 1, x)$, since we only observe the decision to work for responders. Note that under mild regularity conditions,

$$P(y_1 = 1 \cap y_2 = 1|x) = C(p_1(x_1), p_2(x_2))$$

where C is some function (referred to as a copula). Noting that,

$$P(y_1 = 1 \cap y_2 = 1|x) = P(y_2 = 1|y_1 = 1, x)P(y_1 = 1|x)$$

then provided that C is sufficiently regular we can solve $r = C(p_1, p_2)$ for p_2 as a function of p_1 and r, as

$$p_2 = C^{-1}(p_1, r).$$

Using this we can then solve for $p_2(x_2)$ as follows,

$$\begin{aligned} p_2(x_2) &= C^{-1}(p_1(x_1), p_{2|1}(x_2, p_1(x_1))p_1(x_1)) \\ p_{2|1}(x_2, p_1(x_1)) &= P(y_2 = 1|y_1 = 1, x) \end{aligned}$$

Thus we can rewrite the correction term as,

$$\lambda^*(p_1(x_1), p_{2|1}(x_2, p_1(x_1)))$$

for some unknown function λ^*.

We employ a simplified three-step estimator based on the above heuristics. In the first step we estimate $p_1(x_1)$ using a linear probability model of y_1 on x_1. In the second step we estimate a partially linear probability model to estimate $p_{2|1}(x_2, p_1(x_1)))$, which is assumed to have the form,

$$p_{2|1}(x_2, p_1(x_1))) = x_2'\theta_2 + h(p_1(x_1))$$

This involves a regression of y_2 (for responders) on x_2 and polynomial functions of $\hat{p}_1(x_1)$ (estimated in the first step). This is a kind of response bias (or selection) corrected version of the decision to work. Finally, in the third step we also assume that the mean function for earnings (for responding workers) is of the partially linear form, with $\mu(x_3) = x_3'\theta_3$ and with the usable correction term $\lambda^*(p_1(x_1), p_{2|1}(x_2, p_1(x_1)))$ being approximated by two-dimensional polynomials in $\hat{p}_1(x_1)$ and $\hat{p}_{2|1}(x_2, \hat{p}_1(x_1)))$. This particular formulation of the model reinforces the importance of having exclusion restrictions on x_3 relative to x_2 and on x_2 relative to x_1, since without such restrictions there would be perfect multicollinearity in the second and third steps of the three step procedure. In what follows we consider only the cases where the polynomials are all linear or all quadratic in the various correction terms. Standard errors for these two estimation procedures (for the earnings equation) are obtained by bootstrapping the entire three-step procedure to take full account of the first two steps. A complete justification of this approach is beyond the scope of this paper but is possible by adapting the results of DNV.

6 Estimates of the Complete Model of the Determinants of Earnings

Before we discuss the various estimates of the complete model, equations (1)-(3), that we have proposed, we present estimates of a model of log-earnings in which the identification comes solely off the nonlinearity in the probits that generate the selectivity parameters λ_1 and λ_2. The variables included in the vectors x_1 and x_2 are thus identical to those in x_3, except that x_2 also includes the variables describing the presence and age of children. (The results are qualitatively the same if we exclude the children variables from x_2.) This approach provides no real identification of the sample non-response; but it has been employed in the general literature on selectivity (a problem first pointed out by Olsen, 1980). Here and in the rest of this section we present results only for the set of x_3 that is included in Column (1) of Table 3. As in that Table, the results here change little when we use the expanded set of regressors.

The first thing to note from the estimates based on attempting to identify the impacts on earnings off the nonlinearity in the selectivity models is the absolutely unsurprising result that the R^2 in this expanded equation is almost identical to the third significant digit to that of the equation in Column (1) of Table 3, which includes all the same variables but excludes these selectivity terms. What is different, however, is nearly the entire set of coefficients. Most of these are far larger than their counterparts in Table 3, and they imply absurdly large differences in the returns to different college majors. Given the unadjusted means in Table 2 and the pattern of variation across majors in the components of x_3, it is very difficult to believe, for example, that social science majors earn 5.96 (exp(1.546 + .239)) times as much as humanities majors, yet that is what these estimates imply. Equally disturbing, the standard errors of the parameter estimates have become much larger, to the point that few of these inflated parameter estimates are statistically significant. At least in this case, trying to identify selectivity off the nonlinearity in the probits in equations (1) and (2) leads to severe problems of multicollinearity in estimating equation (3). These estimates make no sense.

The second column of Table 4 shows the estimates of (3) based on identification that comes from the exclusion of the affinity indicator and the presence of children from (3), and their inclusion in the selectivity equations (1) and (2) respectively. For this first set of estimates we assume that the errors in (1) and (2) are independent—that the correlation of the residuals in them is $\rho = 0$, so that we need not account for cross-equation correlation in the selectivity equations. The selectivity term describing whether the respondent works is not significantly different from zero. This is not surprising—this is a sample that is majority male, and the college-graduate women who are included are more likely to be committed to working in the presence of young children than are American women generally.[15]

The selectivity correction for non-response bias has a significant effect on the respondent's log-earnings. Conditional on the observable variables, however, it appears that

[15] 68 percent of the female graduates with children under age 6 are employed, while 77 percent of those whose youngest child is between 6 and 18 are employed.

there is negative selectivity: Those who are predicted to be more likely to respond based on unobservable characteristics earn less than otherwise (observably) identical individuals. Most important, though, is that accounting for non-response does matter, and our method of identifying it is useful.

The substantive focus is on the differences in earnings across majors. Here the results vary somewhat from the least-squares estimates shown in Table 3; but although accounting for selectivity helps describe earnings, the changes in the implied differences in earnings across majors are not large. The earnings differentials are generally on the order of 10 to 20 percent smaller than was suggested by the least-squares estimates. Also, their levels of statistical significance (tested against the excluded category, education major) are somewhat lower than implied by the least-squares estimates. Nonetheless, differences between the (adjusted) highest-paying majors—"hard" business and engineering—and the (adjusted) lowest-paying majors—humanities and education—are highly significant statistically.

The coefficient estimates on the other variables change even less from their counterparts in Table 3, and there is no general pattern of increase or decrease. Otherwise identical single women earn 9 percent less than men; a two-standard error increase in background income (income in the graduate's high-school district) generates a 12 percent increase in earnings above that of an otherwise identical graduate; and there is a 28 (8) percent marriage premium (penalty) for men (women). As before, there is a roughly 18 percent premium to earning a doctorate (not a great rate of return, even on the three-year law degree that is included in this category), and having only a Masters degree has essentially no impact on earnings (other than the possible option value of allowing pursuit of a doctorate). Each additional weekly hour of work has a large effect on weekly earnings. Moreover, as in the least squares estimates, adding a quadratic in weekly hours to these equations produced the inverse-U shaped relationship that has been noted for broad samples of the labor force (Biddle and Zarkin, 1989), although it did not qualitatively affect the other estimates.

Columns (3) and (4) of Table 4 list the estimates of equation (3) under different

approaches to estimation when the assumption that $\rho = 0$ is relaxed.[16] In Column (3) we add correction terms computed from joint estimates of (1) and (2), including in each of x_1 and x_2 the *TexasEx* and children variables respectively, using the two-step procedure. In Column (4) the correction terms are estimated by the three-step procedure, first estimating δ_1 and then using those estimates to derive estimates of δ_2 and ρ_{12}. These are then included in the log-earnings regression.

The changes between the estimates of δ_3 when we relax the assumption that $\rho = 0$ differ so minutely from those produced under this assumption as not to merit specific comment. Typically there are differences only in the third significant decimal place between the parameter estimates in Columns (3) and (4) and those in Column (2). Moreover, the explanatory power of the equation increases by less than .001. One must infer that in this example the correlation between the errors in predicting non-response and in predicting whether a respondent works is unrelated to the workers' earnings (conditional on their other characteristics).

The estimates shown in Columns (5) and (6) of Table 4 are based on the second general approach to estimating δ_3 that we have derived—-the estimation of propensity scores describing the probabilities of non-response and employment. In Column (5) we specify the propensity scores as linear functions of the underlying variables and enter the estimated scores into the log-earnings regression. Comparing these estimates to those in Columns (2)–(4), we see that they are generally larger (although in most cases the earnings differences by major are still not quite so pronounced as they were in the least-squares estimates shown in Column (1) of Table 3). As with the extension of the standard selectivity correction, this alternative approach too does not greatly affect our inferences about the impacts of the variables of interest.

The final column in Table 4 shows the results of estimating the log-earnings equation when we use the estimated propensity scores for non-response and employment based on quadratics describing the scores (and thus also including an interaction term between

[16]The programs required to estimate the models with $\rho = 0$, $\rho \neq 0$, and those based on propensity scores are available upon request from the authors.

the estimated linear scores for non-response and employment). These estimates, which we view as the most reliable among those presented in the two Tables, change the results from Column (5) so that the adjusted earnings differences across majors are with two exceptions even less than those shown in Column (2).

Excluding the essentially unidentified estimates in Column (1) of Table 4, the differences that we have found between the most extreme sets of estimates matter but are not qualitatively very large. A fair conclusion is that, while it makes sense to undertake the adjustments for both non-response bias and labor-force selectivity, in this sample the non-randomness implied by behavior along both margins is nearly independent of the differences in earnings across college major. The techniques are generally applicable and useful in a wide variety of problems where one believes there is non-response bias in the data and where one can obtain an affinity measure to identify it; but while they do alter our results somewhat, the qualitative changes are not large.

7 How Large Are Earnings Differences by Major?

There is immense popular discussion and even joking about the poor economic prospects of students who choose to major in the liberal arts—particularly the humanities and the less rigorous social sciences.[17] The evidence in Table 2 on mean earnings by college major suggests these differences are indeed huge—with earnings in the highest-paying major being roughly three times those in the lowest-paying. The first column of Table 5 presents the means of the logarithms of earnings by major. The variation even in the means of log-earnings is immense, with the implied mean in the highest-paying major ("hard" business) being nearly 150 percent more than that in the lowest-paying major (education). The standard deviation of these logarithmic means is 0.305—a substantial amount of inequality, considering that the grade-point averages do not differ across major, that the students' average ages are the same, and that they all graduated from the same

[17] On the Tonight Show (September 18, 2003) Jay Leno noted that philosophy majors spend much of class time debating whether the glass is half-full or half-empty. This prepares them for their subsequent careers—as waiters.

institution.

Looking only at the raw means is very misleading. As Table 2 showed too, the qualifications that students bring to the university, including the proxy for their raw ability, SAT, their background and demographic characteristics, differ substantially across majors. So too do their post-college academic attainment and the effort they put forth on their jobs. The second and third columns of Table 5 repeat the coefficients on the indicator variables in the equations describing log-earnings (from Column (1) of Table 3 and Column (6) of Table 4). Almost exactly half of the pay premia in the highest-paying majors—"hard" business and engineering in particular—is accounted for by the differences in endowments and post-college activities. The role of differences in hours worked is particularly large. When these adjustments are taken into account, the degree of inter-major variation in earnings, as measured by the standard deviations of (adjusted) log-earnings, falls to 0.153. A remarkably large fraction of the perceived differences in earnings among workers who made different curriculum choices while in college are not inherent in those choices.[18] An additional 10 percent of the remaining variation in adjusted earnings across majors is accounted for when we adjust for the non-response that created a non-randomly selected sample.

All of the differences in earnings by college major are conditional on the individuals having chosen their major. We have not identified what earnings differences would be if there were random assignment to majors—an experiment that does not seem possible. We have ignored selectivity into major that is related to unobservable characteristics, partly in order to concentrate on identifying non-response bias, partly because we do not believe that we can find good identifiers for choice of major in our data.[19] Whether the estimated differences in relative earnings across majors are larger or smaller than what would be observed with random assignment is not clear—it depends on whether there is

[18]The inference would not be changed greatly if we also accounted for differences in the probability of employment by major: The only significant difference in this sample is between education majors (a lower employment probability) and all others.

[19]Arcidiacono (2004) uses the NELS data to identify the choice of major and gauge the effect of endogenous choice of major on earnings differences among majors (four broad categories) twelve years after graduation.

positive or negative selection into majors on the unobservables. So long as the selectivity is positive, as seems likely, even the much reduced estimated relative earnings differences across majors shown in the final column of Table 4 overestimate the differences that would be observed if students chose their majors randomly.

8 Conclusions and Future Uses

Using a survey that has enabled us to obtain more background information on students than acquired heretofore, we have examined the relationship between college major and labor-market success, as proxied by earnings. The estimates demonstrate that much of the differences among individuals that appear to make some majors so much more attractive than others are uninformative about the value added by particular majors. Rather, we have found that over half of the variance in the logarithm of earnings across majors is accounted for by sorting by ability, high-school performance, parents' economic status, students' demographic characteristics, and the amount of labor supplied to the market. The choices that students make about their college major do affect their earnings, but the impacts of the choices are not extreme.

Although this examination of the relationship between college major and earnings improves on certain aspects of the literature in this area, it has the drawback of being based on only one institution. This drawback allowed us to obtain more pre-college background information and college course information than have been available in any previous study, at the potential cost of having a sample that may not be representative of all U.S. college graduates. The next step for research in this area is to obtain representative samples of graduates from a large enough group of representative universities with the same detailed information that we have obtained to allow this kind of estimation to be conducted accounting for potential differences that are attributable to inter-university differences in choice of major.

In order to infer the impacts on earnings from these survey results we needed to account for possible non-response bias in our sample. To do so we had to develop a way of

identifying the selectivity that is inherent in any social-science sampling procedure. The identifier we use is the affinity of the potential respondent to the survey organization, in our case indicated by the potential respondent's membership in the University's alumni association. An affinity identifier can, we believe, be developed analogously in many surveys, with the results usable in inferring the importance of non-response bias. While it was not very important in our sample, even though the identifier was quite successful, it may be in other cases.

The affinity-identifier solution to the problem of non-ignorable non-response bias leads to the realization that interactions between potential respondents and the particular identity of those conducting a survey generate a larger problem in drawing inferences from that survey. Altering the identity of the organization or individual conducting a survey will alter the mix of respondents (and perhaps too the veracity, or at least the care, with which they respond). That in turn will generate statistics describing the survey results that depend on choices made about the identity of the surveyor. The role of affinity in survey research is a potentially extremely fruitful area for investigation.

REFERENCES

Hyungtaik Ahn and James Powell, "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," Journal of Econometrics, 58 (July 1993): 3-29.

Joseph Altonji, "The Effects of High School Curriculum on Education and Labor Market Outcomes," Journal of Human Resources, 30 (Summer 1995): 409-38.

Peter Arcidiacono, "Ability Sorting and the Returns to College Major," Journal of Econometrics, 121 (July 2004): 343-75.

Orley Ashenfelter and Joseph Mooney, "Graduate Education, Ability and Earnings," Review of Economics and Statistics, 50 (February 1968): 78-86.

Jeff Biddle and Gary Zarkin, "Choice Among Wage-Hours Packages: An Empirical Investigation of Male Labor Supply," Journal of Labor Economics, 7 (October 1989): 415-37.

Dan Black, Lowell Taylor and Seth Sanders, "The Economic Reward to Studying Economics," Economic Inquiry, 43 (July 2003) 365-77.

William Bowen and Derek Bok, The Shape of the River. Princeton, NJ: Princeton University Press, 1998.

Andrew Copas and Vern Farwell, "Dealing with Non-Ignorable Non-Response by Using an 'Enthusiasm-to-Respond' Variable," Journal of the Royal Statistical Society A, 161 (1998): 385-96.

Mitali Das, Whitney Newey and Francis Vella, "Nonparametric Estimation of Sample Selection Models," Review of Economic Studies, 70 (January 2003): 33-58.

Evangelos Falaris and Elizabeth Peters, "Survey Attrition and Schooling Choices," Journal of Human Resources, 33 (Spring 1998): 531-54.

Barry Gerhart, "Gender Differences in Current and Starting Salaries: The Role of Performance, College Major, and Job Title," Industrial and Labor Relations Review, 43 (April 1990): 418-33.

Jeff Grogger and Eric Eide, "Changes in College Skills and the Rise in the College Wage Premium," Journal of Human Resources, 30 (1995): 280-310.

Reuben Gronau, "Wage Comparisons: A Selectivity Bias," Journal of Political Economy, 82 (Nov.-Dec. 1974): 1119-43.

John Ham, "Estimation of a Labour Supply Model with Censoring Due to Unemployment and Underemployment," Review of Economic Studies, 49 (July 1982): 335-354.

James Heckman, "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for Such Models," Annals of Economic and Social Measurement, 5 (Fall 1976): 475-92.

————————-, "Sample Selection Bias as a Specification Error," Econometrica, 47 (January 1979): 153-61.

Lois Joy, "Salaries of Recent Male and Female College Graduates: Educational and Labor Market Effects," Industrial and Labor Relations Review, 56 (July 2003): 606-21.

Roderick Little and Donald Rubin, Statistical Analysis with Missing Data. New York: Wiley, 1987.

Linda Datcher Loury, "The Gender Earnings Gap among College-Educated Workers, Industrial and Labor Relations Review, 50 (July1997): 580-93.

Whitney Newey, "Two Step Series Estimation of Sample Selection Models," Unpublished Paper, MIT 1988.

———————- and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing," in Robert Engle and Daniel McFadden, eds., Handbook of Econometrics, Vol. 4. Amsterdam: North-Holland, 1994, 2111-2245.

Randall Olsen, "A Least Squares Correction for Selectivity Bias," Econometrica, 48 (1980): 1815-20.

Tomas Philipson, "Data Markets and the Production of Surveys," Review of Economic Studies, 64 (January 1997): 47-72.

Heather Rose and Julian Betts, "The Effect of High School Courses on Earnings," Unpublished paper, Public Policy Institute of California, 2002.

Insan Tunali, "A General Structure for Models of Double-Selection and an Application to a Joint Migration/Earnings Process with Remigration," Research in Labor Economics, 8 (1986): 235-82.

Sarah Turner and William Bowen, "Choice of Major: The Changing (Unchanging) Gender Gap," Industrial and Labor Relations Review, 52 (January 1999): 289-313.

Francis Vella, "Estimating Models with Sample Selection Bias: A Survey," Journal of Human Resources, 33 (Winter 1998): 127-69.

David Wise, "Academic Achievement and Job Performance," American Economic Review, 65 (June 1975): 350-66.

Jeffrey Wooldridge, Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press, 2002.

Appendix

Calculation of Correction Terms

Under the condition that $E(\varepsilon_3|\varepsilon_1,\varepsilon_2) = \gamma_1\varepsilon_1 + \gamma_2\varepsilon_2$ one can characterize the coefficients (using the usual linear projection formula) as,

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \frac{\sigma_3}{1-\rho_{12}^2}\begin{pmatrix} \rho_{13}-\rho_{12}\rho_{23} \\ \rho_{23}-\rho_{12}\rho_{13} \end{pmatrix}$$

To find the correction terms $E(\varepsilon_j|\varepsilon_1 > c_1, \varepsilon_2 > c_2)$ we note that the density of $\varepsilon_1|\varepsilon_1 > c_1, \varepsilon_2 > c_2$ is given by,

$$\frac{\int_{c_2}^{\infty}\phi(\varepsilon_1,\varepsilon_2;\Sigma)d\varepsilon_2}{\int_{c_1}^{\infty}\int_{c_2}^{\infty}\phi(\varepsilon_1,\varepsilon_2;\Sigma)d\varepsilon_2 d\varepsilon_1} = \frac{\phi(\varepsilon_1)(1-\Phi((c_2-\rho_{12}\varepsilon_1)/\sqrt{1-\rho_{12}^2})}{\int_{c_1}^{\infty}\int_{c_2}^{\infty}\phi(\varepsilon_1,\varepsilon_2;\Sigma)d\varepsilon_2 d\varepsilon_1}$$

Then,

$$\begin{aligned} E(\varepsilon_1|\varepsilon_1 &> c_1, \varepsilon_2 > c_2) \\ &= \frac{\int_{c_1}^{\infty}\varepsilon_1\phi(\varepsilon_1)(1-\Phi((c_2-\rho_{12}\varepsilon_1)/\sqrt{1-\rho_{12}^2})d\varepsilon_1}{\int_{c_1}^{\infty}\int_{c_2}^{\infty}\phi(\varepsilon_1,\varepsilon_2;\Sigma)d\varepsilon_2 d\varepsilon_1} \\ &= \frac{\phi(c_1)(1-\Phi(c_2^*)) + \rho_{12}\phi(c_2)(1-\Phi(c_1^*))}{\int_{c_1}^{\infty}\int_{c_2}^{\infty}\phi(\varepsilon_1,\varepsilon_2;\Sigma)d\varepsilon_2 d\varepsilon_1} \end{aligned}$$

with the last line following using integration by parts. Then using a similar expression for $E(\varepsilon_2|\varepsilon_1 > c_1, \varepsilon_2 > c_2)$ and the definition of (γ_1, γ_2) we can derive the correction terms detailed in the text.

Preliminary Estimation

Here we derive the likelihood for (y_1, y_2) under the assumption that $(\varepsilon_1, \varepsilon_2) \sim N(0, \Sigma)$. We can write the joint density of (y_1, y_2) as,

$$f(y_1, y_2|x_1, x_2) = f(y_1|x_1, x_2)f(y_2|y_1 = 1, x_1, x_2)^{y_1}$$

where the second term on the right reflects the fact that we only see y_2 when $y_1 = 1$. Now it is straightforward to see that,

$$\begin{aligned} f(y_1|x) &= P(y_1 = 1|x_1)^{y_1}P(y_1 = 0|x_1)^{1-y_1} \\ &= \Phi(x_1'\delta_1)^{y_1}(1-\Phi(x_1'\delta_1))^{1-y_1} \end{aligned}$$

Next for y_2 we have that we can write,

$$\varepsilon_2 = \rho_{12}\varepsilon_1 + v_2$$

where v_2 is indepenent of ε_1 and $v_2 \sim N(0, 1-\rho_{12}^2)$. Then

$$\begin{aligned} P(y_2 &= 1|\varepsilon_1, x) = E(1(x_2'\delta_2 + \varepsilon_2 > 0)|\varepsilon_1, x) \\ &= E(1(x_2'\delta_2 + \rho_{12}\varepsilon_1 + v_2 > 0)|\varepsilon_1, x) = \Phi\left(\left(x_2'\delta_2 + \rho_{12}\varepsilon_1\right)/(1-\rho_{12}^2)^{1/2}\right) \\ P(y_2 &= 0|\varepsilon_1, x) = 1 - \Phi\left(\left(x_2'\delta_2 + \rho_{12}\varepsilon_1\right)/(1-\rho_{12}^2)^{1/2}\right) \end{aligned}$$

Then,

$$\begin{aligned} P(y_2 &= 1|y_1 = 1, x) = E(1(x_2'\delta_2 + \varepsilon_2 > 0)|\varepsilon_1 > -x_1'\delta_1, x) \\ &= E(\Phi\left(\left(x_2'\delta_2 + \rho_{12}\varepsilon_1\right)/(1-\rho_{12}^2)^{1/2}\right)|\varepsilon_1 > -x_1'\delta_1, x) \\ P(y_2 &= 0|y_1 = 1, x) = 1 - E(\Phi\left(\left(x_2'\delta_2 + \rho_{12}\varepsilon_1\right)/(1-\rho_{12}^2)^{1/2}\right)|\varepsilon_1 > -x_1'\delta_1, x) \end{aligned}$$

where,

$$\begin{aligned} &E(\Phi\left(\left(x_2'\delta_2 + \rho_{12}\varepsilon_1\right)/(1-\rho_{12}^2)^{1/2}\right)|\varepsilon_1 > -x_1'\delta_1, x) \\ &= \frac{1}{\Phi(x_1'\delta_1)} \int_{-x_1'\delta_1}^{\infty} \Phi\left(\left(x_2'\delta_2 + \rho_{12}\varepsilon_1\right)/(1-\rho_{12}^2)^{1/2}\right)\phi(\varepsilon_1)d\varepsilon_1 \\ &= P(y_2 = 1|y_1 = 1, x) \end{aligned}$$

with $P(y_2 = 0|y_1 = 1, x)$ being defined similarly. Note that these integrals need to be computed numerically using quadrature. Then we have,

$$f(y_2|y_1 = 1, x) = P(y_2 = 1|y_1 = 1, x)^{y_2} P(y_2 = 0|y_1 = 1, x)^{1-y_2}$$

and the joint density of y_2, y_1 is,

$$f(y_1, y_2|x) = f(y_1) f(y_2|y_1 = 1, x)^{y_1}$$

The method referred to as the "ρ free, two step" approach involves estimating δ_1, δ_2 and ρ_{12} jointly by maximizing the full log-likelihood

$$\log L(\delta_1, \delta_2, \rho_{12}) = \sum_{i=1}^{n} \log(f(y_{1i}, y_{2i}|x_{1i}, x_{2i}))$$

The method referred to as "ρ free, three step" approach involves estimating δ_1 by maximizing,

$$\log L(\delta_1) = \sum_{i=1}^{n} \log f(y_{1i}|x_{1i})$$

and then estimating δ_2 and ρ_{12} jointly by maximizing,

$$\log L(\delta_2, \rho_{12}|\hat{\delta}_1) = \sum_{i=1}^{n} y_{1i} \log \hat{f}(y_2|y_1 = 1, x)$$

where,

$$\hat{f}(y_2|y_1=1,x) = \hat{P}(y_2=1|y_1=1,x)^{y_2}\hat{P}(y_2=0|y_1=1,x)^{1-y_2}$$
$$\hat{P}(y_2=1|y_1=1,x) = \frac{1}{\Phi(x_1'\hat{\delta}_1)}\int_{-x_1'\hat{\delta}_1}^{\infty}\Phi((x_2'\delta_2+\rho_{12}\varepsilon_1)/(1-\rho_{12}^2)^{1/2})\phi(\varepsilon_1)d\varepsilon_1$$
$$\hat{P}(y_2=0|y_1=1,x) = 1-\hat{P}(y_2=1|y_1=1,x)$$

For both of these methods, as noted in Wooldridge (2002), identification of δ_2 is on a more solid foundation (other than through functional form assumptions) if one can have an exclusion restriction such that there is one variable in x_1 that is not in x_2. This is achieved in our case by having the affinity measure in x_1 but not in x_2.

In the third case where "$\rho = 0$" it is easy to see that the joint likelihood simplifies. One can estimate δ_1 by maximizing,

$$\log L(\delta_1) = \sum_{i=1}^{n}\log f(y_{1i}|x_{1i})$$

and then estimate δ_2 by maximizing,

$$\log L(\delta_2) = \sum_{i=1}^{n} y_{1i}\log(\Phi(x_{2i}'\delta_2)^{y_{2i}}(1-\Phi(x_{2i}'\delta_2))^{1-y_{2i}}$$

which are a simple Probit model for y_1 on the whole sample and a Probit model for y_2 on the sample for whom $y_{1i}=1$.

Table 1. Sampling Frame and Response Rate by Major

Major	No. in Sampling Frame (1)	No. of Respondents (2)	Response Rate (percent) (3)
Architecture and Fine Arts	422	74	17.5
Business—Soft[a]	629	164	26.1
Business—Hard	1026	271	26.4
Communications	997	233	23.4
Education	544	128	23.5
Engineering	863	264	30.6
Humanities	759	156	20.6
Natural Sciences and Pharmacology	1188	306	25.8
Nursing and Social Work	165	50	30.3
Plan II[b]	138	45	32.6
Social Sciences[c]	1239	324	26.2
TOTAL	**7970**	**2015**	**25.3**

In Tables 1-5:

[a]The "Hard" business majors are accounting, actuarial science, business engineering, data processing, finance, management information science, and a few miscellaneous descriptions. The "Soft" business majors are all the rest, mostly general business, management and marketing.

[b]A special multidisciplinary honors major.

[c]Anthropology, economics, geography, government, psychology and sociology.

Table 2. Means of Salary and Major Control Variables by Major (Standard Deviations in Parentheses)

Major	Current Salary ($)	SAT	Percent Female	College GPA	HS Area Income ($)	Weekly Hours	Upper Div. Sci. & Math Grades	Masters or Doctorate	Cell Size
Architecture, Fine Arts	53,214 (31,553)	1,184 (151)	0.536 (.503)	3.139 (.420)	38,772 (12,619)	41.804 (10.705)	2.562 (.729)	0.268 (.447)	56
Business--Soft[a]	109,052 (128,053)	1,145 (119)	0.478 (.501)	2.817 (.431)	39,071 (13,092)	43.284 (14.028)	2.557 (.684)	0.313 (.466)	134
Business--Hard	124,372 (136,368)	1,185 (135)	0.318 (.467)	3.055 (.457)	39.612 (14,147)	45.247 (13.816)	2.888 (.616)	0.296 (.458)	223
Communications	77,874 (93,839)	1,128 (135)	0.599 (.491)	2.891 (.439)	37,968 (14,335)	40.967 (14.306)	2.753 (.719)	0.209 (.408)	182
Education	43,233 (24,488)	1,114 (118)	0.791 (.409)	3.045 (.425)	40,551 (14,534)	40.767 (12.311)	2.832 (.815)	0.267 (.445)	86
Engineering	102,293 (90,714)	1,237 (124)	0.185 (.389)	3.026 (.494)	39,609 (13,963)	44.833 (11.908)	2.988 (.529)	0.302 (.460)	222
Humanities	56,524 (37,058)	1,190 (154)	0.524 (.501)	3.064 (.550)	40,259 (13,492)	42.427 (15.255)	2.777 (.987)	0.516 (.502)	124
Natural Sciences, Pharmacology	91,796 (104,227)	1,194 (136)	0.542 (.499)	2.984 (.493)	38,173 (14,517)	42.119 (14.260)	2.900 (.620)	0.392 (.489)	260
Nursing, Social Work	48,900 (36,805)	1,128 (102)	0.850 (.362)	3.055 (.462)	38,095 (11,490)	33.850 (16.143)	3.006 (.738)	0.425 (.501)	40
Plan II[b]	128,290 (177,275)	1,364 (120)	0.579 (.500)	3.626 (.281)	46,901 (17,656)	43.605 (19.748)	3.449 (.989)	0.605 (.495)	38
Social Sciences[c]	79,805 (82,540)	1,138 (147)	0.500 (.501)	2.851 (.542)	39,927 (14,733)	40.030 (15.436)	2.519 (.574)	0.432 (.496)	266
All Fields	88,819 (101,388)	1,178 (143)	0.477 (.500)	2.985 (.498)	39,408 (14,200)	42.316 (14.303)	2.829 (.477)	0.351 (.477)	1631

Table 3. Regression Estimates of the Determinants of Earnings, and Probit Estimates of Survey Response[a]

		Log(Current Earnings)			Respond=1
		(1)	(2)	(3)	(4)
Major:	Architecture and Fine Arts	0.165 (2.26)	0.179 (2.43)	0.170 (2.34)	-0.044 (1.610)
	Business---Soft	0.413 (5.37)	0.421 (5.50)	0.417 (5.56)	0.046 (1.720)
	Business---Hard	0.522 (7.07)	0.514 (6.89)	0.501 (6.83)	0.038 (1.450)
	Communications	0.366 (5.09)	0.359 (4.98)	0.347 (4.95)	0.023 (1.000)
	Engineering	0.372 (3.94)	0.367 (3.85)	0.367 (3.94)	0.086 (2.510)
	Humanities	0.097 (1.35)	0.100 (1.400)	0.099 (1.39)	-0.013 (0.540)
	Plan II	0.393 (2.74)	0.406 (2.79)	0.418 (2.92)	0.087 (2.080)
	Social Sciences	0.314 (4.84)	0.307 (4.72)	0.300 (4.77)	0.052 (2.280)
	Natural Sciences, Pharmacology	0.293 (3.72)	0.299 (3.80)	0.294 (3.83)	0.040 (1.520)
	Nursing, Social Work	0.212 (2.37)	0.229 (2.62)	0.230 (2.71)	0.061 (1.570)
	Class of 1980	0.644 (11.91)	0.680 (11.42)	0.645 (11.03)	0.025 (1.610)
	Class of 1985	0.571 (11.35)	0.590 (11.43)	0.559 (10.97)	-0.009 (0.610)
	Class of 1990	0.495 (10.45)	0.507 (10.57)	0.488 (10.29)	0.041 (2.650)
	Class of 1995	0.319 (7.000)	0.322 (7.08)	0.313 (6.89)	0.033 (2.200)
	GPA	0.024 (0.70)	0.057 (1.56)	0.067 (1.86)	0.027 (2.590)

Table 3, cont.

	Log(Current Earnings)			Respond=1
	(1)	(2)	(3)	(4)
Upper Div. Sci. & Math Credits	0.0014	0.0019	0.0021	0.0001
	(1.22)	(1.61)	(1.77)	(0.210)
Upper Div. Sci. & Math Grades	0.029	0.027	0.023	0.002
	(1.74)	(1.65)	(1.43)	(0.510)
HS Area Income ($000)	0.003	0.003	0.002	0.001
	(3.06)	(2.34)	(2.15)	(1.920)
Masters	-0.019	-0.020	-0.013	
	(0.51)	(0.55)	(0.34)	
Doctorate	0.172	0.158	0.128	
	(2.99)	(2.72)	(2.20)	
Married	0.256	0.252	0.252	
	(5.50)	(5.45)	(5.470)	
Female	-0.082	-0.082	-0.071	0.031
	(1.72)	(1.68)	(1.48)	(3.060)
Married*Female	-0.277	-0.276	-0.284	
	(4.54)	(4.55)	(4.69)	
Weekly Hours	0.028	0.028	0.028	
	(13.87)	(13.79)	(13.94)	
SAT		-0.00012	-0.0011	
		(0.83)	(0.72)	
HS Percentile		-0.0026	-0.0024	
		(1.76)	(1.69)	
Student Loan			0.016	
			(0.49)	
Self Employed			0.224	
			(3.90)	
N	1501	1501	1501	7970
R^2	0.477	0.481	0.490	

[a]Robust t-tatistics in parentheses. The excluded major is Education.

Table 4. Regression Estimates of the Determinants of Earnings with Correction Terms (N=1501)[a]

		Functional Form	$\rho = 0$	ρ free, two-step	ρ free, three-step	Propensity Score Linear	Propensity Score Quadratic
		(1)	(2)	(3)	(4)	(5)	(6)
Major:	Architecture and Fine Arts	-1.025	0.178	0.177	0.175	0.188	0.162
		(0.96)	(2.37)	(2.28)	(2.33)	(2.52)	(2.04)
	Business---Soft	1.504	0.389	0.389	0.389	0.396	0.378
		(1.52)	(5.05)	(4.99)	(5.04)	(5.12)	(4.80)
	Business---Hard	1.417	0.503	0.503	0.502	0.509	0.489
		(1.74)	(6.73)	(6.41)	(6.73)	(6.79)	(6.36)
	Communications	0.919	0.350	0.350	0.349	0.361	0.342
		(1.80)	(4.72)	(4.62)	(4.70)	(4.86)	(4.47)
	Engineering	2.313	0.336	0.336	0.336	0.339	0.316
		(1.32)	(3.47)	(3.38)	(3.47)	(3.50)	(3.28)
	Humanities	-0.239	0.099	0.099	0.098	0.110	0.086
		(0.77)	(1.38)	(1.36)	(1.35)	(1.51)	(1.16)
	Plan II	2.314	0.333	0.334	0.333	0.363	0.338
		(1.33)	(2.23)	(2.23)	(2.23)	(2.43)	(2.25)
	Social Sciences	1.546	0.292	0.292	0.292	0.298	0.279
		(1.39)	(4.50)	(4.43)	(4.50)	(4.57)	(4.18)
	Natural Sciences, Pharmacology	1.234	0.270	0.270	0.270	0.285	0.265
		(1.44)	(3.39)	(3.28)	(3.39)	(3.56)	(3.27)
	Nursing, Social Work	1.585	0.162	0.162	0.159	0.213	0.196
		(1.26)	(1.66)	(1.64)	(1.64)	(2.20)	(2.03)
	Class of 1980	1.229	0.619	0.619	0.617	0.655	0.646
		(2.25)	(9.21)	(8.96)	(9.26)	(10.47)	(10.41)
	Class of 1985	0.325	0.568	0.568	0.566	0.594	0.584
		(1.65)	(9.83)	(9.55)	(9.76)	(10.91)	(10.47)
	Class of 1990	1.447	0.461	0.462	0.460	0.494	0.488
		(1.65)	(8.21)	(8.14)	(8.31)	(9.41)	(9.36)
	Class of 1995	1.091	0.281	0.282	0.279	0.332	0.326
		(1.52)	(4.33)	(4.15)	(4.34)	(5.66)	(5.61)
	GPA	0.699	0.022	0.022	0.023	0.006	0.007
		(1.26)	(0.58)	(0.55)	(0.61)	(0.17)	(0.17)
	Upper Div. Sci. & Math Credits	0.003	0.001	0.001	0.001	0.002	0.002
		(1.60)	(1.08)	(0.47)	(1.08)	(1.27)	(1.31)
	Upper Div. Sci. & Math Grades	0.086	0.027	0.027	0.027	0.027	0.027
		(1.60)	(1.63)	(1.58)	(1.63)	(1.62)	(1.65)

Table 4, cont.

		Functional Form	$\rho = 0$	ρ free, two-step	ρ free, three-step	Propensity Score Linear	Propensity Score Quadratic
		(1)	(2)	(3)	(4)	(5)	(6)
	HS Area Income ($000)	-0.012	0.004	0.004	0.004	0.004	0.004
		(0.82)	(3.31)	(1.49)	(3.32)	(3.22)	(3.14)
	Masters	-0.026	-0.027	-0.027	-0.028	-0.015	-0.015
		(0.67)	(0.71)	(0.69)	(0.73)	(0.40)	(0.39)
	Doctorate	0.167	0.163	0.164	0.163	0.170	0.171
		(2.89)	(2.82)	(2.80)	(2.82)	(2.96)	(2.98)
	Married	0.244	0.245	0.246	0.244	0.266	0.292
		(5.02)	(5.05)	(4.38)	(5.01)	(5.59)	(5.93)
	Female	0.674	-0.094	-0.094	-0.093	-0.102	-0.104
		(0.99)	(1.85)	(1.84)	(1.82)	(2.02)	(2.03)
	Married*Female	-0.247	-0.243	-0.244	-0.241	-0.306	-0.309
		(3.27)	(3.23)	(2.83)	(3.17)	(4.22)	(4.29)
	Weekly Hours	0.028	0.027	0.027	0.027	0.028	0.027
		(14.03)	(13.90)	(13.99)	(13.89)	(13.92)	(13.91)
Correction factors:	λ_1	9.332	-0.185	-0.215	-0.222		
		(1.11)	(1.80)	(1.81)	(1.92)		
	λ_2	-0.129	-0.131	-0.146	-0.164		
		(0.59)	(0.60)	(0.62)	(0.72)		
Propensity Scores:	Responded					0.452	0.681
						(1.91)	(0.29)
	Working					-0.199	3.994
						(0.68)	(1.69)
	Responded^2						0.661
							(0.26)
	Working^2						-2.319
							(1.84)
	Responded* Working						-0.629
							(0.33)
	Constant	-7.014	9.077	9.070	9.070	8.871	7.069
		(0.49)	(38.25)	(35.96)	(38.90)	(27.84)	(6.23)
	R^2	0.478	0.479	0.479	0.479	0.479	0.481

[a]Robust t-statistics in parentheses. Robust and correct t-statistics in parentheses for the efficient model. The excluded major is Education.

Table 5. Earnings Differential by Major (Log Points), Unadjusted and Adjusted

Major	Raw Differential	Adjusted Differential Based on Table 3	Adjusted Differential Based on Table 4
	(1)	(2)	(3)
Architecture, Fine Arts	0.212	.165	0.162
Business--Soft[a]	0.722	.413	0.378
Business--Hard	0.899	.522	0.489
Communications	0.393	.366	0.342
Education	0.000	.000	0.000
Engineering	0.801	.372	0.316
Humanities	0.210	.097	0.086
Plan II	0.614	.393	0.338
Social Sciences	0.410	.314	0.279
Natural Sciences, Pharmacology	0.538	.293	0.265
Nursing, Social Work	0.030	.212	0.196
Standard Deviation	**0.305**	**0.153**	**0.139**