# DISCUSSION PAPER SERIES

No. 9131

## A CAUTIONARY NOTE ON USING INDUSTRY AFFILIATION TO PREDICT INCOME

Jörn-Steffen Pischke and Hannes Schwandt

*LABOUR ECONOMICS*


Centre for Economic Policy Research

**www.cepr.org**

**Available online at:** **www.cepr.org/pubs/dps/DP9131.asp**

*ISSN 0265-8003*

# A CAUTIONARY NOTE ON USING INDUSTRY AFFILIATION TO PREDICT INCOME

**Jörn-Steffen Pischke, London School of Economics and CEPR**
**Hannes Schwandt, Princeton University**

Discussion Paper No. 9131
September 2012

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801, Fax: (44 20) 7183 8820
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 9131

September 2012

# ABSTRACT

## A Cautionary Note on Using Industry Affiliation to Predict Income

Many literatures investigate the causal impact of income on economic outcomes, for example in the context of intergenerational transmission or well-being and health. Some studies have proposed to use employer wage differentials and in particular industry affiliation as an instrument for income. We demonstrate that industry affiliation is correlated with fixed individual characteristics, specifically parents' education and own height, conditional on the covariates typically controlled for in these studies. These results suggest that there is selection into industries based on unobservables. As a result the exclusion restriction in many IV studies of this type is likely violated.

JEL Classification: D31, I14, I3 and J31
Keywords: happiness, health, industry wage differentials and intergenerational mobility

Jörn-Steffen Pischke
Centre for Economic Performance
London School of Economics
Houghton Street
London WC2A 2AE

Email: s.pischke@lse.ac.uk

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=103890

Hannes Schwandt
Center for Health and Wellbeing
315 Wallis Hall
Princeton, NJ 08544-1013

Email: schwandt@princeton.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=176025

Submitted 06 September 2012

# 1 Introduction

Many literatures ask about the causal impact of income on a variety of economic outcomes. One example is intergenerational transmission; the question whether children of richer parents fare better in terms of their own economic performance (Solon, 1999). Another example is the health-income gradient: higher income individuals tend to be healthier. Is this association causal, do richer individuals differ in other dimensions, or does the causality run the opposite way (Cutler, Lleras-Muney, Vogl, 2011)? Furthermore, income is an important correlate in happiness regressions relating individual life satisfaction to various economic determinants (Blanchflower and Oswald, 2004). Does higher income cause happiness?

Sorting out causality in these and other applications is difficult. A number of studies in the literature have used employer wage differentials and particularly industry affiliation as an instrumental variable for individual or family income. Examples of such studies are Shea (2000) who uses industry affiliation, union membership, and job loss as instruments for parents' income in order to learn about intergenerational transmission. Luttmer (2005) uses interactions of industry and occupation as instruments for income in happiness regressions, an avenue also followed by Luechinger (2009). This idea has been extended by Pischke (2011), who focuses on industry differentials only, and has also been used by Li et al. (2011).

Industry wage differentials in particular are large and remain after controlling for many covariates including individual fixed effects (Krueger and Summers, 1988; Gibbons and Katz, 1992) and ability test scores (Blackburn and Neumark, 1992). As a result, researchers using these wage differentials as instruments for income may have a reasonable claim of identifying causal effects of income not contaminated by unobserved personal attributes and selection. In this note we provide some evidence to caution against this optimism. Following the same specifications as in Pischke (2011) we show

that industry differentials correlate with mother's education and own height. Both of these variables are pre-determined by the time industry affiliation is measured, and should therefore not be affected by industry choice.

The statistical association with industry wage effects which we find therefore suggests selection in industry choice which correlates with personal characteristics related to parental background and height. These characteristics might, for example, be unmeasured cognitive or non-cognitive skills or personality traits. We conclude by discussing why we think the types of variables we use, mother's education and height, might be useful as specification checks but offer little hope in improving inference in regression or IV frameworks. We conjecture that this remains true for even for better measures of the underlying omitted skills or traits, like ability test scores.

# 2 Data

In order to study the effect of parental education we use data from the 1972 to 2006 waves of the US General Social Survey (GSS). The GSS is a repeated cross-sectional survey, carried out every one to two years. The GSS is also the main dataset analyzed in Pischke (2011) and we use the same sample restrictions. The primary sample consists of employed men aged $20 - 64$ without missing values for any of the covariates used in the analysis.

Our outcome measures of primary interest are happiness and parental education. Happiness is transformed as explained in Pischke (2011), so that all results can be interpreted in terms of standard deviation units. As main parental education variable we use the highest year of completed schooling of the respondent's mother. We exclude missing values and those individuals reporting zero years of education. We focus on mother's education because there are fewer missing values in mother's education than in father's education. Using father's education or combining the education of both parents yields comparable results.

Family income, our main regressor, is a bracketed variable and we assign midpoints as described in Pischke (2011). Industry and occupation affiliation of the respondent are aggregated into 32 and 22 categories, respectively. Details are given in the appendix.

The wives sample consists of married females aged 20 to 64 with employed husbands. The industry and occupation variables used in this sample refer to the husbands, while other controls refer to the respondent.

In order to corroborate our results on parental education from the GSS we also present results on body height using the US National Health Interview Survey (NHIS). The NHIS is also a repeated cross-section carried out every year with about 30,000 to 50,000 respondents per wave. We use waves from 1974 to 2009 and restrict the sample to employed men aged 20-64.

Our outcome measures of primary interest in the NHIS are self-reported health and self-reported body height. Self-reported health is the answer to the question "Would you say your health in general is excellent, very good, fair, or poor?" We transform this health measure the same way as happiness in the GSS such that results can be interpreted in terms of standard deviations. Body height is measured in centimeters.

We group industry and occupation affiliations of the respondent into 28 and 18 categories, respectively, These groups are relatively commensurate with the GSS classifications, although a precise correspondence is not possible. Details are in the appendix.

# 3 Results

Table 1 displays regressions of happiness on the logarithm of family income for the sample of men from the GSS. Each column shows results for a different set of included controls. The specifications in each column are estimated by OLS, 2SLS and using Ackerberg and Devereux's (2009) Improved Jackknifed IV estimator (IJIVE). In the first four columns we reestimate the baseline

specifications from Pischke (2011) in the sample without missing values for mother's education. In columns (5) to (8) we repeat these regressions including mother's education as an additional control.

The OLS specifications in the first row indicate a positive and significant association of family income with happiness that is robust to the inclusion of a broad set of control variables. Using industry affiliation as instrument for family income results in somewhat higher and less precisely estimated coefficients. The coefficients in the first four columns are very similar to the estimates in Pischke (2011) although the IV results are slightly higher.

In columns (5) to (8) we repeat the baseline regressions including the years of education of the respondent's mother as additional control variable. Maternal education is an important component of a respondent's family background. If there is strong selection into industries due to unobservables related to a respondent's family background then the inclusion of mother's education should substantially change our IV estimates. This is not the case: coefficient estimates are virtually identical to those in the earlier columns. These results might suggest that omitted variable bias due to unobserved heterogeneity in family background is not particularly important. In other words, selection into industries based on family background does not seem to be a major issue.

But this is not the end of the story. In Table 2 we look at this potential selection issue from a different angle. Instead of including mother's education as a control we use this proxy for family background as dependent variable. The first four columns display the baseline results from the previous table. In columns (5) to (8) we repeat these regressions using as dependent variable mother's education instead of the respondent's happiness. In columns (5) and (6) the pattern of results is remarkably similar to those for happiness. There is a strong correlation between income and mother's education. IV estimates are larger than OLS estimates. It is difficult to think of these results as causal effects: income differences due to industry affiliation should not affect

predetermined parental education. So the obvious conclusion to draw is that these estimates reflect selection bias. Estimates are lower in columns (7) and (8) where we include occupation dummies as controls. But standard errors are large so that this is probably just due to sampling variation as we don't find a similar pattern in the NHIS data described below.

In Table 3 we turn to results for married women. Pischke (2011) uses this group as a specification check. If the correlations between industry income differentials and happiness are due to selection we should see less of an association between husbands industry affiliation and wife's happiness. In particular, for working wives we can control for the woman's own industry, which should be a reasonable guard against selection effects. In the sample we analyze here the IV estimates for income in the regressions using happiness for working wives are a lot lower than those for men. This is not the case in the original sample used in Pischke (2011) where the results for working wives differed little from those for the husbands. In any case, there is no evidence in either sample that controlling for wives' own industry affiliation (which should take care of any selection) lowers the coefficient on income.

Turning to the results using mother's education as dependent variable we find much more consistent results across specifications. The IV results are generally large and well above the OLS results. Controlling for wives' own industry does nothing to the results. Pischke (2011) took this type of evidence as support for the causal interpretation of the results on happiness. Our results for mother's education cast doubt on the usefulness of the wives sample as a specification check since a causal interpretation does not make sense for this dependent variable.

Using the same sample, in Table 4 we turn to the type of instrument set used in Luttmer (2005) combining both industry affiliation and occupation. We find that mother's education is even more strongly related to occupation than industry wage differentials. This suggests that neither variable is useful in order to generate exogenous variation in income.

In Table 5 we explore other employer characteristics as instruments, including union status used in Shea (2001). Our inference here is hindered by the fact that the IV results for mother's education are very imprecisely estimated. The point estimates using union status and firm size are negative and numerically large but not statistically significant. As a result, we can say relatively little about these employer attributes from the small GSS samples.

In order to corroborate our findings further we turn to data from the NHIS to look at health as an outcome. While health is an important outcome in its own right another advantage of the NHIS is the much larger sample sizes compared to the GSS. Unfortunately, the NHIS offers less employment related information. So we return to industry affiliation as the instrument here. We check the health results using height as an alternative outcome which should not be affected by industry affiliation. Body height is largely fixed by age 20 (Hamill, 1977).

Columns (1) to (3) of Table 6 reveal IV coefficients slightly below the OLS results. Taken at face value, these results suggest that a sizeable portion of the association between health and income is due to a causal effect running from income to health with something of an upward bias to the OLS results. However, as with our findings for mother's education, the results for height shed doubt on this interpretation. The association between income and height becomes stronger in the IV estimates cautioning against interpreting these results causally. The results in this larger sample are more precisely estimated that those from the GSS, and they are very stable independently of what controls are included.

In Table 6 we use a sample of men aged 20 – 64, similar to the sample from the GSS used for tables 1 and 2. One caveat with this age group is that older men tend to shrink slightly, and this may be related to income. To guard against this we also present results for men aged 20 – 29 in Table 7, an age group for whom this should not be a concern. The large sample sizes in the NHIS easily facilitate such cuts of the data. The results are qualitatively

very similar to those in Table 6, so differential shrinkage does not seem to be a factor in our findings.

# 4 Interpretation

Our results mirror earlier findings by Blackburn and Neumark (1992) regarding the association of industry wage differentials and ability measures. These authors found that ability measures like IQ type test scores are strongly correlated with industry wage differentials. On the other hand, controlling for the test scores hardly alters estimated industry effects in a wage regression. We find the same pattern of results for mother's education and height. How do these results make sense, and what do they tell us about various empirical strategies to uncover causal effects in the type of applications we have discussed, which use income as a (potentially endogenous) regressor?

We discuss these issues in the following statistical framework. Consider the regression equation

$$h_i = \beta y_i + \gamma x_i + \varepsilon_i \tag{1}$$

where $h_i$ is an outcome like health, happiness, or children's income, $y_i$ is own income, $x_i$ is a confounder, and $\varepsilon_i$ is a regression residual, orthogonal to $y_i$ and $x_i$. Conditional on the confounder $x_i$ the regression coefficient $\beta$ is the causal effect of income on $h_i$. $x_i$ may correspond to difficult to observe attributes like ability, personal traits, etc. We will also assume that conditional on the confounder $x_i$ industry affiliation is random. Hence, we can interpret our IV results as replacing income $y_i$ in (1) by its prediction using industry wage effects.

Income is related to the confounder by the regression equation

$$y_i = \delta x_i + v_i. \tag{2}$$

The confounder $x_i$ is not directly observed. Instead, we observe a proxy

correlated with $x_i$ characterized by the regression

$$m_i = \lambda x_i + \ u_i \tag{3}$$

where $m_i$ might be mother's education or own height. We assume $cov(v_i, u_i) = 0$, i.e. only $x_i$ links income with the measure $m_i$.

Here, we give some intuition for our results within this framework. We derive the relevant implications formally in the appendix. Our regression results indicate that income matters for mother's education (or height) both in the OLS and IV regressions (Table 2, cols. (5) to (8)). This implies both $\lambda$ and $\delta$ have to be positive, since $x_i$ is the only link between these two variables.

Our next finding is that the estimated effect of income on the outcome $h_i$ is basically the same whether we control for $m_i$ in the regression or not (Table 1, cols (1) to (4) versus (5) to (8)). One explanation for the results is that the noise in the measure $m_i$ is large (i.e. $\sigma_u^2 \gg 0$). This noise is a classical measurement error and does not matter when $m_i$ is on the left hand side, so the measure $m_i$ is still useful to glean information on whether there is any bias in the OLS or IV regressions (from the $m_i$ on $y_i$ regression). But the noise leads to attenuation when using $m_i$ on the right hand side of the regression, so it becomes essentially useless as a control in this case. This seems to make sense for measures like mother's education and particularly height, which clearly must be very crude variables to get at the true underlying confounder $x_i$. Variables like ability test scores should be better proxies for $x_i$, in the sense that $\lambda$ is larger and/or $\sigma_u^2$ smaller. However, the Blackburn and Neumark (1992) results indicate that even ability test scores exhibit pretty much the same features. As a result the available variables of this type seem to offer little mileage as controls in regression studies. Our insights here are not new; they are very much reminiscent of Griliches (1977) in his discussion of estimating the returns to schooling.

In order to interpret the IV results in this context think of replacing income with the part of income predicted by industry wage effects. For this

predicted income, the relative size of $\delta$ and $\sigma_v^2$ may be different compared to raw income (which is relevant for the OLS results). The IV results for happiness are above the OLS estimates, and this is true for the regression of mothers education as well. Within this particular framework this has to imply that $\delta$ becomes more important relative to $\sigma_v^2$ for predicted income. This implies that both the OLS and IV results are biased upwards but the IV results are biased even more.

Let us make this slightly more precise. In terms of our framework in eqs. (1) and (2) the variable $v_i$ is a valid instrument for income. In other words, the variation in income due to a valid instrument needs to satisfy the condition $\delta = 0$. This motivates using the IV regression of $m_i$ on $y_i$ as a specification check. In practice, few instruments might satisfy the condition $\delta = 0$ literally. Moreover, most instruments use little of the variation in the endogenous regressor. As a result, both $\delta$ and $\sigma_v^2$ will be lower in the IV case. We show in the appendix that IV is less biased than OLS whenever $\delta / \left(\delta^2 + \sigma_v^2\right)$ is smaller for the income predicted by the instrument than for raw income. $\delta$ may be very small for predicted income but this is of little comfort with relatively weak instruments, an insight which goes back to Bound, Jaeger and Baker (1995). These authors show that the relative bias of IV is related to the size of the partial $R^2$ or $F$-statistic on the excluded instruments in the first stage. These metrics are not just important to assess small sample bias in two stage least squares regressions but also asymptotic bias due to small violations of the exclusion restrictions in the form of $\delta \neq 0$.

# 5 Conclusion

In this note we have assessed the usefulness of industry wage differentials as instruments in regression models for happiness or health with income as an endogenous regressor. Our conclusion is broadly negative: we do not believe that industry wage differentials offer a useful source of variation in income

to establish causal effects. This is based on OLS and IV regressions using fixed personal characteristics as left hand side variables in the respective regressions, which show large effects. One obvious, though hardly novel conclusion from our work is that a healthy degree of doubt about the use of instrumental variables is often warranted, even when the IV regressions pass some purported specification tests. Using fixed personal characteristics as dependent variable offers a useful specification check. We argue that this is the case even for variables which might be poorly measured and are of little value as control variables. Such variables should be available in many data sets.

# References


[1] Ackerberg, Daniel A., and Paul J. Devereux (2009) "Improved JIVE Estimators for Overidentified Linear Models with and without Heteroskedasticity," *Review of Economics and Statistics*, 91, No. 2 (May), 351-362.

[2] Blackburn, McKinley, and David Neumark (1992) "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials," *Quarterly Journal of Economics*, Vol. 107, No. 4 (Nov.), 1421-1436.

[3] Blanchflower, David G. and Andrew J. Oswald (2004) "Well-being over time in Britain and the USA," *Journal of Public Economics*, Volume 88, Issues 7–8, July , 1359–1386.

[4] Bound, John, David A. Jaeger, and Regina M. Baker (1995) "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak," *Journal of the American Statistical Association*, Vol. 90, No. 430, June, 443-450.

[5] Cutler, David M., Adriana Lleras-Muney, and Tom Vogl (2011). "Socioeconomic Status and Health: Dimensions and Mechanisms", in Sherry Glied and Peter Smith (eds.) *Oxford Handbook of Health Economics*, Oxford: Oxford University Press.

[6] Gibbons, Robert and Lawrence F. Katz (1992) "Does Unmeasured Ability Explain Inter-Industry Wage Differentials?" *Review of Economic Studies*, 59, No. 3 (July), 515-535.

[7] Griliches, Zvi (1977) "Estimating the Returns to Schooling - Some Econometric Problems," *Econometrica*, vol. 45, January, 1-22

[8] Hamill, Peter V. V. (1977) "NCHS growth curves for children," Vital and health statistics: Series 11, Data from the National Health Survey; no. 165, Hyattsville, Md: National Center for Health Statistics.

[9] Krueger, Alan B., and Lawrence H. Summers (1988) "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica*, Vol. 56, No. 2 (March), 259-293.

[10] Li, Hongbin, Pak Wai Liu, Maoliang Ye, and Junsen Zhang (2011) "Does Money Buy Happiness? Evidence from Twins in Urban China," Working Paper, Tsinghua University (April).

[11] Luechinger, Simon (2009) "Valuing Air Quality Using the Life Satisfaction Approach," *Economic Journal*, 119, No. 536 (March), 482-515.

[12] Luttmer, Erzo F. P. (2005) "Neighbors as Negatives: Relative Earnings and Well-Being," *Quarterly Journal of Economics*, 120, No. 3 (August), 963-1002.

[13] Pischke, Jörn-Steffen (2011) "Money and Happiness: Evidence from the Industry Wage Structure," NBER Working Paper 17056 (May).

[14] Segal, Carmit (forthcoming) "Working When No One is Watching: Motivation, Test Scores, and Economic Success," *Management Science*.

[15] Shea, John (2000) "Does parents' money matter?" *Journal of Public Economics*, 77, No. 2, (August), 155-184.

[16] Solon, Gary (1999) "Intergenerational mobility in the labor market," chapter 29 in: Orley Ashenfelter and David Card (eds.) *Handbook of Labor Economics*, vol. 3A, Amsterdam: Elsevier.

# 6 Appendix

The variables $h_i$, $y_i$, $m_i$, and $x_i$ are defined by the regression equations

$$\begin{aligned} h_i &= \beta y_i + \gamma x_i + \varepsilon_i \\ y_i &= \delta x_i + v_i \\ m_i &= \lambda x_i + u_i. \end{aligned}$$

Normalize $x_i$ so that $var(x_i) = 1$.

Consider the regression of $y_i$ on $m_i$

$$\begin{aligned} y_i &= \pi m_i + \widetilde{y}_i \\ \pi &= \frac{cov(m_i, y_i)}{var(m_i)} = \frac{cov(\lambda x_i + u_i, \delta x_i + v_i)}{\lambda^2 + \sigma_u^2} \\ &= \frac{\lambda\delta}{\lambda^2 + \sigma_u^2} \end{aligned}$$

so that the residual of $y_i$ after filtering out $m_i$ is given by

$$\widetilde{y}_i = y_i - \pi m_i = y_i - \frac{\lambda\delta}{\lambda^2 + \sigma_u^2} m_i.$$

The variances are

$$\begin{aligned} var(y_i) &= \delta^2 + \sigma_v^2 = \left(\frac{\lambda\delta}{\lambda^2 + \sigma_u^2}\right)^2 \left(\lambda^2 + \sigma_u^2\right) + var(\widetilde{y}_i) \\ var(\widetilde{y}_i) &= \delta^2 + \sigma_v^2 - \frac{\lambda^2\delta^2}{\lambda^2 + \sigma_u^2} \\ &= \frac{\left(\lambda^2 + \sigma_u^2\right)\left(\delta^2 + \sigma_v^2\right) - \lambda^2\delta^2}{\lambda^2 + \sigma_u^2} \\ &= \frac{\lambda^2\sigma_v^2 + \sigma_u^2\left(\delta^2 + \sigma_v^2\right)}{\lambda^2 + \sigma_u^2} \end{aligned}$$

Finally note that the regression coefficient of $x_i$ on $\widetilde{y}_i$ is

$$\begin{aligned} b_{x\widetilde{y}} &= \frac{cov(x_i\,,\widetilde{y}_i)}{var(\widetilde{y}_i)} = \frac{cov\left(x_i\,,y_i-\frac{\lambda\delta}{\lambda^2+\sigma_u^2}m_i\right)}{var(\widetilde{y}_i)} \\ &= \frac{cov\left(x_i\,,\delta x_i+v_i-\frac{\lambda\delta}{\lambda^2+\sigma_u^2}\left(\lambda x_i+\ u_i\right)\right)}{var(\widetilde{y}_i)} \\ &= \frac{\delta-\frac{\lambda^2\delta}{\lambda^2+\sigma_u^2}}{var(\widetilde{y}_i)} = \frac{\delta\left(\lambda^2+\sigma_u^2\right)-\lambda^2\delta}{\lambda^2\sigma_v^2+\sigma_u^2\left(\delta^2+\sigma_v^2\right)} \\ &= \frac{\delta\sigma_u^2}{\lambda^2\sigma_v^2+\sigma_u^2\left(\delta^2+\sigma_v^2\right)}. \end{aligned}$$

Our primary interest is the regression

$$h_i = by_i + cm_i + e_i$$

and we get

$$\begin{aligned} b &= \frac{cov(h_i\,,\widetilde{y}_i)}{var(\widetilde{y}_i)} = \frac{cov(\beta y_i+\gamma x_i+\varepsilon_i\,,\widetilde{y}_i)}{var(\widetilde{y}_i)} \\ &= \frac{\beta var(\widetilde{y}_i)+\gamma cov(x_i\,,\widetilde{y}_i)}{var(\widetilde{y}_i)} \\ &= \beta+\gamma b_{x\widetilde{y}} = \beta+\gamma\frac{\delta\sigma_u^2}{\lambda^2\sigma_v^2+\sigma_u^2\left(\delta^2+\sigma_v^2\right)}. \end{aligned}$$

and running

$$\begin{aligned} h_i &= b^*y_i+e_i^* \\ b^* &= \beta+\gamma b_{xy} = \beta+\gamma\frac{cov(x_i\,,y_i)}{var(y_i)} \\ &= \beta+\gamma\frac{cov\left(x_i\,,\delta x_i+v_i\right)}{\delta^2+\sigma_v^2} \\ &= \beta+\gamma\frac{\delta}{\delta^2+\sigma_v^2} \end{aligned}$$

This corresponds to the regression coefficient we get from regressing $h_i$ on $y_i$ without controlling for $m_i$, i.e. the regressions in Tables 1 and 2, cols. (1) to (4).

Finally, consider the regression of $m_i$ on $y_i$, corresponding to the regressions in columns (5) to (8) of Table 2:

$$\begin{aligned} m_i &= dy_i + e_i^m \\ d &= \frac{cov(m_i\,, y_i)}{var(y_i)} = \frac{\lambda\delta}{\delta^2 + \sigma_v^2}. \end{aligned}$$

This demonstrates the claim that $d > 0$ implies $\lambda > 0$ and $\delta > 0$.

In the OLS results we basically find $b = b^*$ i.e. controlling for $m_i$ hardly matters in the regression of $h_i$ on $y_i$ (Table 1, cols (1) to (4) versus (5) to (8)). If we make $\sigma_u^2$ large we get

$$\lim_{\sigma_u^2 \to \infty} b_{x\tilde{y}} = \frac{\delta}{\delta^2 + \sigma_v^2} = b_{xy}.$$

On the other hand, $\sigma_u^2$ does not figure in the expression for $d$, so poor measurement does not affect the regression of $m_i$ on $y_i$.

**Coding of industries in the GSS**

| | Sector | GSS 1972 – 1988<br>1970 Census codes | GSS 1989 – 2006<br>1980 Census codes |
|---|---|---|---|
| 1 | Agriculture, forestry, fishery | 17-28 | 10-31 |
| 2 | Mining | 47-57 | 40-50 |
| 3 | Construction | 67-77 | 60 |
| 4 | Lumber, wood, furniture | 107-118 | 230-242 |
| 5 | Stone, clay, glass | 119-138 | 250-262 |
| 6 | Metal | 139-169 | 270-301 |
| 7 | Machinery, exc. electrical | 177-198, 258 | 310-332 |
| 8 | Electrical machinery | 199-209 | 340-350 |
| 9 | Transportation equipment | 219-238 | 351-370 |
| 10 | Professional equipment | 239-259 | 371-382 |
| 11 | Food and tobacco | 268-299 | 100-130 |
| 12 | Textile, apparel, leather | 307-327, 388-397 | 132-152, 220-222 |
| 13 | Paper | 328-337 | 160-162 |
| 14 | Printing | 338-339 | 171-172 |
| 15 | Chemicals | 347-369 | 180-192 |
| 16 | Petroleum and rubber | 377-387 | 200-212 |
| 17 | Other manufacturing | 259, 398 | 390-392 |
| 18 | Transportation | 407-429 | 400-432 |
| 19 | Communication | 447-449 | 440-442 |
| 20 | Utilities | 467-479 | 460-472 |
| 21 | Wholesale trade | 507-588 | 500-571 |
| 22 | Retail trade | 607-698 | 580-691 |
| 23 | Finance, insurance, real estate | 707-718 | 700-712 |
| 24 | Business services | 727-748 | 721-742 |
| 25 | Repair services | 749-759 | 750-760 |
| 26 | Personal services | 769-798 | 761-791 |
| 27 | Recreation services | 807-809 | 800-802 |
| 28 | Health | 828-848 | 812-840 |
| 29 | Legal services | 849 | 841 |
| 30 | Education | 857-868 | 842-860 |
| 31 | Religious and welfare services | 877-879 | 861-871, 880 |
| 32 | Other services | 869, 887-897 | 872, 881-892 |
| 33 | Public administration | 907-937 | 907-937 |

**Coding of occupations in the GSS**

| | Occupation | GSS 1972 – 1988<br>1970 Census codes | GSS 1989 – 2006<br>1980 Census codes |
|---|---|---|---|
| 1 | Administrative and managerial | 1, 56, 201-246 | 3-37 |
| 2 | Engineers | 2, 6-26 | 43-63 |
| 3 | Math and computer scientists | 2-5, 34-36, 55 | 64-68 |
| 4 | Natural scientists | 42-54 | 69-83 |
| 5 | Health professionals | 61-73 | 84-89 |
| 6 | Health treatment occupations | 74-76 | 95-106 |
| 7 | Post-secondary teachers | 102-140 | 113-154 |
| 8 | Teachers, exc. post-secondary | 141-145 | 155-159 |
| 9 | Counsellors, librarians, archivists | 32-33, 174 | 163-165 |
| 10 | Social scientists, urban planners | 91-96 | 166-173 |
| 11 | Social and religious workers | 86-90, 100, 101 | 174-177 |
| 12 | Lawyers and judges | 30-31 | 178-179 |
| 13 | Writers, artists, athletes | 175-194 | 183-199 |
| 14 | Technicians and support occupations | 80-85, 150-173 | 203-235 |
| 15 | Sales occupations | 260-296 | 243-285 |
| 16 | Clerical and admin. support occupations | 301-396 | 303-389 |
| 17 | Private household workers | 980-986 | 403-407 |
| 18 | Protective services workers | 960-976 | 413-427 |
| 19 | Service workers, exc. 17 and 18 | 901-954 | 433-469 |
| 20 | Farming occupations | 801-846 | 473-499 |
| 21 | Crafts and repair workers | 401-586 | 503-699 |
| 22 | Operators and laborers | 601-796 | 703-889 |

**Coding of industries in the NHIS**

| | | NHIS 1974 - 1994 | | NHIS 1995-2003 | NHIS 2004-2009 |
|---|---|---|---|---|---|
| | Sector | 1970 Census codes | 1980 Census codes | | |
| 1 | Agriculture, forestry, fishery | 17-28 | 10-31 | 1, 2 | 1-5 |
| 2 | Mining | 47-57 | 40-50 | 10 | 6-8 |
| 3 | Construction | 67-77 | 60 | 20 | 10 |
| 4 | Lumber, wood, furniture, stone, clay, glass | 107-138 | 230-262 | 40 | 17, 30 |
| 5 | Metal | 139-169 | 270-301 | 41 | 24, 25 |
| 6 | Machinery, exc. electrical | 177-198, 258 | 310-339 | 43, 42 | 26 |
| 7 | Electrical machinery | 199-209 | 340-350 | 44 | 27, 28 |
| 8 | Transportation equipment | 219-238 | 351-370 | 45 | 29 |
| 9 | Food and tobacco | 268-299 | 100-130 | 30 | 11, 12 |
| 10 | Textile, apparel, leather | 307-327, 388-397 | 132-152,220-228 | 31 | 13-16 |
| 11 | Printing and paper | 328-339 | 160-175 | 32 | 18, 19, 50 |
| 12 | Chemicals, petroleum, rubber | 347-369, 377-387 | 180-219 | 33 | 20-22 |
| 13 | Other manufacturing | 239-259, 398 | 259,371-399 | 46, 34 | 31, 23 |
| 14 | Transportation | 407-429 | 400-439 | 50-52 | 47-49 |
| 15 | Communication | 447-449 | 440-449 | 53 | 52-53, 51 |
| 16 | Utilities | 467-479 | 450-499 | 54 | 9 |
| 17 | Wholesale | 507-588 | 500-579 | 60 | 32-34 |
| 18 | Retail | 607-698 | 580-699 | 61-65 | 35-46 |
| 19 | FIRE | 707-718 | 700-719 | 70-71 | 54-60 |
| 20 | Business services | 727-748 | 720-750 | 75 | 62 |
| 21 | Repair services | 749-759 | 751-760 | 76 | 74 |
| 22 | Personal services | 769-798 | 761-792 | 77-78 | 75 |
| 23 | Recreation and entertainment | 807-809 | 800-811 | 79 | 69-71 |
| 24 | Health | 828-848 | 812-840 | 80-81 | 65-68 |
| 25 | Educational | 857-868 | 842-861 | 82-83 | 64 |
| 26 | Religious and welfare services | 877-879 | 862-871,880 | 84 | 76 |
| 27 | Other services, inc. legal | 869, 887-897, 849 | 841,872-877,881-899 | 85 | 72, 73, 77, 61, 63 |
| 28 | Public administration | 907-937 | 900-990 | 90 | 78 |

**Coding of occupations in the NHIS**

| | | NHIS 1974 - 1994 | | NHIS 1995-2003 | NHIS 2004-2009 |
|---|---|---|---|---|---|
| | Occupations | 1970 Census code | 1980 Census code | | |
| 1 | Administrative and managerial | 1,200-249, 56 | 3-39 | 1-3 | 1-5 |
| 2 | Engineers | 2-27,34-55,57-58 | 42-83 | 4-6 | 7-10,12-13 |
| 3 | Health professionals | 60-73 | 84-89 | 7 | 29 |
| 4 | Health treatment occs | 74-79 | 90-106 | 8 | 32-34 |
| 5 | Teachers | 32-33,102-149,174 | 110-165 | 9 | 20-23 |
| 6 | Other professional occs | 30-31,86-101 | 166-179 | 11 | 14,16-19 |
| 7 | Writers, artists, athletes | 175-199 | 180-199 | 10 | 25-28 |
| 8 | Technicians and support | 80-85,150-173 | 200-239 | 12-13 | 11,15,30-31 |
| 9 | Sales occupations | 250-297 | 240-299 | 14-16 | 53-57 |
| 10 | Clearical and admin support | 301-399 | 300-399 | 17-21 | 58-64 |
| 11 | Services: protective | 950-989 | 413-432 | 23-24 | 35-38 |
| 12 | Services: food | 910-919 | 433-444 | 25 | 39-42 |
| 13 | Services: cleaning, private hh | 900-909,980-986 | 401-410,448-455 | 22,27 | 43-45 |
| 14 | Services: personal | 931-959 | 456-469 | 28 | 46-52 |
| 15 | Services: health | 921-930 | 445-447 | 26 | 32-34 |
| 16 | Farming | 801-846 | 470-499 | 29-31 | 65-68 |
| 17 | Crafts and repair | 400-599 | 500-699 | 32-34 | 74-77 |
| 18 | Operators and laborers | 600-798,848-899 | 703-899 | 35-41 | 78-93 |

**Table 1**
**Regressions of Happiness on ln of Family Income for Men**
**General Social Survey, 1972 – 2006**
(Standard errors in parentheses)

| Estimation method | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| OLS | 0.172 | 0.128 | 0.124 | 0.097 | 0.171 | 0.126 | 0.122 | 0.096 |
| | (0.015) | (0.015) | (0.016) | (0.015) | (0.015) | (0.015) | (0.016) | (0.015) |
| 2SLS | 0.266 | 0.187 | 0.293 | 0.303 | 0.265 | 0.182 | 0.295 | 0.305 |
| | (0.075) | (0.078) | (0.095) | (0.092) | (0.075) | (0.078) | (0.095) | (0.092) |
| IJIVE | 0.276 | 0.192 | 0.321 | 0.335 | 0.275 | 0.187 | 0.323 | 0.338 |
| | (0.082) | (0.087) | (0.111) | (0.107) | (0.083) | (0.087) | (0.111) | (0.108) |
| First stage *F*-statistic | 11.77 | 10.88 | 7.06 | 7.04 | 11.72 | 10.77 | 7.08 | 7.06 |
| Baseline controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 4 marital status dummies | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ |
| 21 occupation dummies | | | ✓ | ✓ | | | ✓ | ✓ |
| 4 job satisfaction dummies | | | | ✓ | | | | ✓ |
| Mother's education | | | | · | ✓ | ✓ | ✓ | ✓ |

Weighted regressions using GSS sampling weight. The coefficient on ln(family income) is displayed. Baseline controls are age, age squared, dummies for black and other non-white race, eight education dummies, and 25 year dummies. Instruments are 32 industry dummies. Number of observations is 10,547. Heteroskedasticity robust standard errors in parentheses.

**Table 2**
**Regressions of Happiness and Mother's Education on ln of Family Income for Men**
**General Social Survey, 1972 – 2006**
(Standard errors in parentheses)

| | Dependent Variable | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Happiness | | | | Mother's education | | | |
| Estimation method | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| OLS | 0.172 | 0.128 | 0.124 | 0.097 | 0.166 | 0.220 | 0.173 | 0.165 |
| | (0.015) | (0.015) | (0.016) | (0.015) | (0.047) | (0.048) | (0.049) | (0.048) |
| 2SLS | 0.266 | 0.187 | 0.293 | 0.303 | 0.266 | 0.479 | 0.050 | 0.054 |
| | (0.075) | (0.078) | (0.095) | (0.092) | (0.244) | (0.257) | (0.306) | (0.307) |
| IJIVE | 0.276 | 0.192 | 0.321 | 0.335 | 0.282 | 0.515 | 0.040 | 0.047 |
| | (0.082) | (0.087) | (0.111) | (0.107) | (0.268) | (0.286) | (0.355) | (0.356) |
| First stage *F*-statistic | 11.77 | 10.88 | 7.06 | 7.04 | 11.77 | 10.88 | 7.06 | 7.04 |
| Baseline controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 4 marital status dummies | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ |
| 21 occupation dummies | | | ✓ | ✓ | | | ✓ | ✓ |
| 4 job satisfaction dummies | | | | ✓ | | | | ✓ |

Weighted regressions using GSS sampling weight. The coefficient on ln(family income) is displayed. Baseline controls are age, age squared, dummies for black and other non-white race, eight education dummies, and 25 year dummies. Instruments are 32 industry dummies. Number of observations is 10,547. Heteroskedasticity robust standard errors in parentheses.

**Table 3**
**Regressions of Happiness on ln of Family Income for Married Men and Women**
**General Social Survey, 1972 – 2006**
(Standard errors in parentheses)

| | Dependent Variable | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Happiness | | | | | Mother's education | | | | |
| | Sample | | | | | Sample | | | | |
| | Husbands | Wives | Wives, not working | Wives, working | | Husbands | Wives | Wives, not working | Wives, working | |
| Estimation method | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| OLS | 0.189 | 0.105 | 0.114 | 0.102 | 0.103 | 0.387 | 0.314 | 0.391 | 0.245 | 0.217 |
| | (0.023) | (0.019) | (0.027) | (0.026) | (0.027) | (0.072) | (0.060) | (0.085) | (0.088) | (0.091) |
| 2SLS | 0.180 | 0.042 | 0.127 | -0.010 | 0.010 | 0.654 | 0.777 | 0.564 | 0.537 | 0.431 |
| | (0.109) | (0.080) | (0.103) | (0.107) | (0.122) | (0.358) | (0.258) | (0.327) | (0.352) | (0.401) |
| IJIVE | 0.179 | 0.036 | 0.126 | -0.024 | -0.003 | 0.688 | 0.824 | 0.605 | 0.587 | 0.482 |
| | (0.125) | (0.087) | (0.123) | (0.124) | (0.146) | (0.407) | (0.280) | (0.391) | (0.404) | (0.478) |
| First stage *F*-statistic | 8.85 | 11.75 | 5.78 | 7.54 | 6.03 | 8.85 | 11.75 | 5.78 | 7.54 | 6.03 |
| Baseline controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 32 industry dummies (wives) | | | | | ✓ | | | | | ✓ |
| Number of observations | 6,672 | 7,540 | 3,062 | 4,478 | 4,478 | 6,672 | 7,540 | 3,062 | 4,478 | 4,478 |

Weighted regressions using GSS sampling weight. The coefficient on ln(family income) is displayed. Baseline controls are age, age squared, dummies for black and other non-white race, eight education dummies, and 25 year dummies. Instruments are 32 industry dummies for husband's industry affiliation. Heteroskedasticity robust standard errors in parentheses.

**Table 4**
**Comparison with Luttmer (2005)**
**Men, General Social Survey, 1972 – 2006**
(Standard errors in parentheses)

| | Dependent Variable | | | | | |
|---|---|---|---|---|---|---|
| | Happiness | | | Mother's education | | |
| Estimation method | (1) | (2) | (3) | (4) | (5) | (6) |
| OLS | 0.172 | 0.172 | 0.172 | 0.166 | 0.166 | 0.166 |
| | (0.015) | (0.015) | (0.015) | (0.047) | (0.047) | (0.047) |
| 2SLS | 0.266 | 0.328 | 0.283 | 0.266 | 1.238 | 0.273 |
| | (0.075) | (0.072) | (0.048) | (0.244) | (0.239) | (0.143) |
| IJIVE | 0.276 | 0.344 | 0.339 | 0.282 | 1.326 | 0.337 |
| | (0.082) | (0.077) | (0.071) | (0.268) | (0.256) | (0.213) |
| First stage *F*-statistic | 11.77 | 20.64 | 3.19 | 11.77 | 20.64 | 3.19 |
| Instruments | Industry | Occupation | Ind*Occ | Industry | Occupation | Ind*Occ |

Weighted regressions of happiness on ln(family income) using GSS sampling weight. The coefficient on ln(family income) is displayed. All regressions include controls for are age, age squared, dummies for black and other race, eight education dummies, and 25 year dummies. Instruments are 32 industry, 21 occupation dummies, or their interactions. Number of observations is 10,547. Heteroskedasticity robust standard errors in parentheses.

**Table 5**
**Alternative Instruments Using Employer Differentials**
**Men, General Social Survey, 1972 – 2006**
(Standard errors in parentheses)

| | Dependent variable | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Happiness | | | | Mother's education | | | |
| | Instruments | | | | Instruments | | | |
| | Union status | Firm size | Union and firm size | Union and industry | Union status | Firm size | Union and firm size | Union and industry |
| Estimation method | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| OLS | 0.175 | 0.173 | 0.161 | 0.175 | 0.116 | 0.052 | 0.067 | 0.116 |
| | (0.018) | (0.022) | (0.026) | (0.018) | (0.056) | (0.073) | (0.091) | (0.056) |
| 2SLS | 0.242 | 0.208 | 0.321 | 0.274 | -0.785 | -0.435 | -0.656 | 0.169 |
| | (0.184) | (0.160) | (0.169) | (0.078) | (0.603) | (0.491) | (0.553) | (0.251) |
| IJIVE | 0.243 | 0.215 | 0.342 | 0.284 | -0.798 | -0.485 | -0.749 | 0.174 |
| | (0.187) | (0.175) | (0.190) | (0.086) | (0.611) | (0.538) | (0.616) | (0.277) |
| First stage *F*-statistic | 66.89 | 11.32 | 9.60 | 11.8 | 66.89 | 11.32 | 9.60 | 11.80 |
| Baseline controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Number of observations | 7,365 | 4,444 | 3,082 | 7,365 | 7,365 | 4,444 | 3,082 | 7,365 |

Weighted regressions of happiness on ln(family income) using GSS sampling weight. The coefficient on ln(family income) is displayed. Baseline controls are age, age squared, dummies for black and other non-white race, eight education dummies, and 25 year dummies. Instruments are a dummy for union status, six dummies for firm size categories, and/or 32 industry dummies. Heteroskedasticity robust standard errors in parentheses.

**Table 6**
**Regressions of Health and Height on ln of Family Income for Men**
**NHIS, 1974-2009**
(standard errors in parentheses)

| | Dependent variable | | | | | |
|---|---|---|---|---|---|---|
| | Health | | | Height | | |
| Estimation method | (1) | (2) | (3) | (4) | (5) | (6) |
| OLS | 0.135 | 0.130 | 0.124 | 0.770 | 0.781 | 0.731 |
| | (0.002) | (0.002) | (0.002) | (0.018) | (0.018) | (0.019) |
| 2SLS | 0.071 | 0.060 | 0.097 | 1.017 | 0.991 | 1.241 |
| | (0.010) | (0.010) | (0.012) | (0.088) | (0.091) | (0.114) |
| IJIVE | 0.071 | 0.060 | 0.097 | 1.018 | 0.992 | 1.242 |
| | (0.010) | (0.010) | (0.013) | (0.088) | (0.091) | (0.115) |
| First stage F-statistic | 493.3 | 478.0 | 333.7 | 493.3 | 478.0 | 333.7 |
| Baseline controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 4 marital status dummies | | ✓ | ✓ | | ✓ | ✓ |
| 17 occupation dummies | | | ✓ | | | ✓ |

Weighted regressions using NHIS sampling weight. The coefficient on ln(family income) is displayed. Baseline controls are age, age squared, dummies for non-white, six education dummies, three region dummies, and 25 year dummies. Instruments are 27 industry dummies. Number of observations is 458,601. Heteroskedasticity robust standard errors in parentheses.

**Table 7**
**Regressions of Health and Height on ln of Family Income for Men Aged 20-29**
**NHIS, 1974-2009**
(standard errors in parentheses)

| | Dependent variable | | | | | |
|---|---|---|---|---|---|---|
| | Health | | | Height | | |
| Estimation method | (1) | (2) | (3) | (4) | (5) | (6) |
| OLS | 0.095 | 0.094 | 0.093 | 0.606 | 0.607 | 0.572 |
| | (0.003) | (0.003) | (0.003) | (0.031) | (0.031) | (0.031) |
| 2SLS | 0.061 | 0.048 | 0.076 | 0.867 | 0.874 | 0.805 |
| | (0.016) | (0.016) | (0.020) | (0.155) | (0.156) | (0.197) |
| IJIVE | 0.061 | 0.047 | 0.076 | 0.869 | 0.876 | 0.808 |
| | (0.016) | (0.016) | (0.020) | (0.156) | (0.157) | (0.199) |
| First stage F-statistic | 114.2 | 113.9 | 67.98 | 114.2 | 113.9 | 67.98 |
| Baseline controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 4 marital status dummies | | ✓ | ✓ | | ✓ | ✓ |
| 17 occupation dummies | | | ✓ | | | ✓ |

Weighted regressions using NHIS sampling weight. The coefficient on ln(family income) is displayed. Baseline controls are age, age squared, dummies for non-white, six education dummies, three region dummies, and 25 year dummies. Instruments are 27 industry dummies. Number of observations is 121,344. Heteroskedasticity robust standard errors in parentheses.