Solutions to Problem Set 3 – KAN-CCMVV2401U

1 (a) We estimate the model from column (2) but with *KWW* in place of *IQ*. The coefficient on *educ* becomes about .058 (se ≈ .006), so this is similar to the estimate obtained with *IQ*, although slightly larger and more precisely estimated.

(b) When *KWW* and *IQ* are both used as proxies, the coefficient on *educ* becomes about .049 (se ≈ .007). Compared with the estimate when only *KWW* is used as a proxy, the return to education has fallen by almost a full percentage point.

(c) The *t* statistic on *IQ* is about 3.08 while that on *KWW* is about 2.07, so each is significant at the 5% level against a two-sided alternative. They are jointly very significant, with $F_{2,925} \approx 8.59$ and *p*-value ≈ .0002.

2 (a) The regression of log(*wage*) on *sibs* gives

$$\begin{aligned} \log(wage) &= 6.861 \quad -.0279\ sibs \\ &\quad (0.022) \quad (.0059) \end{aligned}$$

$$n = 935, \quad R^2 = .023.$$

This is a reduced form simple regression equation. It shows that, controlling for no other factors, one more sibling in the family is associated with monthly salary that is about 2.8% lower. The *t* statistic on *sibs* is about –4.72. Of course *sibs* can be correlated with many things that should have a bearing on wage including, as we already saw, years of education.

(b) It could be that older children are given priority for higher education, and families may hit budget constraints and may not be able to afford as much education for children born later. The simple regression of *educ* on *brthord* gives

$$\begin{aligned} educ &= 14.15 \quad -.283\ brthord \\ &\quad (0.13) \quad (.046) \end{aligned}$$

$$n = 852, \quad R^2 = .042.$$

(Note that *brthord* is missing for 83 observations.) The equation predicts that every one-unit increase in *brthord* reduces predicted education by about .28 years. In particular, the difference in predicted education for a first-born and fourth-born child is about .85 years.

(c) When *brthord* is used as an IV for *educ* in the simple wage equation we get

$$\begin{aligned} \log(wage) &= 5.03 \quad +.131\ educ \\ &\quad (0.43) \quad (.032) \end{aligned}$$

$$n = 852.$$

(The *R*-squared is negative.) This is much higher than the OLS estimate (.060) and even above the estimate when *sibs* is used as an IV for *educ* (.122). Because of missing data on *brthord*, we are using fewer observations than in the previous analyses.

(d) In the reduced form equation

$$educ = \pi_0 + \pi_1 sibs + \pi_2 brthord + v,$$

we need $\pi_2 \neq 0$ in order for the β_j to be identified. We take the null to be H_0: $\pi_2 = 0$, and look to reject H_0 at a small significance level. The regression of *educ* on *sibs* and *brthord* (using 852 observations) yields $\hat{\pi}_2 = -.153$ and $se(\hat{\pi}_2) = .057$. The t statistic is about –2.68, which rejects H0 fairly strongly. Therefore, the identification assumptions appears to hold.

(e) The equation estimated by IV is

$$\begin{array}{lccc} \log(wage) & = 4.94 & +.137\ educ & +.0021\ sibs \\ & (1.06) & (.075) & (.0174) \end{array}$$

$n = 852.$

The standard error on $\hat{\beta}_{educ}$ is much larger than we obtained in part (c). The 95% CI for β_{educ} is roughly –.010 to .284, which is very wide and includes the value zero. The standard error of $\hat{\beta}_{sibs}$ is very large relative to the coefficient estimate, rendering *sibs* very insignificant.

(f) Letting $educ_i$ be the first-stage fitted values, the correlation between $educ_i$ and $sibs_i$ is about –.930, which is a very strong negative correlation. This means that, for the purposes of using IV, multicollinearity is a serious problem here, and is not allowing us to estimate β_{educ} with much precision.

3 (a) The OLS results are

$$\begin{array}{lccccc} pira = -.198 & +.054\ p401k & +.0087\ inc & -.000023\ inc^2 & -.0016\ age & +.00012\ age^2 \\ (.069) & (.010) & (.0005) & (.000004) & (.0033) & (.00004) \end{array}$$

$n = 9{,}275,\ R^2 = .180$

The coefficient on *p401k* implies that participation in a 401(k) plan is associated with a .054 higher probability of having an individual retirement account, holding income and age fixed.

(b) While the regression in part (a) controls for income and age, it does not account for the fact that different people have different taste for savings, even within

given income and age categories. People that tend to be savers will tend to have both a 401(k) plan as well as an IRA. (This means that the error term, *u*, is positively correlated with *p401k*.) What we would like to know is, for a given person, if that person participates in a 401(k) does it make it less likely or more likely that the person also has an IRA. This ceteris paribus question is difficult to answer by OLS without many more controls for the taste for saving.

(c) First, we need *e*401*k* to be partially correlated with *p*401*k*; not surprisingly, this is not an issue, as being eligible for a 401(k) plan is, by definition, necessary for participation. (The regression in part (d) verifies that they are strongly positively correlated.) The more difficult issue is whether *e*401*k* can be taken as exogenous in the structural model. In other words, is being *eligible* for a 401(k) correlated with unobserved taste for saving? If we think workers that like to save for retirement will match up with employers that provide vehicles for retirement saving, then *u* and *e*401*k* would be positively correlated. Certainly we think that *e*401*k* is less correlated with *u* than is *p*401*k*. But remember, this alone is not enough to ensure that the IV estimator has less asymptotic bias than the OLS estimator; see lecture slides page 22.

(d) The reduced form equation, estimated by OLS but with heteroskedasticity-robust standard errors, is

$$p401k = \underset{(.046)}{.059} + \underset{(.008)}{.689}\, e401k + \underset{(.0003)}{.0011}\, inc - \underset{(.0000027)}{.0000018}\, inc^2 - \underset{(.0022)}{.0047}\, age + \underset{(.000026)}{.000052}\, age^2$$

$n = 9{,}275,\ R^2 = .596$

The *t* statistic on *e*401*k* is over 85, and its coefficient estimate implies that, holding income and age fixed, eligibility in a 401(k) plan increases the probability of participation in a 401(k) by .69. Clearly, *e*401*k* passes one of the two requirements as an IV for *p*401*k*.

(e) When *e*401*k* is used as an IV for *p*401*k* we get the following, with heteroskedasticity-robust standard errors:

$$pira = \underset{(.065)}{-.207} + \underset{(.013)}{.021}\, p401k + \underset{(.0005)}{.0090}\, inc - \underset{(.000004)}{.000024}\, inc^2 - \underset{(.0032)}{.0011}\, age + \underset{(.00004)}{.00011}\, age^2$$

$n = 9{,}275,\ R^2 = .180$

The IV estimate of β_{p401k} is less than half as large as the OLS estimate, and the IV estimate has a *t* statistic roughly equal to 1.62. The reduction in $\hat{\beta}_{p401k}$ is what we expect given the unobserved taste for saving argument made in part (b). But we still do not estimate a tradeoff between participating in a 401(k) plan and participating in an IRA. This conclusion has prompted some in the literature to claim that 401(k) saving is additional saving; it does not simply crowd out saving in other plans.

(f) After obtaining the reduced form residuals from part (d), say $\hat{v}_i$, we add these to the structural equation and run OLS. The coefficient on $\hat{v}_i$ is .075 with a heteroskedasticity-robust $t = 3.92$. Therefore, there is strong evidence that *p401k* is endogenous in the structural equation (assuming, of course, that the IV, *e*401*k*, is exogenous).