# Homework 3

ECON 7023: Econometrics II
Maghfira Ramadhani
February 21, 2022

Spring 2023

## Chapter 5

### Problem 5.4

Use the data in CARD.RAW for this problem

a. Estimate a log($wage$) equation by OLS with $educ, exper, exper^2, black, south, smsa, reg661$ through $reg668$, and $smsa66$ as explanatory variables. Compare your results with Table 2, Column (2) in Card (1995).
Answer:

Figure 1: Result from Card (1995)

Table 2: Estimated Regression Models for Log Hourly Earnings

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| 1. Education | 0.074<br>(0.004) | 0.075<br>(0.003) | 0.073<br>(0.004) | 0.074<br>(0.004) | 0.073<br>(0.004) |
| 2. Experience | 0.084<br>(0.007) | 0.085<br>(0.007) | 0.085<br>(0.007) | 0.085<br>(0.007) | 0.085<br>(0.007) |
| 3. Experience-Squared /100 | -0.224<br>(0.032) | -0.229<br>(0.032) | -0.230<br>(0.032) | -0.226<br>(0.032) | -0.229<br>(0.032) |
| 4. Black Indicator | -0.190<br>(0.017) | -0.199<br>(0.018) | -0.194<br>(0.019) | -0.194<br>(0.019) | -0.189<br>(0.019) |
| 5. Live in South | -0.125<br>(0.015) | -0.148<br>(0.026) | -0.146<br>(0.026) | -0.145<br>(0.026) | -0.146<br>(0.026) |
| 6. Live in SMSA | 0.161<br>(0.015) | 0.136<br>(0.020) | 0.136<br>(0.020) | 0.137<br>(0.020) | 0.138<br>(0.020) |
| 7. Region in 1966 (8 indicators) | no | yes | yes | yes | yes |
| 8. Live in SMSA in 1966 | no | yes | yes | yes | yes |
| 9. Parental Education[a] (main effects) | no | no | yes | yes | yes |
| 10. Interacted Parental Education Classes[b] | no | no | no | yes | yes |
| 11. Family Structure[c] (2 indicators) | no | no | no | no | yes |
| 12. R-squared | 0.291 | 0.300 | 0.301 | 0.303 | 0.304 |
| 13. P-value for family background effects | -- | -- | 0.235 | 0.462 | 0.165 |

Notes: Standard errors in parentheses. Sample size is 3010. The dependent variable in all cases is the log of hourly wages in 1976. The mean and standard deviation of the dependent variable are 6.262 and 0.444.

[a]Variables representing years of education of mother and father, plus indicators for missing mother's or father's education.

[b]Indicators for 8 classes of mother's and father's education.

[c]Indicators for father and mother present at age 14, and single mother at age 14.

From Figure 1, we know that the result in Column (2) the return to education is 7.5% with standard error of 0.03% and is statistically significant. Comparing to the result that we have in Table 1 from estimating log($wage$) by OLS with $educ, exper, exper^2, black, south, smsa, reg661$ through $reg668$, and $smsa66$ as explanatory variables, we also get 7.5% return of education with standard error of 0.03% and is statistically significant. Surprisingly, we get the same coefficient and standard error with those in Card (1995) even though the model specifications are different.

b. Estimate a reduced form equation for $educ$ containing all explanatory variables from part a and the dummy variable $nearc4$. Do $educ$ and $nearc4$ have a practically and statistically significant partial correlation? (See also Table 3, Column (1) in Card (1995).)
Answer:

Figure 2: Result from Card (1995)

Table 3: Reduced Form and Structural Estimates of Education and Earnings Models

| | Reduced Form Models: Education | | Earnings | | Structural Models of Earnings | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| **A: Treat Experience and Experience Squared as Exogenous** | | | | | | |
| 1. Live Near College in 1966 | 0.320 (0.088) | 0.322 (0.083) | 0.042 (0.018) | 0.045 (0.018) | -- | -- |
| 2. Education | -- | -- | -- | -- | 0.132 (0.055) | 0.140 (0.055) |
| 3. Family Background Variables[a] | no | yes | no | yes | no | yes |
| **B: Treat Experience and Experience Squared as Endogenous[b/]** | | | | | | |
| 4. Live Near College in 1966 | 0.382 (0.114) | 0.365 (0.105) | 0.047 (0.019) | 0.048 (0.019) | -- | -- |
| 5. Education | -- | -- | -- | -- | 0.122 (0.046) | 0.132 (0.049) |
| 6. Family Background Variables[a] | no | yes | no | yes | no | yes |

Notes: standard errors in parentheses. Sample size is 3010. The dependent variable in columns 1 and 2 is completed education in 1976 (mean and standard deviation: 13.263 and 2.677). The dependent variable in columns 3-6 is the log of hourly wages in 1976 (mean and standard deviation: 6.262 and 0.444). All models include a black indicator, indicators for southern residence and residence in an SMSA in 1976, indicators for region in 1966 and living in an SMSA in 1966, as well as experience and experience squared.

[a] 14 variables representing mother's and father's education, indicators for missing father's or mother's education, interactions of mother's and father's education, and dummies for family structure at age 14.

[b] In these models, experience is treated as endogenous. Instruments for experience and experience squared are age and age squared.

Table 2 Column (1) shows the reduced from estimates for $educ$ containing all explanatory variables from part a and the dummy variable $nearc4$. Our coefficient of interest is on $nearc4$ which is a dummy variable for someone living near a 4-year college. The estimates show that $educ$ and $nearc4$ are correlated with coefficient of 0.320 with a standard error of 0.0088. The result is statistically very significant. Thus the variable $nearc4$ satisfy the relevance condition for a good IV. Comparing with the result in Card (1995) as shown in Figure 2, the result that we get is the same as in Column (1).

c. Estimate the log($wage$) equation by IV, using $nearc4$ as an instrument for $educ$. Compare the 95 percent confidence interval for the return to education with that obtained from part a. (See also Table 3, Column (5) in Card (1995).)
Answer:

Table 1: Regression result for Problem 5.4.a.

| | (1) |
|---|---|
| years of schooling, 1976 | 0.075*** |
| | (0.003) |
| age - educ - 6 | 0.085*** |
| | (0.007) |
| $\text{exper}^2$ | -0.002*** |
| | (0.000) |
| =1 if black | -0.199*** |
| | (0.018) |
| =1 if in south, 1976 | -0.148*** |
| | (0.026) |
| =1 in in SMSA, 1976 | 0.136*** |
| | (0.020) |
| regional dummy, 1966 | -0.119*** |
| | (0.039) |
| reg662 | -0.022 |
| | (0.028) |
| reg663 | 0.026 |
| | (0.027) |
| reg664 | -0.063* |
| | (0.036) |
| reg665 | 0.009 |
| | (0.036) |
| reg666 | 0.022 |
| | (0.040) |
| reg667 | -0.001 |
| | (0.039) |
| reg668 | -0.175*** |
| | (0.046) |
| =1 if in SMSA, 1966 | 0.026 |
| | (0.019) |
| Constant | 4.739*** |
| | (0.072) |
| Observations | 3010 |

Standard errors in parentheses

Data: CARD.DTA

Wooldridge (2011)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Regression results for (1) Problem 5.4.b. and (2) Problem 5.4.d

| | (1) | (2) |
|---|---|---|
| age - educ - 6 | -0.413*** | -0.412*** |
| | (0.034) | (0.034) |
| $\text{exper}^2$ | 0.001 | 0.001 |
| | (0.002) | (0.002) |
| =1 if black | -0.936*** | -0.945*** |
| | (0.094) | (0.094) |
| =1 if in south, 1976 | -0.052 | -0.042 |
| | (0.135) | (0.136) |
| =1 in in SMSA, 1976 | 0.402*** | 0.401*** |
| | (0.105) | (0.105) |
| regional dummy, 1966 | -0.210 | -0.169 |
| | (0.202) | (0.204) |
| reg662 | -0.289** | -0.269* |
| | (0.147) | (0.148) |
| reg663 | -0.238* | -0.190 |
| | (0.143) | (0.146) |
| reg664 | -0.093 | -0.038 |
| | (0.186) | (0.189) |
| reg665 | -0.483** | -0.437** |
| | (0.188) | (0.190) |
| reg666 | -0.513** | -0.502** |
| | (0.210) | (0.210) |
| reg667 | -0.427** | -0.378* |
| | (0.206) | (0.208) |
| reg668 | 0.314 | 0.382 |
| | (0.242) | (0.245) |
| =1 if in SMSA, 1966 | 0.025 | 0.000 |
| | (0.106) | (0.107) |
| =1 if near 4 yr college, 1966 | 0.320*** | 0.321*** |
| | (0.088) | (0.088) |
| =1 if near 2 yr college, 1966 | | 0.123 |
| | | (0.077) |
| Constant | 16.849*** | 16.773*** |
| | (0.211) | (0.216) |
| Observations | 3010 | 3010 |

Standard errors in parentheses

Data: CARD.DTA

Wooldridge (2011)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3 Column (1) shows the log(*wage*) equation by IV, using *nearc*4 as an instrument for *educ*. The estimated return to education that we get is 13.2% with standard error of 5.5%. The 95 percent confidence in the 2SLS estimation is 2.37% to 23.93% while in the OLS estimation it is 6.78% to 8.15%. In this case, in the OLS since there is indication of endogeneity problem, the estimator may be inconsistent even though the confidence interval is smaller but we still can not believe it directly.

OLS Regression

```
. reg lwage educ exper expersq black south smsa reg661-reg668 smsa66
```

| Source | SS | df | MS | | | |
|---|---|---|---|---|---|---|
| | | | | Number of obs | = | 3,010 |
| | | | | F(15, 2994) | = | 85.48 |
| Model | 177.695591 | 15 | 11.8463727 | Prob > F | = | 0.0000 |
| Residual | 414.946054 | 2,994 | .138592536 | R-squared | = | 0.2998 |
| | | | | Adj R-squared | = | 0.2963 |
| Total | 592.641645 | 3,009 | .196956346 | Root MSE | = | .37228 |

| lwage | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| educ | .0746933 | .0034983 | 21.35 | 0.000 | .0678339 | .0815527 |
| exper | .084832 | .0066242 | 12.81 | 0.000 | .0718435 | .0978205 |
| expersq | -.002287 | .0003166 | -7.22 | 0.000 | -.0029079 | -.0016662 |
| black | -.1990123 | .0182483 | -10.91 | 0.000 | -.2347927 | -.1632318 |
| south | -.147955 | .0259799 | -5.69 | 0.000 | -.1988952 | -.0970148 |
| smsa | .1363845 | .0201005 | 6.79 | 0.000 | .0969724 | .1757967 |
| reg661 | -.1185698 | .0388301 | -3.05 | 0.002 | -.194706 | -.0424335 |
| reg662 | -.0222026 | .0282575 | -0.79 | 0.432 | -.0776088 | .0332036 |
| reg663 | .0259703 | .0273644 | 0.95 | 0.343 | -.0276846 | .0796251 |
| reg664 | -.0634942 | .0356803 | -1.78 | 0.075 | -.1334546 | .0064662 |
| reg665 | .0094551 | .0361174 | 0.26 | 0.794 | -.0613623 | .0802725 |
| reg666 | .0219476 | .0400984 | 0.55 | 0.584 | -.0566755 | .1005708 |
| reg667 | -.0005887 | .0393793 | -0.01 | 0.988 | -.077802 | .0766245 |
| reg668 | -.1750058 | .0463394 | -3.78 | 0.000 | -.265866 | -.0841456 |
| smsa66 | .0262417 | .0194477 | 1.35 | 0.177 | -.0118905 | .0643739 |
| _cons | 4.739377 | .0715282 | 66.26 | 0.000 | 4.599127 | 4.879626 |

2SLS Regression using *nearc*4 as instrument for *educ*

```
. ivreg lwage exper expersq black south smsa reg661-reg668 smsa66 (educ = nearc4)
Instrumental variables 2SLS regression
```

| Source | SS | df | MS | | | |
|---|---|---|---|---|---|---|
| | | | | Number of obs | = | 3,010 |
| | | | | F(15, 2994) | = | 51.01 |
| Model | 141.146813 | 15 | 9.40978752 | Prob > F | = | 0.0000 |
| Residual | 451.494832 | 2,994 | .150799877 | R-squared | = | 0.2382 |
| | | | | Adj R-squared | = | 0.2343 |
| Total | 592.641645 | 3,009 | .196956346 | Root MSE | = | .38833 |

| lwage | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| educ | .1315038 | .0549637 | 2.39 | 0.017 | .0237335 | .2392742 |
| exper | .1082711 | .0236586 | 4.58 | 0.000 | .0618824 | .1546598 |
| expersq | -.0023349 | .0003335 | -7.00 | 0.000 | -.0029888 | -.001681 |
| black | -.1467757 | .0538999 | -2.72 | 0.007 | -.2524603 | -.0410912 |
| south | -.1446715 | .0272846 | -5.30 | 0.000 | -.19817 | -.091173 |
| smsa | .1118083 | .031662 | 3.53 | 0.000 | .0497269 | .1738898 |
| reg661 | -.1078142 | .0418137 | -2.58 | 0.010 | -.1898007 | -.0258278 |
| reg662 | -.0070465 | .0329073 | -0.21 | 0.830 | -.0715696 | .0574767 |
| reg663 | .0404445 | .0317806 | 1.27 | 0.203 | -.0218694 | .1027585 |
| reg664 | -.0579172 | .0376059 | -1.54 | 0.124 | -.1316532 | .0158189 |
| reg665 | .0384577 | .0469387 | 0.82 | 0.413 | -.0535777 | .130493 |
| reg666 | .0550887 | .0526597 | 1.05 | 0.296 | -.0481642 | .1583416 |
| reg667 | .026758 | .0488287 | 0.55 | 0.584 | -.0689832 | .1224992 |
| reg668 | -.1908912 | .0507113 | -3.76 | 0.000 | -.2903238 | -.0914586 |
| smsa66 | .0185311 | .0216086 | 0.86 | 0.391 | -.0238381 | .0609003 |
| _cons | 3.773965 | .934947 | 4.04 | 0.000 | 1.940762 | 5.607169 |

```
Instrumented: educ
 Instruments: exper expersq black south smsa reg661 reg662 reg663 reg664
              reg665 reg666 reg667 reg668 smsa66 nearc4
```

Table 3: Regression results for (1) Problem 5.4.c. and (2) Problem 5.4.d.

| | (1) | (2) |
|---|---|---|
| years of schooling, 1976 | 0.132** | 0.157*** |
| | (0.055) | (0.053) |
| age - educ - 6 | 0.108*** | 0.119*** |
| | (0.024) | (0.023) |
| $\text{exper}^2$ | -0.002*** | -0.002*** |
| | (0.000) | (0.000) |
| =1 if black | -0.147*** | -0.123** |
| | (0.054) | (0.052) |
| =1 if in south, 1976 | -0.145*** | -0.143*** |
| | (0.027) | (0.028) |
| =1 in in SMSA, 1976 | 0.112*** | 0.101*** |
| | (0.032) | (0.032) |
| regional dummy, 1966 | -0.108*** | -0.103** |
| | (0.042) | (0.043) |
| reg662 | -0.007 | -0.000 |
| | (0.033) | (0.034) |
| reg663 | 0.040 | 0.047 |
| | (0.032) | (0.033) |
| reg664 | -0.058 | -0.055 |
| | (0.038) | (0.039) |
| reg665 | 0.038 | 0.052 |
| | (0.047) | (0.048) |
| reg666 | 0.055 | 0.070 |
| | (0.053) | (0.053) |
| reg667 | 0.027 | 0.039 |
| | (0.049) | (0.050) |
| reg668 | -0.191*** | -0.198*** |
| | (0.051) | (0.053) |
| =1 if in SMSA, 1966 | 0.019 | 0.015 |
| | (0.022) | (0.022) |
| Constant | 3.774*** | 3.340*** |
| | (0.935) | (0.895) |
| Observations | 3010 | 3010 |

Standard errors in parentheses

Data: CARD.DTA

Wooldridge (2011)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

d. Now use $nearc2$ along with $nearc4$ as instruments for $educ$. First estimate the reduced form for $educ$, and comment on whether $nearc4$ is more strongly related to $educ$. How do the 2SLS estimates compare with the earlier estimates?
Answer:
Table 2 Column (2) shows the reduced form estimates when adding $nearc2$ and $nearc4$ together. The coefficient for $nearc2$ is 0.123 with standard error of 0.077 which is not statistically significant. Compared to previous result, the coefficient for $nearc4$ is now increased to 0.321 with relatively similar standard error. Table 3 Column (2) shows the 2SLS estimates with $nearc2$ and $nearc4$ as instruments. The return to education increase to about 15.7% with increased significance level compared to previous result when using only $nearc4$.

e. For as subset of the men in the sample, IQ score is available. Regress $iq$ on $nearc4$. Is IQ score uncorrelated with $nearc4$?
Answer:
Table 4 Column (1) show the regression result, it shows that IQ score is correlated with $nearc4$.

Table 4: Regression results for (1) Problem 5.4.e. and (2) Problem 5.4.f.

| | (1) | (2) |
|---|---|---|
| =1 if near 4 yr college, 1966 | 2.596$^{***}$ | 0.868 |
| | (0.745) | (0.822) |
| =1 if in SMSA, 1966 | | 1.355$^{*}$ |
| | | (0.803) |
| regional dummy, 1966 | | 4.768$^{***}$ |
| | | (1.547) |
| reg662 | | 5.808$^{***}$ |
| | | (0.902) |
| reg669 | | 1.845 |
| | | (1.152) |
| Constant | 100.611$^{***}$ | 99.385$^{***}$ |
| | (0.627) | (0.702) |
| Observations | 2061 | 2061 |

Standard errors in parentheses
Data: CARD.DTA
Wooldridge (2011)
$^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

f. Now regress $iq$ on $nearc4$ along with $smsa66, reg661, reg662$, and $reg669$. Are $iq$ and $nearc4$ partially correlated? What do you conclude about the importance of controlling for the 1966 location and regional dummies in the $\log(wage)$ equation when using $nearc4$ as an IV for $educ$?
Answer:
Table 4 Column (2) show the regression result, now after controlling for $smsa66, reg661, reg662$, and $reg669$ the coefficient on $nearc4$ become insignificant, in other word the effect or correlation still exists and the same, but statistically dissapears. Thus, when using $nearc4$ as IV for educ, it is important to add control variables also in the $\log(wage)$ equation.

## Problem 5.7

Consider model (5.45) where $v$ has zero mean and is uncorrelated with $x_1, \ldots, x_K$ and $q$. The unobservable $q$ is thought to be correlated with at least some of the $x_j$. Assume without loss of generality that $\mathrm{E}(q) = 0$. You have a single indicator of $q$, written as $q_1 = \delta_1 q + a_1, \delta_1 \neq 0$, where $a_1$ has zero mean and is uncorrelated with each of $x_j, q$ and $v$. In addition, $z_1, z_2, \dots, z_M$ is a set of variables that are (1) redundant in the structural equation (5.45) and (2) uncorrelated with $a_1$.

a. Suggest an IV method for consistently estimating the $\beta_j$. Be sure to discuss what is needed for identification.
Answer:
Recall equation (5.45): $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \gamma q + v$. It is given that $q_1 = \delta_1 q + a_1 \Leftrightarrow q = \frac{1}{\delta_1} q_1 - \frac{1}{\delta_1} a_1$. Substitute it to (5.45) we will have

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \gamma \left( \frac{1}{\delta_1} q_1 - \frac{1}{\delta_1} a_1 \right) + v$$

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \frac{\gamma}{\delta_1} q_1 + v - \frac{\gamma}{\delta_1} a_1 \quad (1)$$

We also have $z_1, z_2, \dots, z_M$ is a set of variables that are redundant in the structural equation (5.45) and uncorrelated with $a_1$. Thus, we have $z_1, z_2, \dots, z_M$ are uncorrelated with the error, $v$. Also, $a_1$ is uncorrelated with each of $x_j, q$ and $v$. Thus, we have $x_j$ uncorrelated with $v - \frac{\gamma}{\delta_1} a_1$. These conditions satisfy the exlclusion conditions for IV. Therefore we can now estimate equation (1) to get consistent estimation for $\beta_j$ and $\frac{\gamma}{\delta_1}$ by 2SLS estimation using instruments $x_1, \dots, x_K, z_1, \dots, z_M$. Another identification that we need to show is that at least one of the instruments in $z_1, \dots, z_M$ appears in $q_1$ to satisfy the rank condition.

b. If equation (5.45) is a $\log(wage)$ equation, $q$ is ability, $q_1$ is IQ or some other test score, and $z_1, \dots, z_M$ are family of variables, such as parents' education and number of siblings, describe the economic assumption needed for consistency of the IV procedure in part a.
Answer:
Suppose $y$ is $\log(wage)$, $z_1, \dots, z_M$ are family of variables, such as parents' education and number of siblings. The first condition is for family background to be exogenous in equation (5.45), or that the family backgrounds variables are redundant in (5.45) after we control for ability using the indicator $q_1$. The second condition, the rank condition, we need the family backgrounds variable to be correlated with the indicator $q_1$ or $IQ$. It is common to say that family background will be partially correlated with ability.

c. Carry out this procedure using the data in NLS80.RAW. Include among the explanatory variables $exper, tenure, educ, married, south, urban$, and $black$. First use $IQ$ as $q_1$ and then $KWW$. Include in the $z_h$ the variables $meduc, feduc$, and $sibs$. Discuss the results.
Answer:
2SLS Regression using $meduc, feduc, sibs$ as instrument for $iq$

```
. ivreg lwage exper tenure educ married south urban black (iq = meduc feduc sibs)

Instrumental variables 2SLS regression

      Source |       SS           df       MS      Number of obs   =       722
-------------+----------------------------------   F(8, 713)       =     25.81
       Model |  19.6029198         8  2.45036497   Prob > F        =    0.0000
    Residual |  107.208996       713  .150363248   R-squared       =    0.1546
-------------+----------------------------------   Adj R-squared   =    0.1451
       Total |  126.811916       721  .175883378   Root MSE        =    .38777

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          iq |   .0154368   .0077077     2.00   0.046     .0003044    .0305692
       exper |   .0162185   .0040076     4.05   0.000     .0083503    .0240867
      tenure |   .0076754   .0030956     2.48   0.013     .0015979    .0137529
        educ |   .0161809   .0261982     0.62   0.537    -.035254     .0676158
     married |   .1901012   .0467592     4.07   0.000     .0982991    .2819033
       south |  -.047992    .0367425    -1.31   0.192    -.1201284    .0241444
       urban |   .1869376   .0327986     5.70   0.000     .1225442    .2513311
       black |   .0400269   .1138678     0.35   0.725    -.1835294    .2635832
       _cons |   4.471616    .468913     9.54   0.000        3.551    5.392231
------------------------------------------------------------------------------
Instrumented: iq
 Instruments: exper tenure educ married south urban black meduc feduc sibs
```

2SLS Regression using $meduc, feduc, sibs$ as instrument for $kww$

```
. ivreg lwage exper tenure educ married south urban black (kww = meduc feduc sibs)

Instrumental variables 2SLS regression

      Source |       SS           df       MS      Number of obs   =       722
-------------+----------------------------------   F(8, 713)       =     25.70
       Model |   19.820304         8    2.477538   Prob > F        =    0.0000
    Residual |  106.991612       713  .150058361   R-squared       =    0.1563
-------------+----------------------------------   Adj R-squared   =    0.1468
       Total |  126.811916       721  .175883378   Root MSE        =    .38737

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         kww |   .0249441   .0150576     1.66   0.098    -.0046184    .0545067
       exper |   .0068682   .0067471     1.02   0.309    -.0063783    .0201147
      tenure |   .0051145   .0037739     1.36   0.176    -.0022947    .0125238
        educ |   .0260808   .0255051     1.02   0.307    -.0239933    .0761549
     married |   .1605273   .0529759     3.03   0.003     .0565198    .2645347
       south |   -.091887   .0322147    -2.85   0.004    -.1551341   -.0286399
       urban |   .1484003   .0411598     3.61   0.000     .0675914    .2292093
       black |  -.0424452   .0893695    -0.47   0.635    -.2179041    .1330137
       _cons |   5.217818   .1627592    32.06   0.000     4.898273    5.537362
------------------------------------------------------------------------------
Instrumented: kww
 Instruments: exper tenure educ married south urban black meduc feduc sibs
```

Table 5: Regression results for Problem 5.7.c.

| | (1) | (2) |
|---|---|---|
| IQ score | 0.015** | |
| | (0.008) | |
| years of work experience | 0.016*** | 0.007 |
| | (0.004) | (0.007) |
| years with current employer | 0.008** | 0.005 |
| | (0.003) | (0.004) |
| years of education | 0.016 | 0.026 |
| | (0.026) | (0.026) |
| =1 if married | 0.190*** | 0.161*** |
| | (0.047) | (0.053) |
| =1 if live in south | -0.048 | -0.092*** |
| | (0.037) | (0.032) |
| =1 if live in SMSA | 0.187*** | 0.148*** |
| | (0.033) | (0.041) |
| =1 if black | 0.040 | -0.042 |
| | (0.114) | (0.089) |
| knowledge of world work score | | 0.025* |
| | | (0.015) |
| Constant | 4.472*** | 5.218*** |
| | (0.469) | (0.163) |
| Observations | 722 | 722 |

Standard errors in parentheses

Data: NLS80.DTA

Wooldridge (2011)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The regression results are shown in Table (5) in Column (1) we use $IQ$ and in Column (2) we use $KWW$ as indicator. The return to education in both estimates is statistically not significant. We may suspect that the family background in this case not satisfy the redundancy conditions mentioned in part b, or they may be correlated with $a_1$.

## Problem 5.11

A model with a single endogenous explanatory variable can be written as

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1, \quad \mathrm{E}(\mathbf{z}'u_1) = 0,$$

where $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$. Consider the following two-step method, intended to mimic 2SLS:

a. Regress $y_2$ on $\mathbf{z}_2$, and obtain the fitted values, $\tilde{y}_2$. (That is, $\mathbf{z}_1$ is omitted from the first-stage regression.)

b. Regress $y_1$ on $\mathbf{z}_1, \tilde{y}_2$ to obtain $\tilde{\boldsymbol{\delta}}_1$ and $\tilde{\alpha}_1$.

Show that $\tilde{\boldsymbol{\delta}}_1$ and $\tilde{\alpha}_1$ are generally inconsistent. When would $\tilde{\boldsymbol{\delta}}_1$ and $\tilde{\alpha}_1$ be consistent? (Hint: Let $y_2^0$ be the population linear projection of $y_2$ on $\mathbf{z}_2$, and let $a_2$ be the projection error: $y_2^0 = \mathbf{z}_2\boldsymbol{\lambda}_2 + a_2, \mathrm{E}(\mathbf{z}_2'a_2) = \mathbf{0}$. For simplicity, pretend that $\boldsymbol{\lambda}_2$ is known rather than estimated; that is, assume that $\tilde{y}_2$ is actually $y_2^0$. Then, write:

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2^0 + \alpha_1 a_2 + u_1$$

and check whether the composite error $\alpha_1 a_2 + u_1$ is uncorrelated with the explanatory variables.)
Answer:
Following the hint, let $y_2^0$ be the population linear projection of $y_2$ on $\mathbf{z}_2$, and let $a_2$ be the projection error: $y_2^0 = \mathbf{z}_2\boldsymbol{\lambda}_2 + a_2, \mathrm{E}(\mathbf{z}_2'a_2) = \mathbf{0}$. Assume that $\boldsymbol{\lambda}_2$ is known. Write $y_2 = y_2^0 + a_2$ and substitute it into the original equation, we have

$$\begin{aligned} & y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \\ \Leftrightarrow\ & y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(y_2^0 + a_2) + u_1 \\ \Leftrightarrow\ & y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2^0 + \underbrace{\alpha_1 a_2 + u_1}_{\text{composite error}}. \end{aligned}$$

The conditions for consistent estimation is that $\alpha_1 a_2 + u_1$ be orthogonal to $\mathbf{z}_1$ and $y_2^0$. It is given that $E(\mathbf{z}'u_1) = 0$ which also means $E(\mathbf{z}_1'u_1) = 0, E(\mathbf{z}_2'u_1) = 0$. From the linear projection by construction we have $E(\mathbf{z}_2'a_2) = 0$. However, since $\mathbf{z}_1$ is omitted in step a, we will have $E(\mathbf{z}_2'a_2) \neq 0$. Thus our OLS regression in step b will produce inconsistent estimation. This problem happens because we did not include all exogenous variable in the first stage regression.

## Problem 5.13

Consider the simple regression model

$$y = \beta_1 + \beta_1 x + u$$

and let $z$ be a *binary* instrumental variable for $x$.

a. Show that the IV estimator $\widehat{\beta}_1$ can be written as

$$\widehat{\beta}_1 = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0),$$

where $\bar{y}_0$ and $\bar{x}_0$ are the sample averages of $y_i$ and $x_i$ over the part of the sample with $z_i = 0$, $\bar{y}_1$ and $\bar{x}_1$ are the sample averages of $y_i$ and $x_i$ over the part of the sample with $z_i = 1$. This estimator, knows as a **grouping estimator**, was first suggested by Wald (1940).
Answer:
Recall our IV estimator

$$\widehat{\beta}_{IV} = \left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'y_i\right).$$

In the case of simple regression model with single IV, our estimator can be written as

$$
\begin{aligned}
\widehat{\beta}_1 &= \text{Cov}(z,x)^{-1}\text{Cov}(z,y) \\
&= \left(N^{-1}\sum_{i=1}^{N}(z_i-\bar{z})(x_i-\bar{x})\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}(z_i-\bar{z})(y_i-\bar{y})\right) \\
&= \left(\sum_{i=1}^{N}(z_i-\bar{z})(x_i-\bar{x})\right)^{-1}\left(\sum_{i=1}^{N}(z_i-\bar{z})(y_i-\bar{y})\right) \\
&= \left(\sum_{i=1}^{N}z_i(x_i-\bar{x})\right)^{-1}\left(\sum_{i=1}^{N}z_i(y_i-\bar{y})\right) \\
&= \left(\sum_{i=1}^{N}z_ix_i-\bar{x}\sum_{i=1}^{N}z_i)\right)^{-1}\left(\sum_{i=1}^{N}z_iy_i-\bar{y}\sum_{i=1}^{N}z_i\right) && [\text{Let } N_1=\sum_{i=1}^{N}1(z_i=1), N_0=\sum_{i=1}^{N}1(z_i=0)] \\
&= (N_1\bar{x}_1-\bar{x}N_1)^{-1}(N_1\bar{y}_1-\bar{y}N_1) && [\text{Note } N\bar{y}=N_0\bar{y}_0+N_1\bar{y}_1, \text{the same for } x] \\
&= (\bar{x}_1-\bar{x})^{-1}(\bar{y}_1-\bar{y}) \\
&= \frac{(\bar{y}_1-((N_0/N)\bar{y}_0+(N_1/N)\bar{y}_1))}{(\bar{x}_1-((N_0/N)\bar{x}_0+(N_1/N)\bar{x}_1))} \\
&= \frac{(((N-N_1)/N)\bar{y}_1-(N_0/N)\bar{y}_0)}{(((N-N_1)/N)\bar{x}_1-(N_0/N)\bar{x}_0)} && [\text{Note } N-N_1=N_0] \\
&= (\bar{y}_1-\bar{y}_0)/(\bar{x}_1-\bar{x}_0).
\end{aligned}
$$

□

b. What is the interpretation of $\widehat{\beta}_1$ if $x$ is also binary, for example, representing participation in a social program?
Answer:
When $x$ is also binary, $\bar{x}_1$ represents the fraction of observations receiving treatment when $z_i = 1$ and $\bar{x}_0$ represents the fraction of observations receiving treatment when $z_i = 0$. We can see $z_1 = 1$ as eligibility to receive a treatment. Suppose one of the observations have $z_i = 1, x_i = 1$ it means they are eligible and receive treatment. Thus $\bar{x}_1$ represent the fraction of eligible people participating in the treatment, and $\bar{x}_0$ represent the fraction of non-eligible people participating in the treatment. We can see $z_i = 1$ as being offered scholarship and $x_i = 1$ as taking the scholarship offer. We can interpret $\widehat{\beta}_1$ as the difference in mean outcome between $z = 1$ and $z = 0$ divided by the difference in participation rate between those groups, it is also known as average treatment effect.

# Chapter 6

## Problem 6.3

Consider a model for individual data to test whether nutrition affects productivity (in a developing country):

$$\log(produc) = \delta_0 + \delta_1 exper + \delta_2 exper^2 + \delta_3 educ + \alpha_1 calories + \alpha_2 protein + u_1, \tag{6.57}$$

where *produc* is some measure of worker productivity, *calories* is calorie intake per day, and *protein* is a measure of protein intake per day. Assume here that $exper$, $exper^2$, and $educ$ are all exogenous. The variables *calories* and *protein* are possibly correlated with $u_1$ (See Strauss and Thomas (1995) for discussion). Possible instrumental variables for *calories* and *protein* are regional prices of various goods, such as grains, meats, breads, dairy products, and so on.

a. Under what circumstances do prices make good IVs for *calories* and *protein*? What if prices reflect quality of food?

Answer:
Recall two conditions for a good IV, first is that prices must be partially correlated with *calories* and *protein* (the rank condition), and the second, prices is not correlated with the error term $u_1$ (or exogenous in equation (6.57)). We can check the first condition by running reduced form. For the second condition, we must show that prices are not systematically related with individual productivity. If prices reflect quality of food, then it may cause prices to be correlated with the error term, since in equation (6.57) quality of food are omitted and appear in the error $u_1$.

b. How many prices are needed to identify equation (6.57)?
Answer:
Since we suspect *calories* and *protein* to be endogenous then we must have two instruments minimum.

c. Suppose we have $M$ prices, $p_1, \ldots, p_M$. Explain how to test the null hypothesis that *calories* and *protein* are exogenous in equation (6.57)?
Answer:
We can do the following steps:

(a) Estimate two reduced forms: (1) Regress *calories* on $exper, exper^2, educ$ and instruments $p_1, \ldots, p_M$ and obtain the residual $\widehat{v}_1$, (2) Regress *protein* on $exper, exper^2, educ$ and instruments $p_1, \ldots, p_M$ and obtain the residual $\widehat{v}_2$

(b) Run OLS regression: Regress $\log(produc)$ on $exper, exper^2, educ, \widehat{v}_1, \widehat{v}_2$.

(c) Test the joint significance on $\widehat{v}_1, \widehat{v}_2$ using F-test.

## Problem 6.7

For this problem use the data in HPRICE.RAW, which is a subset of the data used by Kiel and McClain (1995). The file contains housing prices and characteristics for two years, 1978 and 1981, for homes sold in North Andover, Massachusetts. In 1981, construction on a garbage incinerator began. Rumors about the incinerator being built were circulating in 1979, and it is for this reason that 1978 is used as the base year. By 1981 it was very clear that the incinerator would be operating soon.

a. Using the 1981 cross section, estimate a bivariate, constant elasticity model relating housing price to distance from the incinerator. Is this regression appropriate for determining the causal effects of incinerator on housing prices? Explain.
Answer:
Table 6 shows the estimates for a bivariate constant elasticity model. The regression result is statistically significant and convincing with an elasticity of 0.365. However, we must suspect that the model may have endogeneity problem that will cause our OLS estimator to be inconsistent. Thus, the inference result may not be meaningful. In this case the endogeneity problem may arise from simultaneity, it may be that the incinerator site is determined to be in area where the housing price is already low.

b. Pooling the two years of data, consider the model

$$\log(price) = \delta_0 + \delta_1 y81 + \delta_2 \log(dist) + \delta_3 y81 \cdot \log(dist) + u.$$

If the incinerator has a negative effect on housing prices for homes closer to the incinerator, what sign is $\delta_3$? Estimate this model and test the null hypothesis that the incinerator had no effect on housing prices.
Answer:
If the incinerator has a negative effect on housing prices for homes closer to the incinerator, then $\delta_3$ should be positive. The expected sign means the house built farther from the incinerator will have higher prices in the year that the incinerator will be operating (1981). Our hypothesis will be

$$H_0 : \delta_3 = 0, \quad H_1 : \delta_3 > 0$$

Table 6: Regression result for Problem 6.7.a.

| | (1) |
|---|---|
| log(dist) | 0.365*** |
| | (0.066) |
| Constant | 8.047*** |
| | (0.646) |
| Observations | 142 |

Standard errors in parentheses
Data: HPRICE.DTA
Wooldridge (2011)
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Regression results: (1) Problem 6.7.b. and (2) 6.7.c

| | (1) | (2) |
|---|---|---|
| y81 | -0.011 | -0.230 |
| | (0.805) | (0.488) |
| log(dist) | 0.317*** | 0.087* |
| | (0.052) | (0.052) |
| y81 x log(dist) | 0.048 | 0.062 |
| | (0.082) | (0.050) |
| log(intst) | | 0.963*** |
| | | (0.326) |
| $\text{lintst}^2$ | | -0.059*** |
| | | (0.019) |
| log(area) | | 0.355*** |
| | | (0.051) |
| log(land) | | 0.110*** |
| | | (0.025) |
| age of house | | -0.007*** |
| | | (0.001) |
| $\text{age}^2$ | | 0.000*** |
| | | (0.000) |
| rooms in house | | 0.047*** |
| | | (0.017) |
| bathrooms | | 0.096*** |
| | | (0.027) |
| Constant | 8.058*** | 2.306 |
| | (0.508) | (1.774) |
| Observations | 321 | 321 |

Standard errors in parentheses
Data: HPRICE.DTA
Wooldridge (2011)
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7 Column (1) shows the estimates for the model in part b. The sign of the estimates on the interaction terms is positive as expected, but it is not large and not significant (P-value of 0.556). Therefore, we can not reject the null hypothesis that construction of the incinerator have no effect on housing price.

c. Add the variables $\log(intst), [\log(intst)]^2, \log(area), \log(land), age, age^2, rooms, baths$ to the model in part b, and test for an incinerator effect. What do you conclude?
Answer:
Table 7 Column (2) shows the estimates for the model in part b and adding the following variables: $\log(intst), [\log(intst)]^2$, $\log(area), \log(land), age, age^2, rooms, baths$ to the model. The coefficient on the interaction term in now larger, with elasticity of 0.062 with an improved statistical significance. The P-value is 0.214 for two-sided or equivalent with 0.107 for one-sided hypothesis. This is almost significant at 10% but still the evidence of detrimental effect of incinerator being built to house prices is weak.

# Non-textbook Problem

Derive the variance of the IV estimator for the case of heteroskedasticity.
Answer:
Suppose we have the following model $y = \mathbf{x}\boldsymbol{\beta} + u$, that have some endogenous regressors that we will be estimating by IV estimation using instruments $z$. Recall the IV estimator

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{IV} &= [(\mathbf{x}'\mathbf{z})(\mathbf{z}'\mathbf{z})^{-1}(\mathbf{z}'\mathbf{x})]^{-1}[(\mathbf{x}'\mathbf{z})(\mathbf{z}'\mathbf{z})^{-1}(\mathbf{z}'\mathbf{y})] \\
&= \left[\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{z}_i\right)\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{x}_i\right)\right]^{-1} \\
&\quad \left[\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{z}_i\right)\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'y_i\right)\right].
\end{aligned}$$

Substitute $y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i$, we obtain

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{IV} = \boldsymbol{\beta} + &\left[\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{z}_i\right)\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{x}_i\right)\right]^{-1} \\
&\left[\underbrace{\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{z}_i\right)}_{C'}\underbrace{\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}}_{D^{-1}}\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'u_i\right)\right] \\
\Leftrightarrow \widehat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta} &= [C'D^{-1}C]^{-1}C'D^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'u_i\right) \\
\Leftrightarrow \sqrt{N}(\widehat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) &= [C'D^{-1}C]^{-1}C'D^{-1}\left(N^{-\frac{1}{2}}\sum_{i=1}^{N}\mathbf{z}_i'u_i\right).
\end{aligned}$$

Note that We take the consistency of IV for granted and focus on deriving the variance. We have that $[C'D^{-1}C]^{-1}C'D^{-1} = o_p(1)$, and by Central Limit Theorem, we have $N^{-\frac{1}{2}}\sum_{i=1}^{N}\mathbf{z}_i'u_i = O_p(1)$. Applying

Central Limit Theorem we have

$$\left(N^{-\frac{1}{2}}\sum_{i=1}^{N}\mathbf{z}_i'u_i\right) \xrightarrow{d} \mathbb{N}(0, \mathrm{E}(u^2\mathbf{z}'\mathbf{z}))$$

$$\Leftrightarrow \sqrt{N}(\widehat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) = [C'D^{-1}C]^{-1}C'D^{-1}\left(N^{-\frac{1}{2}}\sum_{i=1}^{N}\mathbf{z}_i'u_i\right) \xrightarrow{d} \mathbb{N}(0, [C'D^{-1}C]^{-1}[C'D^{-1}\mathrm{E}(u^2\mathbf{z}'\mathbf{z})D^{-1}C][C'D^{-1}C]^{-1}).$$

Finally we have the asymptotic distribution is

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) \xrightarrow{d} \mathbb{N}(0, [C'D^{-1}C]^{-1}[C'D^{-1}\mathrm{E}(u^2\mathbf{z}'\mathbf{z})D^{-1}C][C'D^{-1}C]^{-1}),$$

written differently, the variance of IV estimator under heteroskedasticity is

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{IV}) = [C'D^{-1}C]^{-1}\left[C'D^{-1}\left(N^{-1}\sum_{i=1}^{N}\widehat{u}_i^2\mathbf{z}_i'\mathbf{z}_i\right)D^{-1}C\right][C'D^{-1}C]^{-1}.$$

$\square$