PS6 Solution

Chunyu Qu

Nov 1, 2021

6.1

Answer

Recap 5.4. Use the data in CARD.RAW

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).

```
# Run OLS
CARD = read.csv("D:/Google Drive/Fordham/2019 Spring/AE/MITDataCSV/CARD.csv", header = TRUE)

lm54a_ols = lm(lwage ~ educ + exper + expersq + black + south + smsa +reg661 + reg662 + reg663 + reg664
summary(lm54a_ols)
```

```
##
## Call:
## lm(formula = lwage ~ educ + exper + expersq + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66, data = CARD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62326 -0.22141  0.02001  0.23932  1.33340
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.7393766  0.0715282  66.259  < 2e-16 ***
## educ         0.0746933  0.0034983  21.351  < 2e-16 ***
## exper        0.0848320  0.0066242  12.806  < 2e-16 ***
## expersq     -0.0022870  0.0003166  -7.223 6.41e-13 ***
## black       -0.1990123  0.0182483 -10.906  < 2e-16 ***
## south       -0.1479550  0.0259799  -5.695 1.35e-08 ***
## smsa         0.1363846  0.0201005   6.785 1.39e-11 ***
## reg661      -0.1185697  0.0388301  -3.054 0.002281 **
## reg662      -0.0222026  0.0282575  -0.786 0.432092
## reg663       0.0259703  0.0273644   0.949 0.342670
## reg664      -0.0634942  0.0356803  -1.780 0.075254 .
## reg665       0.0094551  0.0361174   0.262 0.793504
```

```
## reg666       0.0219476  0.0400984   0.547 0.584182
## reg667      -0.0005887  0.0393793  -0.015 0.988073
## reg668      -0.1750058  0.0463394  -3.777 0.000162 ***
## smsa66       0.0262417  0.0194477   1.349 0.177327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3723 on 2994 degrees of freedom
## Multiple R-squared:  0.2998, Adjusted R-squared:  0.2963
## F-statistic: 85.48 on 15 and 2994 DF,  p-value: < 2.2e-16
```

The estimated return to education is about 7.5%, with a very large t statistic. These reproduce the estimates from Table 2, Column (2) in Card (1995).

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).)

```
# Run Reduced Form regression
lm54b_rf = lm(educ ~ nearc4 + exper + expersq + black + south + smsa +reg661 + reg662 + reg663 + reg664
summary(lm54b_rf)
```

```
##
## Call:
## lm(formula = educ ~ nearc4 + exper + expersq + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66, data = CARD)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.545 -1.370 -0.091  1.278  6.239
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.8485239  0.2111222  79.805  < 2e-16 ***
## nearc4       0.3198989  0.0878638   3.641 0.000276 ***
## exper       -0.4125334  0.0336996 -12.241  < 2e-16 ***
## expersq      0.0008686  0.0016504   0.526 0.598728
## black       -0.9355287  0.0937348  -9.981  < 2e-16 ***
## south       -0.0516126  0.1354284  -0.381 0.703152
## smsa         0.4021825  0.1048112   3.837 0.000127 ***
## reg661      -0.2102710  0.2024568  -1.039 0.299076
## reg662      -0.2889073  0.1473395  -1.961 0.049992 *
## reg663      -0.2382099  0.1426357  -1.670 0.095012 .
## reg664      -0.0930890  0.1859827  -0.501 0.616742
## reg665      -0.4828875  0.1881872  -2.566 0.010336 *
## reg666      -0.5130857  0.2096352  -2.448 0.014442 *
## reg667      -0.4270887  0.2056208  -2.077 0.037880 *
## reg668       0.3136204  0.2416739   1.298 0.194490
## smsa66       0.0254805  0.1057692   0.241 0.809644
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.941 on 2994 degrees of freedom
```

```
## Multiple R-squared:  0.4771, Adjusted R-squared:  0.4745
## F-statistic: 182.1 on 15 and 2994 DF,  p-value: < 2.2e-16
```

The important coefficient is on nearc4. Statistically, educ and nearc4 are partially correlated, and in a way that makes sense: holding other factors in the reduced form fixed, someone living near a four-year college at age 16 has, on average, almost one-third a year more education than a person not near a four-year college at age 16. This is not trivial a effect, so nearc4 passes the requirement that it is partially correlated with educ.

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).)

```r
# Run IV regression
lm54c_iv = ivreg(lwage ~ educ + exper + expersq + black + south + smsa +reg661 + reg662 + reg663 + reg6
summary(lm54c_iv)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ + exper + expersq + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66 | exper + expersq + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66 + nearc4, data = CARD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83164 -0.24075  0.02428  0.25208  1.42760
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.7739661  0.9349469   4.037 5.56e-05 ***
## educ         0.1315038  0.0549637   2.393  0.01679 *
## exper        0.1082711  0.0236586   4.576 4.92e-06 ***
## expersq     -0.0023349  0.0003335  -7.001 3.12e-12 ***
## black       -0.1467758  0.0538999  -2.723  0.00650 **
## south       -0.1446715  0.0272846  -5.302 1.23e-07 ***
## smsa         0.1118084  0.0316620   3.531  0.00042 ***
## reg661      -0.1078142  0.0418137  -2.578  0.00997 **
## reg662      -0.0070465  0.0329073  -0.214  0.83046
## reg663       0.0404445  0.0317806   1.273  0.20325
## reg664      -0.0579171  0.0376059  -1.540  0.12364
## reg665       0.0384576  0.0469387   0.819  0.41267
## reg666       0.0550887  0.0526597   1.046  0.29559
## reg667       0.0267580  0.0488287   0.548  0.58373
## reg668      -0.1908912  0.0507113  -3.764  0.00017 ***
## smsa66       0.0185311  0.0216086   0.858  0.39119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3883 on 2994 degrees of freedom
## Multiple R-Squared: 0.2382,  Adjusted R-squared: 0.2343
## Wald test: 51.01 on 15 and 2994 DF,  p-value: < 2.2e-16
```

The estimated return to education has increased to about 13.2%, but notice how wide the 95% confidence interval is: 2.4% to 23.9%. By contrast, the OLS confidence interval is about 6.8% to 8.2%, which is much

tighter. Of course, OLS could be inconsistent, in which case a tighter CI is of little value. But the estimated return to education is higher with IV, something that seems a bit counterintuitive.
One possible explanation is that educ suffers from classical errors-in-variables. Therefore, while OLS would tend to overestimate the return to schooling because of omitted "ability," classical measurement error in educ leads to an attenuation bias. Measurement error may help explain why the IV estimate is larger, but it is not entirely convincing. It seems unlikely that educ satisfies the CEV assumptions. For example, if we think the measurement error is due to truncation - people are asked about highest grade completed, not actual years of schooling - then educ is always less than or equal to educ???. And the measurement error could not be independent of educ???. If we think the mismeasurement is due to is unobserved quality of schooling, it seems likely that quality of schooling part of the measurement error is positively correlated with actual amount of schooling. This, too, violates the CEV assumptions.
Another possibility for the much higher IV estimate comes out of the recent treatment effect literature, which is covered in Section 21.4. Of course, we must also remember that the point estimates - particularly the IV estimate - are subject to substantial sampling variation. At this point, we do not even know of OLS and IV are statistically different from each other. See Problem 6.1.

d. Now use nearc2 along with nearc4 as instruments for educ. First estimate the reduced form for educ, and comment on whether nearc2 or nearc4 is more strongly related to educ. How do the 2SLS estimates compare with the earlier estimates?

```
# Run Reduced Form regression
lm54d_rf = lm(educ ~ nearc2+ nearc4 + exper + expersq + black + south + smsa +reg661 + reg662 + reg663 +
summary(lm54d_rf)
```

```
##
## Call:
## lm(formula = educ ~ nearc2 + nearc4 + exper + expersq + black +
##     south + smsa + reg661 + reg662 + reg663 + reg664 + reg665 +
##     reg666 + reg667 + reg668 + smsa66, data = CARD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5851 -1.3845 -0.0823  1.2765  6.2930
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.677e+01  2.163e-01  77.528  < 2e-16 ***
## nearc2       1.230e-01  7.743e-02   1.589 0.112256
## nearc4       3.206e-01  8.784e-02   3.650 0.000267 ***
## exper       -4.123e-01  3.369e-02 -12.237  < 2e-16 ***
## expersq      8.479e-04  1.650e-03   0.514 0.607379
## black       -9.452e-01  9.391e-02 -10.065  < 2e-16 ***
## south       -4.191e-02  1.355e-01  -0.309 0.757162
## smsa         4.014e-01  1.048e-01   3.830 0.000131 ***
## reg661      -1.688e-01  2.041e-01  -0.827 0.408286
## reg662      -2.690e-01  1.478e-01  -1.820 0.068884 .
## reg663      -1.902e-01  1.458e-01  -1.305 0.192022
## reg664      -3.772e-02  1.892e-01  -0.199 0.841990
## reg665      -4.371e-01  1.903e-01  -2.297 0.021703 *
## reg666      -5.022e-01  2.097e-01  -2.395 0.016679 *
## reg667      -3.775e-01  2.079e-01  -1.816 0.069511 .
## reg668       3.820e-01  2.454e-01   1.557 0.119683
## smsa66       7.825e-05  1.069e-01   0.001 0.999416
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.94 on 2993 degrees of freedom
## Multiple R-squared:  0.4776, Adjusted R-squared:  0.4748
## F-statistic:   171 on 16 and 2993 DF,  p-value: < 2.2e-16
```

```
# 2. Run IV regression
lm54d_iv = ivreg(lwage ~ educ + exper + expersq + black + south + smsa +reg661 + reg662 + reg663 + reg6
summary(lm54d_iv)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ + exper + expersq + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66 | exper + expersq + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66 + nearc2 + nearc4, data = CARD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93841 -0.25068  0.01932  0.26519  1.46998
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.3396875  0.8945377   3.733 0.000192 ***
## educ         0.1570593  0.0525782   2.987 0.002839 **
## exper        0.1188149  0.0228061   5.210 2.02e-07 ***
## expersq     -0.0023565  0.0003475  -6.781 1.43e-11 ***
## black       -0.1232778  0.0521500  -2.364 0.018147 *
## south       -0.1431945  0.0284448  -5.034 5.08e-07 ***
## smsa         0.1007530  0.0315193   3.197 0.001405 **
## reg661      -0.1029760  0.0434224  -2.371 0.017779 *
## reg662      -0.0002287  0.0337943  -0.007 0.994602
## reg663       0.0469556  0.0326490   1.438 0.150484
## reg664      -0.0554084  0.0391828  -1.414 0.157437
## reg665       0.0515041  0.0475678   1.083 0.279006
## reg666       0.0699968  0.0533049   1.313 0.189237
## reg667       0.0390596  0.0497499   0.785 0.432446
## reg668      -0.1980371  0.0525350  -3.770 0.000167 ***
## smsa66       0.0150626  0.0223360   0.674 0.500132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4053 on 2994 degrees of freedom
## Multiple R-Squared: 0.1702,  Adjusted R-squared: 0.166
## Wald test: 47.07 on 15 and 2994 DF,  p-value: < 2.2e-16
```

When nearc2 is added to the reduced form of educ it has a coefficient (standard error) of .123 (.077), compared with .321 (.089) for nearc4. Therefore, nearc4 has a much stronger ceteris paribus relationship with educ; nearc2 is only marginally statistically significant once nearc4 has been included. The joint F test gives F ??? 7.89 with p-value ??? .004.
The 2SLS estimate of the return to education becomes about 15.7%, with 95% CI given by 5.4% to 26%. The CI is still very wide.

Now answer 6.1,

a. test the null hypothesis that educ is exogenous.

Answer

```
# Obtain estimated v2 and u1 first
v2_hat = resid(lm54d_rf)
u_hat1 = resid(lm54d_iv)

lm61a = lm(lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 + reg663 + reg664 + r
summary(lm61a)
```

```
##
## Call:
## lm(formula = lwage ~ educ + exper + expersq + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66 + v2_hat, data = CARD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63371 -0.22173  0.01706  0.23886  1.32811
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.3396875  0.8214340   4.066 4.91e-05 ***
## educ         0.1570593  0.0482814   3.253 0.001155 **
## exper        0.1188149  0.0209423   5.673 1.53e-08 ***
## expersq     -0.0023565  0.0003191  -7.384 1.98e-13 ***
## black       -0.1232778  0.0478882  -2.574 0.010092 *
## south       -0.1431945  0.0261202  -5.482 4.55e-08 ***
## smsa         0.1007530  0.0289435   3.481 0.000507 ***
## reg661      -0.1029760  0.0398738  -2.583 0.009854 **
## reg662      -0.0002287  0.0310325  -0.007 0.994121
## reg663       0.0469556  0.0299809   1.566 0.117411
## reg664      -0.0554084  0.0359807  -1.540 0.123679
## reg665       0.0515041  0.0436804   1.179 0.238447
## reg666       0.0699968  0.0489487   1.430 0.152821
## reg667       0.0390596  0.0456842   0.855 0.392625
## reg668      -0.1980371  0.0482417  -4.105 4.15e-05 ***
## smsa66       0.0150626  0.0205106   0.734 0.462775
## v2_hat      -0.0828005  0.0484086  -1.710 0.087286 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3722 on 2993 degrees of freedom
## Multiple R-squared:  0.3005, Adjusted R-squared:  0.2968
## F-statistic: 80.37 on 16 and 2993 DF,  p-value: < 2.2e-16
```

The t statistic on $\hat{v}_2$ is ???1.71, which is not significant at the 5% level against a two-sided alternative. The negative correlation between u1 and educ is essentially the same finding that the 2SLS estimated return to education is larger than the OLS estimate. In any case, I would call this marginal evidence that educ is endogenous. The quandary is that the OLS and 2SLS.

b. Test the the single overidentifying restriction in this example.

Answer

```
# Then we regress estimated u1 on all the variables with educ replacing by IVs
lm61b = lm(u_hat1 ~ exper + expersq + black + south + smsa + reg661 + reg662 + reg663 + reg664 + reg665

# Obtain R square value
Rsq = summary(lm61b)$r.squared
pchisq(Rsq * nrow(CARD), df=1, lower.tail=FALSE)
```

```
## [1] 0.2639051
```

The test statistic is the sample size times the R-squared from this regression, or about 1.25. The p-value, obtained from chi-square distribution, is about .264, so the instruments pass the over identification test.

6.2

In Problem 5.8b, test the null hypothesis that educ and IQ are exogenous in the equation estimated by 2SLS.

Answer

Answer 5.8 a-b first

Consider a model with unobserved heterogeneity (q) and measurement error in an explanatory variable:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K^* + q + v$$

where $e_K = x_K - x_K^*$ is the measurement error and we set the coefficient on q equal to one without loss of generality. The variable q might be correlated with any of the explanatory variables, but an indicator, $q_1 = \delta_0 + \delta_1 q + a_1$, is available. The measurement error e_K might be correlated with the observed measure, x_K, In addition to q_1, you might also have variables $z_1, ...z_M, M \geq 2$, that are uncorrelated with v, a_1, e_K.

a. Suggest an IV procedure for consistently estimating the β_j. Why is $M \geq 2$ required? (Hint: Plug in q_1 for q and x_K for x_K^*, and go from there.)

Answer

Plug in the indicator q_1 for q and the measurement x_K for $x_K^?$?? , being sure to keep track of the errors:

$$y = \gamma_0 + \beta_1 x_1 + ... + \beta_K x_K + \gamma_1 q_1 + v - \beta_K e_K + \gamma_1 a_1$$

$$= \gamma_0 + \beta_1 x_1 + ... + \beta_K x_K + \gamma_1 q_1 + u$$

where $\gamma_1 = 1/\delta_1$. Now, if the variables $z_1, ...z_M$ are redundant in the structural equation (so they are uncorrelated with v), and uncorrelated with the measurement error e_K and the indicator error a_1 we can use these as IVs for x_K and q_1 in 2SLS.
We need $M \geq 2$ because we have two explanatory variables, x_q and q_1, that are possibly correlated with the composite error u.

b. Apply this method to the model estimated in Example 5.5, where actual education, say $educ^*$, plays the role of x_K^* . Use IQ as the indicator of q=ability, and KWW, meduc, feduc, and sibs as the elements of z.

Answer

```r
# Read the data
NLS80 = read.csv("D:/Google Drive/Fordham/2019 Spring/AE/MITDataCSV/NLS80.csv", header = TRUE)
# Run the reduced forms
lm_rd1 = lm(educ ~ exper + tenure + married + south + urban + black + kww + meduc + feduc + sibs, data=
v21_hat = resid(lm_rd1)
length(NLS80$educ)
```

```
## [1] 935
```

```r
length(v21_hat)
```

```
## [1] 722
```

```r
# Since the lengths of the response variable and the predictor are not the same, there might be NA valu
NLS80omit = na.omit(NLS80)
# Run the regression again
lm_rd1 = lm(educ ~ exper + tenure + married + south + urban + black + kww + meduc + feduc + sibs, data=
v21_hat = resid(lm_rd1)

lm_rd2 = lm(iq ~ exper + tenure + married + south + urban + black + kww + meduc + feduc + sibs, data=NL
v22_hat = resid(lm_rd2)

# Run the 2nd stage regression
lm_2s = lm(lwage ~ exper + tenure + married + south + urban + black + educ + iq + v21_hat + v22_hat, da
summary(lm_2s )
```

```
##
## Call:
## lm(formula = lwage ~ exper + tenure + married + south + urban +
##     black + educ + iq + v21_hat + v22_hat, data = NLS80omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0900 -0.1953  0.0055  0.2347  1.2369
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.909488   0.494295   9.932  < 2e-16 ***
## exper        0.033545   0.011764   2.852  0.00449 **
## tenure       0.006143   0.003151   1.949  0.05168 .
## married      0.217867   0.047637   4.573 5.74e-06 ***
## south       -0.095903   0.056212  -1.706  0.08847 .
## urban        0.175886   0.033475   5.254 2.02e-07 ***
## black       -0.269254   0.221863  -1.214  0.22534
## educ         0.181555   0.112004   1.621  0.10551
## iq          -0.012445   0.019932  -0.624  0.53259
## v21_hat     -0.137434   0.112348  -1.223  0.22166
## v22_hat      0.015495   0.019968   0.776  0.43804
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3539 on 652 degrees of freedom
```

```
## Multiple R-squared:  0.274,  Adjusted R-squared:  0.2628
## F-statistic: 24.61 on 10 and 652 DF,  p-value: < 2.2e-16
```

Now work on 6.2

```
linearHypothesis(lm_2s , c("v21_hat=0", "v22_hat=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## v21_hat = 0
## v22_hat = 0
##
## Model 1: restricted model
## Model 2: lwage ~ exper + tenure + married + south + urban + black + educ +
##     iq + v21_hat + v22_hat
##
##   Res.Df    RSS Df Sum of Sq     F   Pr(>F)
## 1    654 82.970
## 2    652 81.665  2    1.3054 5.211 0.005686 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore, the test finds fairly strong evidence for endogeneity of at least one of educ and IQ, although this conclusion relies on the instruments being truly exogenous. If you look back at Problem 5.8, this IV solution did not seem to work very well. So we still do not know what should be treated as exogenous in this method.

6.3

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

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

where produc is some measure of worker productivity, calories is caloric 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 proteins? What if prices reflect quality of food?

Answer

We need prices to satisfy two requirements. First, calories and protein must be partially correlated with prices of food. While this is easy to test separately by estimating the two reduced forms, the rank condition could still be violated. (Problem 5.15c contains a sufficient condition for the rank condition to hold.) In addition, we must also assume prices are exogenous in the productivity equation. Ideally, prices vary because of things like transportation costs that are not systematically related to regional variations in individual productivity. A potential problem is that prices reflect food quality and that features of the food other than calories and protein appear in the disturbance u1.

b. How many prices are needed to identify equation (6.57)?

Answer

Since there are two endogenous explanatory variables we need at least two prices.

c. Suppose we have M prices, p1; . . . ; pM. Explain how to test the null hypothesis that calories and protein are exogenous in equation (6.57).

Answer

We would first estimate the two reduced forms for calories and protein by regressing each on a constant, exper,$exper^2$ and educ and the M prices $p_1, ...p_M$, we obtain $\hat{v_{21}}, \hat{v_{22}}$. Then we would run the regression log(produc) on 1,exper,$exper^2$ and educ, $\hat{v_{21}}, \hat{v_{22}}$, and do a joint significance test on $\hat{v_{21}}, \hat{v_{22}}$. We could use a standard F test or use a heteroskedasticity-robust test.

6.8

The data in FERTIL1.RAW are a pooled cross section on more than a thousand U.S. women for the even years between 1972 and 1984, inclusive; the data set is similar to the one used by Sander (1992). These data can be used to study the relationship between women's education and fertility.

a. Use OLS to estimate a model relating number of children ever born to a woman (kids) to years of education, age, region, race, and type of environment reared in. You should use a quadratic in age and should include year dummies. What is the estimated relationship between fertility and education? Holding other factors fixed, has there been any notable secular change in fertility over the time period?

Answer

```
FERTIL1 = read.csv("D:/Google Drive/Fordham/2019 Spring/AE/MITDataCSV/FERTIL1.csv", header = TRUE)

# Run the OLS
lm68a= lm(kids ~ educ+age+agesq+black+east+northcen+west+farm+othrural+town+smcity+y74+y76+y78+y80+y82+y
summary(lm68a)
```

```
##
## Call:
## lm(formula = kids ~ educ + age + agesq + black + east + northcen +
##     west + farm + othrural + town + smcity + y74 + y76 + y78 +
##     y80 + y82 + y84, data = FERTIL1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9878 -1.0086 -0.0767  0.9331  4.6548
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.742457   3.051767  -2.537 0.011315 *
## educ        -0.128427   0.018349  -6.999 4.44e-12 ***
## age          0.532135   0.138386   3.845 0.000127 ***
## agesq       -0.005804   0.001564  -3.710 0.000217 ***
## black        1.075658   0.173536   6.198 8.02e-10 ***
## east         0.217324   0.132788   1.637 0.101992
## northcen     0.363114   0.120897   3.004 0.002729 **
## west         0.197603   0.166913   1.184 0.236719
## farm        -0.052557   0.147190  -0.357 0.721105
## othrural    -0.162854   0.175442  -0.928 0.353481
```

```
## town         0.084353   0.124531   0.677 0.498314
## smcity       0.211879   0.160296   1.322 0.186507
## y74          0.268183   0.172716   1.553 0.120771
## y76         -0.097379   0.179046  -0.544 0.586633
## y78         -0.068666   0.181684  -0.378 0.705544
## y80         -0.071305   0.182771  -0.390 0.696511
## y82         -0.522484   0.172436  -3.030 0.002502 **
## y84         -0.545166   0.174516  -3.124 0.001831 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.555 on 1111 degrees of freedom
## Multiple R-squared:  0.1295, Adjusted R-squared:  0.1162
## F-statistic: 9.723 on 17 and 1111 DF,  p-value: < 2.2e-16
```

The estimate says that a women with about eight more years of education has about one fewer child (gotten from .128$times$8=1.024), other factors fixed. The coefficient is very statistically significant. Also, there has been a notable secular decline in fertility over this period: on average, with other factors held fixed, a women in 1984 had about half a child less -0.545 than a similar woman in 1972, the base year. The effect is also statistically significant with p-value=. 002.

b. Reestimate the model in part a, but use motheduc and fatheduc as instruments for educ. First check that these instruments are sufficiently partially correlated with educ. Test whether educ is in fact exogenous in the fertility equation.

Answer

```
# Run the 1st stage reduced form for educ
lm68b= lm(educ ~ age+agesq+black+east+northcen+west+farm+othrural+town+smcity+y74+y76+y78+y80+y82+y84+m
summary(lm68b)
```

```
##
## Call:
## lm(formula = educ ~ age + agesq + black + east + northcen + west +
##     farm + othrural + town + smcity + y74 + y76 + y78 + y80 +
##     y82 + y84 + meduc + feduc, data = FERTIL1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9976  -1.3767  -0.2344   1.1540  10.6059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.633344   4.396773   3.101  0.00198 **
## age         -0.224369   0.200001  -1.122  0.26217
## agesq        0.002566   0.002261   1.135  0.25648
## black        0.366782   0.252287   1.454  0.14628
## east         0.248804   0.192014   1.296  0.19533
## northcen     0.091395   0.175774   0.520  0.60320
## west         0.101068   0.242241   0.417  0.67660
## farm        -0.379262   0.214386  -1.769  0.07716 .
## othrural    -0.560814   0.255120  -2.198  0.02814 *
## town         0.061634   0.180783   0.341  0.73322
## smcity       0.080663   0.231739   0.348  0.72785
```

```
## y74          0.006099   0.249827   0.024  0.98053
## y76          0.123910   0.258792   0.479  0.63217
## y78          0.207786   0.262774   0.791  0.42926
## y80          0.382891   0.264243   1.449  0.14762
## y82          0.582040   0.249237   2.335  0.01971 *
## y84          0.425043   0.252901   1.681  0.09311 .
## meduc        0.172301   0.022196   7.763 1.88e-14 ***
## feduc        0.207419   0.025460   8.147 9.99e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.247 on 1110 degrees of freedom
## Multiple R-squared:  0.2869, Adjusted R-squared:  0.2754
## F-statistic: 24.82 on 18 and 1110 DF,  p-value: < 2.2e-16
```

Then we test the partial correlation

```
linearHypothesis(lm68b, c("meduc=0", "feduc=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## meduc = 0
## feduc = 0
##
## Model 1: restricted model
## Model 2: educ ~ age + agesq + black + east + northcen + west + farm +
##     othrural + town + smcity + y74 + y76 + y78 + y80 + y82 +
##     y84 + meduc + feduc
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1112 7180.7
## 2   1110 5606.9  2    1573.9 155.79 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The joint F test shows that educ is significantly partially correlated with meduc and feduc; the t statistics also show this clearly. If we make the test robust to heteroskedasticity of unknown form, the F statistic drops to 131.37 but the p-value is still zero to four decimal places.

To test the null that educ is exogenous, we need to reduced form residuals and then include them in the OLS regression. I suppress the output here:

```
# Obtain estimated v2
v2_hat = resid(lm68b)

lm_2nd = lm(kids ~ educ+age+agesq+black+east+northcen+west+farm+othrural+town+smcity+y74+y76+y78+y80+y8
summary(lm_2nd)
```

```
##
## Call:
## lm(formula = kids ~ educ + age + agesq + black + east + northcen +
##     west + farm + othrural + town + smcity + y74 + y76 + y78 +
```

```
##     y80 + y82 + y84 + v2_hat, data = FERTIL1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9816 -1.0100 -0.0601  0.9181  4.6688
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.241244   3.134883  -2.310 0.021077 *
## educ        -0.152739   0.039201  -3.896 0.000104 ***
## age          0.523554   0.138957   3.768 0.000173 ***
## agesq       -0.005716   0.001570  -3.642 0.000283 ***
## black        1.072952   0.173618   6.180 8.99e-10 ***
## east         0.228555   0.133779   1.708 0.087831 .
## northcen     0.374419   0.121992   3.069 0.002198 **
## west         0.207640   0.167563   1.239 0.215542
## farm        -0.077002   0.151287  -0.509 0.610869
## othrural    -0.195245   0.181449  -1.076 0.282147
## town         0.081810   0.124612   0.657 0.511628
## smcity       0.212500   0.160335   1.325 0.185329
## y74          0.272129   0.172847   1.574 0.115681
## y76         -0.094548   0.179132  -0.528 0.597734
## y78         -0.057254   0.182451  -0.314 0.753727
## y80         -0.053248   0.184614  -0.288 0.773072
## y82         -0.496215   0.176490  -2.812 0.005017 **
## y84         -0.521360   0.177821  -2.932 0.003438 **
## v2_hat       0.031137   0.044363   0.702 0.482906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.555 on 1110 degrees of freedom
## Multiple R-squared:  0.1299, Adjusted R-squared:  0.1158
## F-statistic: 9.206 on 18 and 1110 DF,  p-value: < 2.2e-16
```

The t statistic on v2hat is . 702, so there is little evidence that educ is endogenous in the equation. Still, we can see if 2SLS produces very different estimates:

```
# Generate the estimated educ
educ_hat = fitted.values(lm68b)

# Run the 2nd stage regression
lm_2nd = lm(kids ~ educ_hat+age+agesq+black+east+northcen+west+farm+othrural+town+smcity+y74+y76+y78+y8
summary(lm_2nd)
```

```
##
## Call:
## lm(formula = kids ~ educ_hat + age + agesq + black + east + northcen +
##     west + farm + othrural + town + smcity + y74 + y76 + y78 +
##     y80 + y82 + y84, data = FERTIL1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9611 -1.0591 -0.0576  0.9432  4.8706
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.241244   3.181487  -2.276 0.023032 *
## educ_hat    -0.152739   0.039784  -3.839 0.000130 ***
## age          0.523554   0.141023   3.713 0.000215 ***
## agesq       -0.005716   0.001593  -3.588 0.000347 ***
## black        1.072952   0.176199   6.089 1.56e-09 ***
## east         0.228555   0.135767   1.683 0.092572 .
## northcen     0.374419   0.123806   3.024 0.002550 **
## west         0.207640   0.170054   1.221 0.222336
## farm        -0.077002   0.153536  -0.502 0.616104
## othrural    -0.195245   0.184147  -1.060 0.289252
## town         0.081810   0.126465   0.647 0.517831
## smcity       0.212500   0.162719   1.306 0.191846
## y74          0.272129   0.175417   1.551 0.121107
## y76         -0.094548   0.181795  -0.520 0.603110
## y78         -0.057254   0.185164  -0.309 0.757220
## y80         -0.053248   0.187358  -0.284 0.776307
## y82         -0.496215   0.179114  -2.770 0.005692 **
## y84         -0.521360   0.180464  -2.889 0.003940 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.578 on 1111 degrees of freedom
## Multiple R-squared:  0.103,  Adjusted R-squared:  0.0893
## F-statistic: 7.507 on 17 and 1111 DF,  p-value: < 2.2e-16
```

The estimated coefficient on educ is larger in magnitude than before, but the test for endogeneity shows that we can reasonably attribute the difference between OLS and 2SLS to sampling error.

c. Now allow the effect of education to change over time by including interaction terms such as y74educ, y76educ, and so on in the model. Use interactions of time dummies and parents' education as instruments for the interaction terms. Test that there has been no change in the relationship between fertility and education over time.

Answer

Since there is little evidence that educ is endogenous, we could just use OLS. I did it both ways. First, I just added interactions y74educ, y76educ, . . . , y84educ to the model in part a and used OLS.

```
lm_68c1 = lm(kids ~ educ+age+agesq+black+east+northcen+west+farm+othrural+town+smcity+y74+y76+y78+y80+y8
summary(lm_68c1)
```

```
##
## Call:
## lm(formula = kids ~ educ + age + agesq + black + east + northcen +
##     west + farm + othrural + town + smcity + y74 + y76 + y78 +
##     y80 + y82 + y84 + y74educ + y76educ + y78educ + y80educ +
##     y82educ + y84educ, data = FERTIL1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5343 -1.0340 -0.0823  0.9550  4.6006
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.477302   3.126360  -2.712 0.006801 **
## educ        -0.022515   0.053618  -0.420 0.674628
## age          0.507466   0.138922   3.653 0.000271 ***
## agesq       -0.005525   0.001570  -3.519 0.000451 ***
## black        1.074055   0.173701   6.183 8.82e-10 ***
## east         0.206056   0.133143   1.548 0.121998
## northcen     0.348287   0.121099   2.876 0.004104 **
## west         0.177122   0.167452   1.058 0.290402
## farm        -0.072162   0.147508  -0.489 0.624791
## othrural    -0.191154   0.175934  -1.087 0.277491
## town         0.088229   0.124536   0.708 0.478804
## smcity       0.205358   0.160210   1.282 0.200182
## y74          0.946915   0.904159   1.047 0.295196
## y76          1.019963   0.882034   1.156 0.247777
## y78          1.805985   0.951866   1.897 0.058047 .
## y80          1.114183   0.897601   1.241 0.214762
## y82          1.199807   0.876289   1.369 0.171218
## y84          1.671261   0.899050   1.859 0.063304 .
## y74educ     -0.056425   0.072561  -0.778 0.436958
## y76educ     -0.092100   0.070875  -1.299 0.194053
## y78educ     -0.152387   0.075282  -2.024 0.043187 *
## y80educ     -0.097905   0.070452  -1.390 0.164912
## y82educ     -0.138945   0.068371  -2.032 0.042371 *
## y84educ     -0.176097   0.069915  -2.519 0.011918 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.553 on 1105 degrees of freedom
## Multiple R-squared:  0.1365, Adjusted R-squared:  0.1185
## F-statistic: 7.593 on 23 and 1105 DF,  p-value: < 2.2e-16
```

Some of the interactions, particularly in the last two years, are marginally significant and negative, showing that the effect of education has become stronger over time.

```
linearHypothesis(lm_68c1, c("y74educ=0", "y76educ=0", "y78educ=0", "y80educ=0", "y82educ=0", "y84educ=0
```

```
## Linear hypothesis test
##
## Hypothesis:
## y74educ = 0
## y76educ = 0
## y78educ = 0
## y80educ = 0
## y82educ = 0
## y84educ = 0
##
## Model 1: restricted model
## Model 2: kids ~ educ + age + agesq + black + east + northcen + west +
##     farm + othrural + town + smcity + y74 + y76 + y78 + y80 +
##     y82 + y84 + y74educ + y76educ + y78educ + y80educ + y82educ +
##     y84educ
```

```
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1111 2685.9
## 2   1105 2664.4  6    21.464 1.4836 0.1803
```

But the joint F test for the interaction terms yields p-value =. 180, and so we do not reject the model without the interactions. Still, the possibility that the link between fertility and education has become stronger over time is deserves attention, especially using more recent data.

```
lm_68c2= ivreg(kids ~ age+agesq+black+east+northcen+west+farm+othrural+town+smcity+educ+y74+y76+y78+y80
summary(lm_68c2)
```

```
##
## Call:
## ivreg(formula = kids ~ age + agesq + black + east + northcen +
##     west + farm + othrural + town + smcity + educ + y74 + y76 +
##     y78 + y80 + y82 + y84 + y74educ + y76educ + y78educ + y80educ +
##     y82educ + y84educ | age + agesq + black + east + northcen +
##     west + farm + othrural + town + smcity + y74 + y76 + y78 +
##     y80 + y82 + y84 + meduc + feduc + y74meduc + y76meduc + y78meduc +
##     y80meduc + y82meduc + y84meduc + y74feduc + y76feduc + y78feduc +
##     y80feduc + y82feduc + y84feduc, data = FERTIL1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3018 -1.0384 -0.0828  0.9876  4.6449
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.203119   3.391987  -2.713 0.006767 **
## age          0.514547   0.142993   3.598 0.000334 ***
## agesq       -0.005620   0.001614  -3.482 0.000517 ***
## black        1.090738   0.176431   6.182 8.88e-10 ***
## east         0.222614   0.136732   1.628 0.103788
## northcen     0.369538   0.124232   2.975 0.002998 **
## west         0.176224   0.171479   1.028 0.304329
## farm        -0.109485   0.154969  -0.706 0.480030
## othrural    -0.242762   0.185901  -1.306 0.191868
## town         0.094527   0.126213   0.749 0.454051
## smcity       0.196322   0.162003   1.212 0.225831
## educ         0.026893   0.096370   0.279 0.780252
## y74          1.723865   1.732284   0.995 0.319886
## y76          3.127236   1.802454   1.735 0.083022 .
## y78          4.340946   1.885580   2.302 0.021510 *
## y80          1.122166   1.590342   0.706 0.480578
## y82          2.668207   1.582737   1.686 0.092113 .
## y84          1.792424   1.713527   1.046 0.295770
## y74educ     -0.120057   0.140763  -0.853 0.393898
## y76educ     -0.264544   0.147040  -1.799 0.072272 .
## y78educ     -0.354725   0.150628  -2.355 0.018698 *
## y80educ     -0.101147   0.126609  -0.799 0.424527
## y82educ     -0.254079   0.124907  -2.034 0.042175 *
## y84educ     -0.189629   0.134215  -1.413 0.157976
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.567 on 1105 degrees of freedom
## Multiple R-Squared: 0.1206,  Adjusted R-squared: 0.1023
## Wald test: 5.922 on 23 and 1105 DF,  p-value: < 2.2e-16
```

Then test

```
linearHypothesis(lm_68c2, c("y74educ=0", "y76educ=0", "y78educ=0", "y80educ=0", "y82educ=0", "y84educ=0
```

```
## Linear hypothesis test
##
## Hypothesis:
## y74educ = 0
## y76educ = 0
## y78educ = 0
## y80educ = 0
## y82educ = 0
## y84educ = 0
##
## Model 1: restricted model
## Model 2: kids ~ age + agesq + black + east + northcen + west + farm +
##     othrural + town + smcity + educ + y74 + y76 + y78 + y80 +
##     y82 + y84 + y74educ + y76educ + y78educ + y80educ + y82educ +
##     y84educ | age + agesq + black + east + northcen + west +
##     farm + othrural + town + smcity + y74 + y76 + y78 + y80 +
##     y82 + y84 + meduc + feduc + y74meduc + y76meduc + y78meduc +
##     y80meduc + y82meduc + y84meduc + y74feduc + y76feduc + y78feduc +
##     y80feduc + y82feduc + y84feduc
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   1111
## 2   1105  6 8.4955      0.204
```

Qualitatively, the results are similar to the OLS estimates. The p-value for the joint F test on the interactions is 0.204, which has asymptotic justification under Assumption 2SLS.3, the homoskedasticity assumption - so again there is no strong evidence favoring including of the interactions of year dummies and education.