Introductory Econometrics
Endogeneity
Suggested Solution

by Hieu Nguyen

Fall 2024

1.

Suppose that you wish to estimate the effect of class attendance on student performance. A model to explain standardized outcome on a final exam (stndfnl) in terms of percentage of classes attended (attnd), prior college Grade Point Average (priGPA), and American College Testing score (ACT) is:

$$\text{stndfnl} = \beta_0 + \beta_1\text{attnd} + \beta_2\text{priGPA} + \beta_3\text{ACT} + \epsilon.$$

(a) Why might attnd be suspected to be endogenous in the model?

(b) Let dist be the distance from the students' living quarters to the lecture hall. Do you think dist is uncorrelated with ϵ?

(c) Assuming that dist and ϵ are uncorrelated, what other assumption must dist satisfy in order to be a good instrument for attnd?

(d) Suppose we add the interaction term priGPA $\cdot$ attnd to the model:

$$\text{stndfnl} = \beta_0 + \beta_1\text{attnd} + \beta_2\text{priGPA} + \beta_3\text{ACT} + \beta_4\text{priGPA} \cdot \text{attnd} + u.$$

If attnd is correlated with ϵ, then, in general, so is priGPA $\cdot$ attnd. What might be a good instrument candidate for priGPA $\cdot$ attnd?

Solution:

(a) We might be worried that attnd is correlated with other factors in ϵ. There are various potential reasons why. E.g., highly motivated students might attend more classes (omitted variable bias). Or some students need to have a part-time job (a piece of information not collected—thus unobservable—by the University) to cover their living expenses, so they might not have enough time to attend classes as well as to study sufficiently for the final exam (omitted influence bias). Hence the OLS regression of stndfnl on attnd may give us a poor estimate of the causal impact of attended classes.

(b) In a multiple regression model, various factors potentially correlated with dist (e.g., students from low-income families or exchange/foreign students may live off-campus) affecting stndfnl can be included directly in the model so that we can control for their potential impact. Then, dist can be reasonably (as it generally is as a typical textbook example) expected to be uncorrelated with ϵ, primarily if the campus rooms are assigned randomly (as is often a best practice at many universities).

(c) It needs to be correlated with attnd. Especially at large universities, some students commute to campus, which may increase the likelihood of missing lectures (bad weather, oversleeping, etc.). Or they may be lazy to commute, or they decide to study efficiently 'at home' instead of commuting. Thus attnd is expected to be negatively correlated with dist. This can be checked by regressing attnd on dist via a t-test.

(d) The idea is that class attendance might have a different effect on students who have performed differently in the past, as measured by priGPA. Thus we need an instrument that is not correlated with attnd but correlated with priGPA. As priGPA, ACT, and dist are assumed exogenous, then (technically) any function of these three variables is also exogenous so that we may think of, e.g., priGPA $\cdot$ dist. Or we may try to find a real-world variable correlated with priGPA. Or what about the level of education of one's mother and father (see seminar #8)?

2.

The data in `fertil2.gdt` includes, for a sample of women in Botswana during 1988, information on the number of children, years of education, age, and religious and economic status variables.

(a) Estimate this model by OLS and briefly comment on results:

$$\text{children} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{age} + \beta_3 \text{age}^2 + \epsilon.$$

If 100 women receive another year of education, how many fewer children are they expected to have?

(b) In lecture #10, we discussed why we might suspect educ to be endogenous in this model. We also suggested `frsthalf` (a dummy variable equal to one if the woman was born in the first six months of a year) to be a good candidate for an instrument for educ. Show its relevance via a first stage regression. Assume that `frsthalf` is uncorrelated with the error term ϵ. Now estimate the model from part (a) by using `frsthalf` as an instrument for educ (= IV estimator, 2SLS). Compare the estimated effect of education with the OLS estimate. Which of the estimators is consistent?

(c) Add the binary explanatory variables `electric`, `tv`, and `bicycle` to the model and assume these are exogenous as well. Estimate the equation by 2SLS directly in Gretl and compare the estimated coefficient of educ with part (b) and with the OLS estimate. Interpret the output of the Hausman test.

Solution:

(a) The model estimated by OLS is:

```
Model 1: OLS, using observations 1{4361
Dependent variable: children

Coefficient     Std. Error    t-ratio    p-value
------------------------------------------------
const           -4.13831      0.240594    -17.2004    0.0000
educ            -0.0905755    0.00592069 -15.2981    0.0000
age             0.332449      0.0165495   20.0882    0.0000
sq age          -0.00263082   0.000272592 -9.6511    0.0000

Mean dependent var  2.267828    S.D. dependent var  2.222032
Sum squared resid   9284.147    S.E. of regression  1.459746
R^2                0.568724   Adjusted R^2       0.568427
F (3, 4357)        1915.196   P-value(F )        0.000000
Log-likelihood     -7835.592  Akaike criterion   15679.18
Schwarz criterion  15704.71   Hannan{Quinn       15688.19
```

Following our expectation, educ has a significantly negative effect, and estimated coefficients of age and age^2 support its polynomial functional form. Interpretation of the model was discussed in detail during the seminar.

If we interpret the estimated effect of one additional year of educ literally, it suggests reducing the estimated number of children by 0.09. But this is impossible for any particular woman (children is a typical example of a discrete count variable). A standard economic interpretation is that average fertility decreases by 0.09 children given one more year of education. More reasonably worded: if 100 women receive another year of education, the estimated model suggests that there will be nine fewer children among this group in the future.

(b) To check the relevance of the `frsthalf` instrument and to run the 2SLS manually; we start with the first stage regression:

```
Model 2: OLS, using observations 1{4361
Dependent variable: educ

Coefficient     Std. Error    t-ratio    p-value
-------------------------------------------------
const           9.69286       0.598069   16.2069   0.0000
age             -0.107950     0.0420402  -2.5678   0.0103
sq age          -0.000505567  0.000692940 -0.7296  0.4657
frsthalf        -0.852285     0.112830   -7.5537   0.0000

Mean dependent var  5.855996    S.D. dependent var  3.927075
Sum squared resid   60001.14    S.E. of regression  3.710957
R^2                0.107651    Adjusted R^2        0.107037
F (3, 4357)        175.2068    P-value(F )         3.0e{107
Log-likelihood     -11904.53   Akaike criterion    23817.05
Schwarz criterion  23842.57    Hannan{Quinn        23826.06
```

We can see that `frsthalf` is a relevant instrument. It is correlated with the endogenous explanatory variable educ: Cov(educ, `frsthalf`) $\neq 0$ because it is statistically strongly significant in the first stage regression (t-test). But note that the R^2 of the model is rather small. Now save fitted values as `educ_hat2` (the fitted value from model 2), follow Save—Fitted values in the Gretl Model 1 menu.

We continue with the second stage regression:

```
Model 3: OLS, using observations 1{4361
Dependent variable: children

Coefficient     Std. Error    t-ratio    p-value
-------------------------------------------------
const           -3.38781      0.550340   -6.1558   0.0000
educ hat2       -0.171499     0.0533921  -3.2121   0.0013
age             0.323605      0.0179310  18.0473   0.0000
sq age          -0.00267228   0.000280805 -9.5165  0.0000

Mean dependent var  2.267828    S.D. dependent var  2.222032
Sum squared resid   9759.726    S.E. of regression  1.496667
R^2                0.546632    Adjusted R^2        0.546320
F (3, 4357)        1751.100    P-value(F )         0.000000
Log-likelihood     -7944.521   Akaike criterion    15897.04
Schwarz criterion  15922.56    Hannan{Quinn        15906.05
```

Contrasting the OLS and 2SLS estimated models, we observe a potential reduction of the positive OLS bias of educ—the estimated effect is now much more extensive (in the expected negative direction). The SE of the IV estimate for educ is also much bigger, about nine times! This produces a relatively wide 95

But be aware that the SEs and test statistics obtained this way are generally invalid. The reason is that the theoretical composite error term of the second stage model also includes the error term from the first stage. Still, the second stage SEs, if estimated 'manually,' are based on only residuals of the second stage (OLS does not know that we estimated the first stage before). It is thus preferred to run 2SLS jointly to account for residuals from both stages. This is an automated standard option in any regression package:

```
Model 4: TSLS, using observations 1{4361
Dependent variable: children

Instrumented: educ
Instruments: const frsthalf age sq age

Coefficient     Std. Error    z          p-value
-------------------------------------------------
const           -3.38781      0.548150    -6.1804   0.0000
educ            -0.171499     0.0531796   -3.2249   0.0013
age             0.323605      0.0178596   18.1194   0.0000
sq age          -0.00267228   0.000279687 -9.5545   0.0000

Mean dependent var  2.267828    S.D. dependent var  2.222032
Sum squared resid   9682.216    S.E. of regression  1.490712
R^2                 0.552676   Adjusted R^2        0.552368
F (3, 4357)         1765.119   P-value(F )         0.000000
Log-likelihood      -47917.57  Akaike criterion    95843.15
Schwarz criterion  95868.67    Hannan{Quinn        95852.15
```

Not much has changed, but now the SEs are valid, and we can use them for hypothesis testing.

To answer which of the estimators is consistent, we need to test whether educ is an endogenous variable, more generally, whether there is evidence of endogeneity in the data. We know that 2SLS is consistent in both cases but inefficient if there is no endogeneity. On the other hand, OLS is inconsistent under endogeneity. We use the Hausman test, which output is automatically attached to the automated Gretl 2SLS regression:

```
Hausman test {
Null hypothesis: OLS estimates are consistent
Asymptotic test statistic: 2(1) = 2.4501
with p-value = 0.117517
```

Hypotheses vaguely:

$$H_0 : \text{OLS estimator is consistent vs } H_A : \text{OLS estimator is inconsistent},$$

$$H_0 : \text{no endogeneity vs } H_A : \text{endogeneity}.$$

Hausman (Wald) test statistic:

$$H(orW) = (\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})'(\mathrm{Var}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}))^{-1}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}) \sim \chi^2_{k+1}.$$

The critical value for the 'default' 5% significance level and four d.f.: 9.49 (would be 3.84 for one d.f.) ¿ 2.45 ⇒ in any case, we do not reject the H_0 ⇒ the consistency of the OLS estimator is not rejected at 5%. Hence, it should be used preferably because it is efficient.

This result does not necessarily mean (and definitely not prove) that educ is exogenous. The p-value suggests that we would reject H_0 at the 12% significance level, so our conclusion in the previous paragraph is not very statistically strong. Because the sample size is large, we might rather suspect insufficient explanatory power of the instrument (remind the not very large R^2 of the first stage model and see also sample correlation between educ and `frsthalf`) to cause this situation. There might be other not yet considered omitted variables/influences in the model causing additional bias, maybe in the opposite direction, thus decreasing the power of the Hausman test to recognize endogeneity. Finally, a linear model might not be the correct functional form for the discrete count nature of the dependent variable. Finally, the discrete count nature of the dependent variable is likely to be a problematic issue for linear regression (will be discussed in lecture #11). All these aspects might decrease the overall power of the test.

Advice about what to do when there is uncertainty as to whether an explanatory variable is endogenous or not is somewhat mixed. The prevailing attitude is probably summarized by Wooldridge (2010) who suggests: "We find evidence of endogeneity of ... at the 10% significance level against a two-sided alternative, and so 2SLS is probably a good idea (assuming that we trust the instruments.)". Moreover, Guggenberger (2010) advises that if testing the coefficient of the endogenous explanatory variable is the objective, we should avoid considering the Hausman test result and use 2SLS.

(c) Model estimated by the 2SLS routine:

```
Model 5: TSLS, using observations 1{4361 (n = 4356)
Missing or incomplete observations dropped: 5

Dependent variable: children
Instrumented: educ

Instruments: const frsthalf age sq age electric tv bicycle

Coefficient     Std. Error    z          p-value
-------------------------------------------------
const           -3.59133      0.645089   -5.5672   0.0000
educ            -0.163981     0.0655269  -2.5025   0.0123
age             0.328145      0.0190587  17.2176   0.0000
sq age          -0.00272216   0.000276559 -9.8430  0.0000
electric        -0.106531     0.165965   -0.6419   0.5209
tv              -0.00255501   0.209230   -0.0122   0.9903
bicycle         0.332072      0.0515264  6.4447    0.0000

Mean dependent var  2.268365    S.D. dependent var  2.222073
Sum squared resid   9511.714    S.E. of regression  1.478886
R^2                0.559569    Adjusted R^2        0.558961
F (6, 4349)        921.7086    P-value(F )         0.000000
Log-likelihood     -47487.53   Akaike criterion    94989.05
Schwarz criterion  95033.71    Hannan{Quinn        95004.82
```

Hausman test – Null hypothesis: OLS estimates are consistent Asymptotic test statistic: 2(1) = 1.87295 with p-value = 0.171137

The interpretation of the Hausman test is precisely the same as in the previous part, with similar caveats. 'No endogeneity' H_0 is technically not rejected. We thus compare the results with a model estimated by OLS:

```
Model 6: OLS, using observations 1{4361 (n = 4356)
Missing or incomplete observations dropped: 5

Dependent variable: children

Coefficient     Std. Error    t-ratio    p-value
-------------------------------------------------
const           -4.38978      0.240317   -18.2666  0.0000
educ            -0.0767093    0.00635259 -12.0753  0.0000
age             0.340204      0.0164417  20.6915   0.0000
sq age          -0.00270808   0.000270551 -10.0095 0.0000
```

```
electric          -0.302729     0.0761869  -3.9735    0.0001
tv                -0.253144     0.0914374  -2.7685    0.0057
bicycle           0.317895      0.0493661  6.4395     0.0000

Mean dependent var  2.268365    S.D. dependent var  2.222073
Sum squared resid   9116.101    S.E. of regression  1.447804
R^2                0.576060    Adjusted R^2        0.575475
F (6, 4349)        984.9211    P-value(F )         0.000000
Log-likelihood     -7789.323   Akaike criterion    15592.65
Schwarz criterion  15637.30    Hannan{Quinn        15608.41
```

Adding `electric`, `tv`, and `bicycle` to the model slightly reduces the estimated effect of educ in both cases. In the model estimated by OLS, the coefficient on `tv` implies that other factors fixed, four families that own a television will have about one fewer child than four families without a TV. A causal interpretation is that TV provides an alternative form of recreation.

A comparison of the models and a general idea behind the inclusion of new variables was discussed in detail during the seminar (e.g., only 14% of women have electricity in their homes, which can be considered as a proxy for good economic background of the family; television ownership can be a proxy for different things, including income and perhaps geographic location, what about an intuitive logic of the bicycle, which is strongly statistically significant in both models?).

Interestingly, the effects of `electric` and `tv` substantially drop in the model estimated by 2SLS. This supports our suspicion about the functional form of the model.

3.

A researcher estimated by OLS two specifications of a regression model:

$$y = \alpha + \beta x_1 + \epsilon,$$

$$y = \tilde{\alpha} + \tilde{\beta} x_1 + \tilde{\gamma} x_2 + \tilde{\epsilon}$$

Explain theoretically under what circumstances the following will be true. If some case cannot be true, explain why.

(a) $\hat{\beta} = \hat{\tilde{\beta}}$.

(b) β is statistically significant (at the 5

(c) $\tilde{\beta}$ is statistically significant (at the 5% level) but β is not.

(d) If $\hat{\epsilon}_i$ and $\hat{\tilde{\epsilon}}_i$ are the estimated residuals from the two equations,

$$\sum_{i=1}^{n} \hat{\epsilon}_i{}^2 \geq \sum_{i=1}^{n} \hat{\tilde{\epsilon}}_i^2.$$

Solution:

(a) The estimated coefficients from the two models will be the same if, by omitting x_2, we do not cause any bias of the estimator of β. This happens either if $\gamma = 0$ (x_2 is an irrelevant variable) or when x_1 and x_2 are not correlated. Hence,

$$\hat{\beta} = \hat{\beta} \Leftrightarrow \text{Cov}(x_1, x_2) = 0 \text{ or } \gamma = 0.$$

(b) This will typically happen when x_1 and x_2 are highly correlated. When x_2 is not included in the model, most of its explanatory power is attributed to x_1, and its significance can be overestimated. When x_2 is included, the explanatory power is diluted between the two variables, and x_1 can lose its significance (multicollinearity problem).

(c) This can happen in the situation when the second model is correct (x_2 should be included, i.e., $\gamma \neq 0$). In this case, x_2 is an omitted variable in the first model, and we experience an omitted variable bias. If the omission of x_2 biases the OLS estimator of β enough in the direction towards zero, we are likely to observe the impact of the bias in the numerical value of $\hat{\beta}$ and β can thus become statistically insignificant.

(d) We can consider the two models as a restricted and an unrestricted version of the second model. We know from lectures that the inequality below always holds:

$$\mathrm{RSS}_R \geq \mathrm{RSS}_U.$$