Test Exercise 4

Coursera/Erasmus U., Econometric Methods and Applications

Anthony Nguyen

Notes:

- See website for how to submit your answers and how feedback is organized.
- This exercise uses the datafile `TestExer4_Wage` and requires a computer.
- The dataset `TestExer4_Wage` is available on the website.

Goals and skills being used:

- Obtain insight in consequences of endogeneity
- Practice with identifying causes of endogeneity
- Practice with identifying valid instruments
- Obtain hands-on experience with applying 2SLS and the Sargan test

Questions

A challenging and very relevant economic problem is the measurement of the returns to schooling. In this question, we will use the following variable on 3,010 US men:

- `logw`: log wage
- `educ`: number of years of schooling
- `age`: age of the individual in years
- `exper`: working experience in years
- `smsa`: dummy indicating whether the individual lived in a metropolitan area
- `south`: dummy indicating whether the individual lived in the south
- `nearc`: dummy indicating whether the individual lived near a 4-year college
- `dadeduc`: education of the individual's father (in years)
- `momeduc`: education of the individual's mother (in years)

This data is a selection of the data used by D. Card (1995)[1]

[1] "Using Geographic Variation in College Proximity to Estimate the Return to Schooling". In L.N. Christofides, E.K. Grant, and R. Swidinsky, editors, Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp. Toronto: University of Toronto Press, 1995.

(a) Use OLS to estimate the parameters of the model:

$$logw = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 smsa + \beta_6 south + \varepsilon.$$

Give an interpretation to the estimated β_2 coefficient.

(b) OLS may be inconsistent in this case as `educ` and `exper` may be endogenous. Give a reason why this may be the case. Also indicate whether the estimate in part (a) is still useful.

(c) Give a motivation why `age` and `age^2` can be used as instruments for `exper` and `exper^2`.

(d) Run the first-stage regression for `educ` for the two-stage least squares estimation of the parameters in the model once when `age`, `age^2`, `nearc`, `dadeduc`, and `momeduc` are used as additional instruments. What do you conclude about the suitability of these instruments for schooling?

(e) Estimate the parameters of the model for log wage using two-stage least squares where you correct for the endogeneity of education **and** experience. Compare your result to the estimate in part (a).

(f) Perform the Sargan test for validity of the instruments. What is your conclusion?

Answers

(a) Use OLS to estimate the parameters of the model and give an interpretation to the estimated β_2 coefficient.:

$$logw = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 smsa + \beta_6 south + \varepsilon.$$

```
# Add exper^2 to data frame
TestExer4 <- TestExer4 %>% mutate(exper2 = exper^2)

# Regress logw
mod1 <- lm(logw~educ+exper+exper2+smsa+south, data = TestExer4)

# Print regression results
summary(mod1)
```

```
##
## Call:
## lm(formula = logw ~ educ + exper + exper2 + smsa + south, data = TestExer4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71487 -0.22987  0.02268  0.24898  1.38552
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.6110144  0.0678950  67.914  < 2e-16 ***
## educ         0.0815797  0.0034990  23.315  < 2e-16 ***
## exper        0.0838357  0.0067735  12.377  < 2e-16 ***
## exper2      -0.0022021  0.0003238  -6.800 1.26e-11 ***
## smsa         0.1508006  0.0158360   9.523  < 2e-16 ***
## south       -0.1751761  0.0146486 -11.959  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3813 on 3004 degrees of freedom
## Multiple R-squared:  0.2632, Adjusted R-squared:  0.2619
## F-statistic: 214.6 on 5 and 3004 DF,  p-value: < 2.2e-16
```

From the regression results, we can see that the β_2 coefficient is equal to 0.0815797. In other words, for each additional year of education, the log wage earned increases by 0.082.

We can convert this into a percentage term by taking the exponent, which leads us to say that, for each additional year of education, the wage earned increases by 8.5%.

(b) OLS may be inconsistent in this case as `educ` and `exper` may be endogenous. Give a reason why this may be the case. Also indicate whether the estimate in part (a) is still useful.

A variable is said to be endogenous if it is correlated with the error term. In other words, there is some omitted factor within the error term, ε, that is not included in our model that affects both y and X. If this is the case, then OLS does not properly estimate β, and our model would be biased, inconsistent and no longer useful.

In the example given here, `educ` and `exper` may be endogenous as there are a number of other factors not included in our model that could affect both of these independent variable, as well as the outcome/dependent variable. For example, intelligence, ability or socio-economic status.

(c) Give a motivation why `age` and `age^2` can be used as instruments for `exper` and `exper^2`.

A valid instrument must satisfy two conditions: namely, (1) that the instrument (Z) and the endogenous variable (X) are correlated, and (2) that the instrument (Z) does not correlate with the error term (ε).

With that said, `age` can be used as an instrument for `exper` as it is correlated with experience (the endogenous variable) and it is also clearly exogenous (determined outside of the model) and therefore not correlated to the error term.

(d) Run the first-stage regression for `educ` for the two-stage least squares estimation of the parameters in the model once when `age`, `age^2`, `nearc`, `dadeduc`, and `momeduc` are used as additional instruments. What do you conclude about the suitability of these instruments for schooling?

```
# Add age^2 to dataframe
TestExer4 <- TestExer4 %>% mutate(age2 = age^2)

# Regress educ
mod2 <- lm(educ~age+age2+nearc+daded+momed+smsa+south, data = TestExer4)

# Save fitted values
educ_fit <- fitted(mod2)

# Print regression results
summary(mod2)
```

```
##
## Call:
## lm(formula = educ ~ age + age2 + nearc + daded + momed + smsa +
##     south, data = TestExer4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2777  -1.5450  -0.2224   1.6957   7.2250
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.652354   3.976343  -1.421 0.155277
## age          0.989610   0.278714   3.551 0.000390 ***
## age2        -0.017019   0.004838  -3.518 0.000441 ***
## nearc        0.264554   0.099085   2.670 0.007626 **
## daded        0.190443   0.015611  12.199  < 2e-16 ***
## momed        0.234515   0.017028  13.773  < 2e-16 ***
## smsa         0.529566   0.101504   5.217 1.94e-07 ***
## south       -0.424851   0.091037  -4.667 3.19e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.326 on 3002 degrees of freedom
```

```
## Multiple R-squared:  0.2466, Adjusted R-squared:  0.2448
## F-statistic: 140.4 on 7 and 3002 DF,  p-value: < 2.2e-16
```

Looking at the results of this regression, the variables `age`, `age^2`, `nearc`, `dadeduc`, and `momeduc` are all significantly correlated with `educ`, and would be suitable instruments for `educ`.

(e) Estimate the parameters of the model for log wage using two-stage least squares where you correct for the endogeneity of education and experience. Compare your result to the estimate in part (a).

We already performed the first-stage regression and saved the fitted values for `educ`, so now we just need to do the same for `exper` and `exper2` before we can proceed to the second-stage:

```
# Perform 1st stage regression for `exper`
mod3 <- lm(exper~age+age2+daded+momed+smsa+south+nearc,data = TestExer4)

# Save fitted values for `exper`$
exper_fit <- fitted(mod3)

# Perform 1st stage regression for `exper2`
mod4 <- lm(exper2~age+age2+daded+momed+smsa+south+nearc,data = TestExer4)

# Save fitted values for `exper2`
exper2_fit <- fitted(mod4)
```

Now that we have the fitted values for the three endogenous variables we are replacing with instruments, we can run the second stage regression as follows:

```
# Add fitted values to dataframe
TestExer4 <- TestExer4 %>% mutate(educ_fit = educ_fit, exper_fit = exper_fit, exper2_fit = exper2_fit)

# Perform 2nd stage regression
mod5 <- lm(logw~educ_fit+exper_fit+exper2_fit+smsa+south, data = TestExer4)

# Print regression results
summary(mod5)
```

```
##
## Call:
## lm(formula = logw ~ educ_fit + exper_fit + exper2_fit + smsa +
##     south, data = TestExer4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67797 -0.23820  0.01715  0.26700  1.46756
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.4169039  0.1178606  37.476  < 2e-16 ***
## educ_fit     0.0998429  0.0067128  14.874  < 2e-16 ***
## exper_fit    0.0728669  0.0170667   4.270 2.02e-05 ***
## exper2_fit  -0.0016393  0.0008559  -1.915   0.0555 .
```

```
## smsa         0.1349370  0.0171240   7.880 4.54e-15 ***
## south       -0.1589869  0.0160170  -9.926  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3925 on 3004 degrees of freedom
## Multiple R-squared:  0.2192, Adjusted R-squared:  0.2179
## F-statistic: 168.6 on 5 and 3004 DF,  p-value: < 2.2e-16
```

We can compare our results with what we obtained in part (a) with the following table:

Table 1:

	Dependent variable:	
	logw	
	(1)	(2)
educ	0.082^{***}	
	(0.003)	
exper	0.084^{***}	
	(0.007)	
exper2	-0.002^{***}	
	(0.0003)	
educ_fit		0.100^{***}
		(0.007)
exper_fit		0.073^{***}
		(0.017)
exper2_fit		-0.002^{*}
		(0.001)
smsa	0.151^{***}	0.135^{***}
	(0.016)	(0.017)
south	-0.175^{***}	-0.159^{***}
	(0.015)	(0.016)
Constant	4.611^{***}	4.417^{***}
	(0.068)	(0.118)
R^2	0.263	0.219
Note:	*p<0.1; **p<0.05; ***p<0.01	

Here we can see that the overall picture of the two models is quite similar. The signs of the coefficients remain the same between the original and the fitted estimators, all of them are significant (though the strength of the `exper2_fit` goes down quite a bit), and differences between original and fitted estimators is quite small in all cases. We can also see that the R^2 value in the 2SLS model goes down a little when compared to the OLS model.

(f) Perform the Sargan test for validity of the instruments. What is your conclusion?

To perform the Sargan test, we need use the residuals from our 2SLS, which we can derive by using the coefficients from our second-stage regression ($b_{2\text{SLS}}$) and plugging them back into our original dataset, so that we have:

$$e_{2\text{SLS}} = y - X \cdot b_{2\text{SLS}}$$

From our example here, we would have:

$$e_{2\text{SLS}} = logw - (\beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 smsa + \beta_6 south)$$

```
# Use the attach() function to make it easier to call up all the column names in the following formula
attach(TestExer4)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     educ_fit, exper_fit, exper2_fit
```

```
# Plug in 2nd-stage coefficients to calculate 2SLS residuals
#
# Note that it is important to use the full values here, as using rounded numbers
# will lead to incorrect results when calculating test statistics later on.
e_2SLS <- logw - (4.416903899 + 0.099842919*educ + 0.072866858*exper - 0.001639293*exper2 + 0.134937031

# Add residuals vector to our dataframe for later use
TestExer4 <- TestExer4 %>% mutate(e_2SLS = e_2SLS)

# detach dataframe
detach(TestExer4)
```

Next we regress our 2SLS residual term ($e_{2\text{SLS}}$) by our instruments (Z)

```
# Regress 2SLS residuals by instruments
mod6 <- lm(e_2SLS~age+age2+daded+momed+smsa+south+nearc, data = TestExer4)

summary(mod6)
```

```
##
## Call:
## lm(formula = e_2SLS ~ age + age2 + daded + momed + smsa + south +
##     nearc, data = TestExer4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77698 -0.23327  0.02731  0.25039  1.34302
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1258214  0.6568119   0.192    0.848
## age         -0.0093315  0.0460379  -0.203    0.839
## age2         0.0001591  0.0007991   0.199    0.842
## daded       -0.0041052  0.0025787  -1.592    0.111
```

```
## momed        0.0041134  0.0028126   1.462    0.144
## smsa        -0.0033465  0.0167664  -0.200    0.842
## south        0.0022260  0.0150375   0.148    0.882
## nearc        0.0135079  0.0163668   0.825    0.409
##
## Residual standard error: 0.3843 on 3002 degrees of freedom
## Multiple R-squared:  0.00123,    Adjusted R-squared:  -0.001099
## F-statistic: 0.5282 on 7 and 3002 DF,  p-value: 0.8138
```

Next we can calculate the Sargan test statistic by multiplying the R^2 value obtained in this regression with the number of observations (n) in the dataset.

```
Sargan <- summary(mod6)$r.squared * nrow(TestExer4)
```

Finally, we can check the p-value of our test statistic by using the $\chi^2(m-k)$ distribution, where m is the number of instruments in Z, and k is the number of explanatory variables in X.

```
# Determine p-value of Sargan test using chi_sq distribution
Sargan_pval <- 1-pchisq(Sargan, 2)

print(paste("Sargan value =", round(Sargan,3)))
```

```
## [1] "Sargan value = 3.702"
```

```
print(paste("Sargan p-value =", round(Sargan_pval, 5)))
```

```
## [1] "Sargan p-value = 0.15705"
```

The null hypothesis (H_0) in the Sargan test is that our instruments are *not* correlated with the error term, and here, the p-value is too large to reject the null, meaning that our instruments are not correlated with the error term and hence, they are valid.