ECONOMETRICS FINAL EXAM 2023-24: SOLUTIONS PART A

Universidad Carlos III de Madrid

Answer all questions in 50 minutes

Use the **wage2** dataset from Wooldridge to estimate the returns to education for men using the following model,

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 black + \beta_6 south + \beta_7 urban + u, \quad (1)$$

where *wage* are monthly earnings in dollars, *educ* are the years of education, *exper* are the years of work experience, *tenure* are the years with the current employer, *married* is a binary variable equal to 1 if the worker is married, *black* is a binary variable equal to 1 if the worker is black and *south* and *urban* are also binary variables for living in the south and in a metropolitan area, respectively.

It is argued that *educ* is likely to be endogenous in model (1) because worker's ability is omitted from this equation, while we can take the rest of regressors as exogenous.

1. (40%) The variables *meduc* (years of education of the mother) and *brthord* (birthday order: one for the first child, two for the second one, etc.) are proposed as instruments for *educ*. Argue why *meduc* and *brthord* could be valid instruments for *educ* (10%). Check the relevance of the instruments and explain why is important (15%). Explain how would you check the exogeneity of the proposed instruments and interpret the result of the test (15%).

   .

   ANSWER:

   [10%] We expect *meduc* to be partially correlated with *educ* as more educated mothers typically have children with higher education levels keeping all the other individuals' (exogenous) characteristics fixed (relevance), while we could think that once we control for the actual level of education of the worker, mother's education has no direct impact on wages and is not correlated to omitted ability (exogeneity, $\text{Cov}(meduc, u) = 0$).

   The variable *brthord* could be related to *educ* as the older children may receive more resources than younger ones to attend higher education for some families (relevance), while the birth order might not be related to intrinsic ability omitted in the equation (exogeneity, $\text{Cov}(brthord, u) = 0$). However, it could be also argued that older children receive more attention from parents, increasing their ability, and violating the exogeneity condition as *brthord* would be potentially correlated to u.

   .

   [15%] To check the relevance of $(meduc, brthord)$ we should use the F statistic testing whether in the Reduced Form of *educ*, i.e. the first stage of TSLS,

   $$educ = \pi_0 + \pi_1 meduc + \pi_2 brthord + \pi_3 exper + \pi_4 tenure + \pi_5 married + \pi_6 black + \pi_7 south + \pi_8 urban + u.$$

the two instruments have nonzero coefficients by testing the hypothesis

$$\begin{aligned} H_0 &: \pi_1 = \pi_2 = 0 \\ H_1 &: \pi_1 \neq 0 \text{ and/or } \pi_2 \neq 0 \end{aligned}$$

in the OLS estimation of the Reduced Form of *educ*.

To obtain the F statistic we can use directly GRETL TSLS estimation output

$$\widehat{\text{lwage}} = \underset{(0.34619)}{4.30952} + \underset{(0.021405)}{0.133078}\,\text{educ} + \underset{(0.0061604)}{0.0317955}\,\text{exper} + \underset{(0.0028796)}{0.00659002}\,\text{tenure}$$
$$+ \underset{(0.047683)}{0.204236}\,\text{married} - \underset{(0.046690)}{0.152152}\,\text{black} - \underset{(0.030454)}{0.0582469}\,\text{south} + \underset{(0.032036)}{0.174541}\,\text{urban}$$

Hausman test –

Null hypothesis: OLS estimates are consistent
Asymptotic test statistic: $\chi^2(1) = 9.72835$
with p-value = 0.00181447

Sargan over-identification test –
Null hypothesis: all instruments are valid
Test statistic: LM = 0.172964
with p-value = $P(\chi^2(1) > 0.172964) = 0.677491$

Weak instrument test –
First-stage $F(2, 775) = 44.3177$

.

or, alternatively, estimate explicitly the actual reduced form regression,

$$\widehat{\text{educ}} = \underset{(0.46143)}{14.1242} - \underset{(0.015737)}{0.221153}\,\text{exper} + \underset{(0.013347)}{0.0375681}\,\text{tenure} - \underset{(0.22899)}{0.186191}\,\text{married}$$
$$- \underset{(0.19594)}{0.443012}\,\text{black} - \underset{(0.14438)}{0.000405493}\,\text{south} + \underset{(0.14788)}{0.340227}\,\text{urban} - \underset{(0.042981)}{0.127550}\,\text{brthord}$$
$$+ \underset{(0.025152)}{0.186792}\,\text{meduc}$$
$$T = 784 \quad \bar{R}^2 = 0.3087 \quad F(8, 775) = 54.718 \quad \hat{\sigma} = 1.8307$$

F Test on the Reduced Form of *educ*

Null hypothesis: the regression parameters are zero for the variables $brthord, meduc$

Test statistic: **Robust F(2, 775) = 44.3177**, p-value 5.97541e-19

.

The First Stage F statistic should be larger than 10 to rule out weak instruments in the estimated Reduced Form. Then, because $F = 44.3177 > 10$, the instruments are not weak and therefore are relevant, guaranteeing that the properties of the TSLS estimates are well approximated for large samples by the normal distribution provided by the theory.

.

[15%] Since there is overidentification ($m > k$) because there are more instruments ($m = 2$: *meduc* and *brthord*) than endogenous regressors ($k = 1 : educ$), we can check the instruments exogeneity assumption using the J-test of overidentifying restrictions or the Sargan test provided by the TSLS output.

The J statistic is obtained by running the regression

$$\hat{u} = \delta_0 + \delta_1 meduc + \delta_2 brthord + \delta_2 exper + \delta_3 tenure + \delta_4 married + \delta_5 black + \delta_6 south + \delta_7 urban + u$$

on the TSLS residuals $\hat{u}$, and computing the (robust) F-statistic for the significance of the instruments, i.e. the hypothesis

$$\begin{aligned} H_0 &: \delta_1 = \delta_2 = 0 \\ H_1 &: \delta_1 \neq 0 \text{ and/or } \delta_2 \neq 0, \end{aligned}$$

where the J statistic is

$$J = mF = 2 \cdot 0.0862917 = 0.172\,58$$

using the number of instruments $m = 2$ and the value of the $F = 0.0862917$ statistic.

The J statistic is compared to the 5% critical value from the $\chi^2_{m-k} = \chi^2_1$ distribution equal to $3.84 = 1.96^2$, and since $J < 3.84$ (or equivalently, Sargan's $LM = 0.172964 < 3.84$) we cannot reject the null hypothesis of exogeneity of the instruments, confirming the validity of the instruments.

Not needed to get full marks: The actual TSLS residuals regression for the J test is

$$\begin{aligned} \widehat{u} = & \underset{(0.085187)}{-0.0209885} + \underset{(0.0091699)}{0.00350205}\,\text{brthord} + \underset{(0.0046340)}{0.00115334}\,\text{meduc} + \underset{(0.0032660)}{3.86469\text{e–}05}\,\text{exper} \\ & + \underset{(0.0028066)}{4.71381\text{e–}06}\,\text{tenure} + \underset{(0.047979)}{0.000402907}\,\text{married} + \underset{(0.044715)}{0.000238648}\,\text{black} - \underset{(0.030498)}{0.000126937}\,\text{south} \\ & - \underset{(0.030143)}{3.14898\text{e–}05}\,\text{urban} \end{aligned}$$

Test on Model 2:

Null hypothesis: the regression parameters are zero for the variables: $brthord, meduc$

Test statistic: Robust F(2, 775) = 0.0862917, p-value 0.917335

.

Note that the reported p-value of this F test is incorrect because it uses the wrong degrees of freedom.

2. (25%) Construct a 95% confidence interval for the effect on wages of a worker getting one more year of experience and one more year of tenure, everything else fixed, using the estimation results of the previous question (15%). Interpret the results (10%).

.

ANSWER:

[15%] The effect of interest is

$$\begin{aligned}\Delta \ln(wage) &= \mathbb{E}[\ln(wage)\,|exper+1, tenure+1] - \mathbb{E}[\ln(wage)\,|exper, tenure] \\ &= \beta_2 + \beta_3,\end{aligned}$$

or in percentage terms,

$$\begin{aligned}\%\Delta wage &\approx 100 \cdot \Delta \ln(wage)\,\% \\ &= 100 \cdot (\mathbb{E}[\ln(wage)\,|exper+1, tenure+1] - \mathbb{E}[\ln(wage)\,|exper, tenure])\,\% \\ &= 100\,(\beta_2 + \beta_3)\,\%,\end{aligned}$$

which is estimated using TSLS by

$$100\left(\hat{\beta}_2 + \hat{\beta}_3\right) = 100 \cdot (0.0317955 + 0.00659002)\,\% = 3.838\,6\%,$$

i.e. wages increase on average approximately by 3.8386% when gaining more one year of experience and of tenure.

To compute the confidence interval for $\beta_2 + \beta_3$ we need the Standard Error for the sum of the TSLS estimates $\hat{\beta}_2 + \hat{\beta}_3$:

$$\begin{aligned}SE\left(\hat{\beta}_2 + \hat{\beta}_3\right) &= \sqrt{\widehat{Var}\left(\hat{\beta}_2\right) + \widehat{Var}\left(\hat{\beta}_3\right) + 2\widehat{Cov}\left(\hat{\beta}_2, \hat{\beta}_3\right)} \\ &= \sqrt{SE\left(\hat{\beta}_2\right)^2 + SE\left(\hat{\beta}_3\right)^2 + 2\widehat{Cov}\left(\hat{\beta}_2, \hat{\beta}_3\right)} \\ &= \sqrt{0.0061604^2 + 0.0028796^2 + 2 \cdot (-5.99500 \cdot 10^{-6})} \\ &= 5.852\,6 \times 10^{-3}\end{aligned}$$

and the 95% confidence interval is then

$$0.03838\,6 \pm 1.96 \times 5.852\,6 \times 10^{-3} = \left(2.691\,5 \times 10^{-2}, 4.985\,7 \times 10^{-2}\right).$$

[10%] To interpret the effect, in percentage terms the confidence interval is

$$(2.6915\%,\ 4.9857\%)$$

so the average increment of wages due the extra year in experience and tenure is (approximately) between 2.7% and 5% with a 95% confidence.

3. (35%) Modify the model (1) so that the returns to education can depend on the ethnic origin of the worker and test for this dependence (15%). Discuss whether the new model is under/over-identified and if you could test that the instruments of question 1. (*meduc* and *brthord*) are valid and non weak for the new model (10%). Test whether wages depend on the ethnic origin of the worker (10%).

.

ANSWER:

[15%] To allow for the returns on education to depend on the ethnic origin, model (1) has to be modified including the interaction between *educ* and *black* :

$$\begin{aligned}\log(wage) &= \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 black \\ &\quad + \beta_6 south + \beta_7 urban + \beta_8 educ \cdot black + u,\end{aligned} \quad (2)$$

so that the partial effect of *educ* on $\log(wage)$ is now $\beta_1 + \beta_8 black$, which depends on ethnic origin if $\beta_8 \neq 0$. Then to test

$$\begin{aligned}H_0 &: \beta_8 = 0 \\ H_1 &: \beta_8 \neq 0\end{aligned}$$

we can use a significance t-test from the updated TSLS estimation

$$\begin{aligned}\widehat{\text{lwage}} = &\underset{(0.61752)}{4.09144} + \underset{(0.043935)}{0.149253}\,\text{educ} + \underset{(0.0063334)}{0.0324834}\,\text{exper} + \underset{(0.0030926)}{0.00628784}\,\text{tenure} \\ &+ \underset{(0.049966)}{0.197125}\,\text{married} + \underset{(4.0919)}{1.50577}\,\text{black} - \underset{(0.034307)}{0.0507960}\,\text{south} + \underset{(0.035852)}{0.167536}\,\text{urban} \\ &- \underset{(0.32008)}{0.129621}\,\text{educ} \cdot \text{black}\end{aligned}$$

using instruments (*meduc* and *brthord*) and presuming that both *educ* and $educ \cdot black$ are endogenous, i.e.

$$t(\beta_8) = \frac{\hat{\beta}_8}{SE\left(\hat{\beta}_8\right)} = \frac{-0.129621}{0.32008} = -0.404\,96,$$

which is not significative at the usual levels, not rejecting H_0 and concluding that there is no empirical evidence of the effect of education on wages depending on ethnic origin.

.

[10%] For the identification analysis we expect that the interaction $educ \cdot black$ in model (2) is also endogenous as is *educ*. Now there are $k = 2$ endogenous regressors while we have only $m = 2$ instruments, so the new model is just identified, and still could be estimated by TSLS, but we can not run a test of overidentifying restrictions.

However, despite a real checking of relevance involves coefficients of the reduced forms of both endogenous regressors, we could still run individual F tests to check the instruments relevance for *educ* (exactly the same test as before, since the list of exogenous variables has not changed), and for the interaction $educ \cdot black$. This last one is similar to that for $educ$ using its reduced form

$$\begin{aligned}(educ \cdot black) &= \alpha_0 + \alpha_1 meduc + \alpha_2 brthord + \alpha_3 exper + \alpha_4 tenure + \alpha_5 married \\ &\quad + \alpha_6 black + \alpha_7 south + \alpha_8 urban + u,\end{aligned}$$

which is estimated by OLS as

$$\begin{aligned}\widehat{\text{educblack}} = &\underset{(0.16052)}{0.242048} - \underset{(0.018095)}{0.0429345}\,\text{brthord} + \underset{(0.0092754)}{0.0144118}\,\text{meduc} - \underset{(0.0056914)}{0.0225887}\,\text{exper} \\ &+ \underset{(0.0042490)}{0.00232049}\,\text{tenure} - \underset{(0.10348)}{0.0812026}\,\text{married} + \underset{(0.19914)}{12.7334}\,\text{black} + \underset{(0.051332)}{0.0584107}\,\text{south} \\ &- \underset{(0.028731)}{0.0113449}\,\text{urban}\end{aligned}$$

where the First Stage F test statistic for

$$\begin{aligned}H_0 &: \alpha_1 = \alpha_2 = 0 \\ H_1 &: \alpha_1 \neq 0 \text{ and/or } \alpha_2 \neq 0,\end{aligned}$$

Test on Model 3:

Null hypothesis: the regression parameters are zero for the variables: $(brthord, meduc)$

Test statistic: Robust F(2, 775) = 5.3826, p-value 0.00476931

.

indicates that the instruments are weak for the new interaction regressor because $F < 10$, and the TSLS estimates of model (2) are not very reliable.

.

[10%] For testing whether wages depend on ethnic origin, we need to test the joint significance of the variables $black$ and $educ \cdot black$ in model (2):

$$\begin{aligned}H_0 &: \beta_5 = \beta_8 = 0 \\ H_1 &: \beta_5 \neq 0 \text{ and/or } \beta_8 \neq 0,\end{aligned}$$

using a robust F test in the TSLS output:

.

Test on Model (2):

Null hypothesis: the regression parameters are zero for the variables $(black, educblack)$

Test statistic: Robust F(2, 775) = 5.44754, p-value 0.00447341

.

and comparing $F = 5.44754$ with the 5% critical value from a $\chi^2_2/2$ distribution equal to $\chi^2_{2,0.05}/2 = 3$, we reject the null hypothesis as $F > \chi^2_{2,0.05}/2$ (or by checking that the p-value is less than 0.05), confirming that wages do depend on ethnic origin.

.

.

SOME CRITICAL VALUES: $Z_{0.10} = 1.282$, $Z_{0.05} = 1.645$, $Z_{0.025} = 1.96$, $\chi^2_{1,0.05} = 3.841$, $\chi^2_{1,0.01} = 6.635$, $\chi^2_{2,0.05} = 5.991$, $\chi^2_{2,0.01} = 9.210$, $\chi^2_{3,0.05} = 7.815$, $\chi^2_{3,0.01} 11.345$, $\chi^2_{4,0.05} = 9.488$, $\chi^2_{4,0.01} = 13.277$, $\chi^2_{5,0.05} = 11.071$, $\chi^2_{5,0.01} = 15.086$, $\chi^2_{6,0.05} = 12.592$, $\chi^2_{6,0.01} = 16.812$, where $\mathbb{P}(Z > Z_\alpha) = \alpha$ and $\mathbb{P}\left(\chi^2_m > \chi^2_{m,\alpha}\right) = \alpha$, Z is distributed as a standard normal random variable of zero mean and unit variance, and χ^2_m as a $chi - squared$ with m degrees of freedom.