# FINAL EXAM. ECONOMETRICS

Answer each question in a different booklet in two hours and a half. All exercises have the same grading.

1. We wish to estimate the following wage equation

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + abil + u, \quad (1)$$

where $wage$ are monthly earnings, $educ$ are years of schooling, $exper$ are years of work experience and $u$ satisfies the usual assumptions of the multiple linear regression model, but we do not observe the ability of the worker ($abil$).

We have observations for the scores of two tests ($test1$ and $test2$) which are indicators of the ability ($abil$). We assume that the scores can be written as

$$test1 = \gamma_1 abil + e1, \quad \text{Cov}\,(abil, e1) = 0$$

and

$$test2 = \delta_1 abil + e2, \quad \text{Cov}\,(abil, e2) = 0,$$

where $\gamma_1 > 0$ and $\delta_1 > 0$. Given that it is ability which causes the wage, we can assume that $test1$ and $test2$ are not correlated with $u$, and we also assume that $e1$ and $e2$ are not correlated with any of the explanatory variables in (1).

(a) Explain why an OLS regression of (1) with omitted $abil$ will produce inconsistent estimates and argue whether $test1$ and $test2$ are valid instruments.

(b) If we write $abil$ in terms of the score of the first test and we plug in the result in (1), we obtain

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \alpha_1 test1 + v. \quad (2)$$

Determine the value of $\alpha_1$, write $v$ in terms of $u$ and $e1$, and prove that $test1$ is endogenous in this equation. Would an OLS regression of (2) produce consistent estimates of $\beta_1$?

(c) If additionally we assume that $e1$ and $e2$ are not mutually correlated, would you use $test2$ preferably as an additional control variable or as an instrument for $test1$ in (2)? Explain your answer.

(d) Consider equation (2) and the estimation output of Table 1. Test, if possible, whether $test2$ is a relevant instrument for $test1$. Test, if possible, whether $test2$ is exogenous. Which information is given by the estimation output about whether $test1$ is endogenous or exogenous (assuming that $test2$ is exogenous)?

Table 1: Regression table

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dependent var.: | log($wage$) | log($wage$) | log($wage$) | test1 | test2 | log($wage$) |
| educ | 0.0780 | 0.0573 | 0.0478 | 2.637 | 1.258 | 0.00965 |
| | (0.00680) | (0.00792) | (0.00860) | (0.243) | (0.116) | (0.0178) |
| exper | 0.0163 | 0.0157 | 0.0179 | 0.239 | -0.290 | 0.0145 |
| | (0.0140) | (0.0140) | (0.0138) | (0.398) | (0.222) | (0.0154) |
| exper$^2$ | 0.000152 | 0.000165 | -0.0000685 | -0.0181 | 0.0307 | 0.000194 |
| | (0.000588) | (0.000591) | (0.000587) | (0.0167) | (0.00916) | (0.000656) |
| test1 | | 0.00579 | 0.00468 | | 0.146 | 0.0191 |
| | | (0.000984) | (0.000999) | | (0.0155) | (0.00424) |
| test2 | | | 0.00758 | 0.524 | | |
| | | | (0.00206) | (0.0614) | | |
| Constant | 5.517 | 5.214 | 5.194 | 47.02 | 2.672 | 4.514 |
| | (0.125) | (0.131) | (0.128) | (3.874) | (2.336) | (0.239) |
| Observations | 935 | 935 | 935 | 935 | 935 | 935 |
| $R^2$ | 0.131 | 0.162 | 0.176 | 0.322 | 0.267 | . |

Robust standard errors in parentheses
All regressions are fitted by OLS, except (6), which is fitted by 2SLS with test2 as IV for test1.

2. We want to estimate this equation

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u.$$

to explain the minutes of sleep at night (per week), *sleep*, of a sample of workers, males and females, in terms of *totwrk* (mins worked per week), *educ* (years of schooling), *age* (in years) and *yngkid* (which is a binary variable equal to one if children less than 3 years old are present at home). Assume that $u$ satisfies the usual regression assumptions, including conditional homoskedasticity. Using the appropriate estimation output in Table 2 answer the following questions.

(a) Test whether the same regression model is appropriate for both men and women and whether there is a discrimination against women in the child care duties.

(b) Test whether the effect of *age* on *sleep* depends on gender and find the level of *age* where the expected value of *sleep* is minimum for women, all other factors fixed.

(c) Construct and interpret a 95% confidence interval for the effect over *sleep* of an increment of one year of education for a man.

(d) Test if the average effect on *sleep* of one additional year of *age* is equal to the effect of one year less of *educ* for 20 years old males, everything else fixed.

Table 2: Regression table

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Dependent var.: | sleep | sleep | sleep | sleep | sleep |
| totwrk | -0.146 | -0.163 | -0.182 | -0.183 | -0.182 |
| | (0.0191) | (0.0207) | (0.0293) | (0.0291) | (0.0293) |
| educ | -11.14 | -11.71 | -13.05 | -13.87 | -7.731 |
| | (5.747) | (5.748) | (7.767) | (7.646) | (11.58) |
| age | -8.124 | -8.697 | 7.157 | -9.230 | |
| | (11.86) | (11.79) | (13.63) | (11.78) | |
| $age^2$ | 0.126 | 0.128 | -0.0448 | 0.133 | |
| | (0.137) | (0.136) | (0.156) | (0.136) | |
| yngkid | 17.15 | -0.0228 | 60.38 | 39.54 | 60.38 |
| | (53.93) | (53.91) | (64.52) | (62.57) | (64.52) |
| female | | -87.75 | 590.5 | -226.2 | 590.5 |
| | | (35.54) | (541.6) | (162.4) | (541.6) |
| totwrkf*female | | | 0.0422 | 0.0381 | 0.0422 |
| | | | (0.0412) | (0.0406) | (0.0412) |
| educ*female | | | 2.847 | 5.748 | 2.847 |
| | | | (11.53) | (11.01) | (11.53) |
| age*female | | | -37.51 | | -37.51 |
| | | | (24.91) | | (24.91) |
| $age^2$*female | | | 0.413 | | 0.413 |
| | | | (0.289) | | (0.289) |
| yngkid*female | | | -178.7 | -128.0 | -178.7 |
| | | | (117.6) | (109.6) | (117.6) |
| $age - educ$ | | | | | 7.157 |
| | | | | | (13.63) |
| $age^2 - 41*educ$ | | | | | -0.0448 |
| | | | | | (0.156) |
| Constant | 3825.4 | 3928.6 | 3648.2 | 4010.0 | 3648.2 |
| | (259.3) | (257.9) | (323.0) | (278.7) | (323.0) |
| Observations | 706 | 706 | 706 | 706 | 706 |
| $R^2$ | 0.115 | 0.123 | 0.131 | 0.126 | 0.131 |

Standard errors in parentheses
All regressions are fitted by OLS

3. Consider a simple regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

and let $Z_i$ be a binary instrument for $X_i$.

(a) Show that the 2SLS estimator of $\beta_1$ can be written as

$$\hat{\beta}_1^{2SLS} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0}$$

where $\bar{Y}_1$ and $\bar{X}_1$ denote the means of $Y_i$ and $X_i$ (respectively) over that part of the sample with $Z_i = 1$ and $\bar{Y}_0$ and $\bar{X}_0$ denote the means of $Y_i$ and $X_i$ (respectively) over that part of the sample with $Z_i = 0$.

Hint: denoting by $n_1$ the number of observations for which $Z_i = 1$ and by $n_0$ the number of observations for which $Z_i = 0$, $n = n_1 + n_{0,}$ we can write

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\left(\sum_{i:Z_i=1} Y_i + \sum_{i:Z_i=0} Y_i\right) = \frac{n_1}{n}\bar{Y}_1 - \frac{n_0}{n}\bar{Y}_0.$$

Consider a simple model to estimate the effects of personal computer (PC) ownership on college grade point average for graduating seniors at a university,

$$GPA_i = \beta_0 + \beta_1 PC_i + u_i$$

where $PC_i$ is a binary variable indicating PC ownership.

(b) Why might PC ownership be correlated with $u_i$? Explain why $PC_i$ is likely to be related to parent's annual income. Does this mean that parental income is a good instrumental variable for $PC_i$? Why or why not.

(c) Suppose that, four years ago, the university gave grants to buy computers to half of the incoming students, and the students who received the grants were randomly chosen. Explain how you would use this information to construct an instrumental variable for $PC_i$.

In particular, if you were told

• that among those students who received the grants, 90% of them owned a PC and the group had an average GPA of 3.05 and

• that among those students who did not receive the grants, 75% of them owned a PC and the group had an average GPA of 2.75.

What would your estimate $\hat{\beta}_1^{2SLS}$ be?

(d) Now imagine that the university only gave grants to (randomly selected) students whose parent's family income were lower than a given threshold (and we have a list of students that qualified, but we still do not observe family income). How would you need to modify your model and/or estimation strategy to obtain consistent estimates of $\beta_1$?

**SOME CRITICAL VALUES:** $Z_{0.90} = 1.282$, $Z_{0.95} = 1.645$, $Z_{0.975} = 1.96$, $\chi^2_{2,0.95} = 5.99$, $\chi^2_{2,0.975} = 7.378$, $\chi^2_{3,0.95} = 7.815$, $\chi^2_{3,0.975} = 9.348$, $\chi^2_{4,0.95} = 9.488$, $\chi^2_{4,0.975} = 11.143$, $\chi^2_{5,0.95} = 11.071$, $\chi^2_{5,0.975} = 12.833$, $\chi^2_{6,0.95} = 12.592$, $\chi^2_{6,0.975} = 14.449$, where $\mathbb{P}(Z \leq Z_\alpha) = \alpha$ and $\mathbb{P}\left(\chi^2_m \leq \chi^2_{m,\alpha}\right) = \alpha$, $Z$ is distributed as a standard normal with zero mean and unit variance, and $\chi^2_m$ as a chi-square with $m$ degrees of freedom.