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 α_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 β_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 β_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 β_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 $\mathbb{P}(Z \le Z_\alpha) = \alpha$ and $\mathbb{P}\left(\chi^2_m \le \chi^2_{m,\alpha}\right) = \alpha$, Z is distributed as a standard normal with zero mean and unit variance, and χ^2_m as a chi-square with m degrees of freedom.

FINAL EXAM. ECONOMETRICS
SOLUTIONS

1. *We wish to estimate the following wage equation*

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + abil + u, \tag{3}$$

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* (3).

(a) *Explain why an OLS regression of* (3) *with omitted abil will produce inconsistent estimates and argue whether* $test1$ *and* $test2$ *are valid instruments.*

[50%] Es razonable pensar que $abil$ estará correlada con alguna de los regresores incluídos, en particular con $educ$, por lo que se estaría incumpliendo el supuesto $E\left[u|\, educ, exper\right] = 0$ ya que $\text{Cov}(u, educ) \neq 0$.

[50%] De igual forma, las dos ecuaciones para $test1$ y para $test2$, también implican que $\text{Cov}(test1, abil) \neq 0$ y $\text{Cov}(test2, abil) \neq 0$, y por tanto estarían correladas con el error $abil + u$, es decir, no serían exógenas (aunque previsiblemente estarían correladas con $educ$ y otros regresores potencialmente endógenos).

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

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \alpha_1 test1 + v. \tag{4}$$

Determine the value of α_1*, write* v *in terms of* u *and* $e1$*, and prove that* $test1$ *is endogenous in this equation. Would an OLS regression of* (4) *produce consistent estimates of* β_1*?*

[50%] Despejando $abil$ se obtiene

$$abil = \frac{1}{\gamma_1} test1 - \frac{1}{\gamma_1} e1$$

y sustituyendo en (3) se obtiene

$$\begin{aligned}
\log(wage) &= \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \frac{1}{\gamma_1} test1 - \frac{1}{\gamma_1} e1 + u, \\
\alpha_1 &= \frac{1}{\gamma_1}, \quad v = u - \frac{1}{\gamma_1} e1.
\end{aligned}$$

[50%] En este caso, los regresores originales, $educ$, $exper$ y $exper^2$ están incorrelados con v porque lo están con $e1$ y con u. $test1$ también está incorrelado con u, pero no con $e1$, ya que $\text{Cov}(test1, e1) = Var\,(e1) > 0$, por lo que $test1$ es endógena en (4) y en ese caso el estimador MCO de todos los coeficienets, incluyendo β_1, serán inconsistentes.

(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* (4)*? Explain your answer.*

[50%] Para saber si $test2$ es un buen instrumento hay que comprobar la exogeneidad de $test2$ en (4) y su relevancia para el regresor endógeno en (4) que es $test2$:

- exogeneidad: $Cov\,(test2, v) = 0$, que se cumple porque $test2$ está incorrelado con u (no está omitida en (3)) y con $e1$, ya que se asume que el error $e1$ no depende de ningún regresor en (3) .

- relevancia: $\text{Cov}(test2, test1) = \gamma_1 \delta_1 Var\,(abil) \neq 0$.

[50%] Por tanto $test2$ sería un instrumento válido, pero por esa razón no podría ser una buena variable de control porque no aportaría ninguna información sobre factores omitidos contenidos en el error v, ya que $Cov\,(test2, v) = 0$.

(d) *Consider equation* (4) *and the estimation output of Table 1. Test, if possible, whether* $test2$ *is a relevant instrument for* $test1$.

[25%] Para hacer el contraste hay que comprobar que $test2$ es significativa en la forma reducida de $test1$, regresando $test1$ sobre todas las variables exógenas

$$test1 = \pi_0 + \pi_1 educ + \pi_2 exper + \pi_3 exper^2 + \pi_4 test2 + w.$$

Se realizaría el contraste de

$$\begin{aligned} H_0 &: \pi_4 = 0 \\ H_1 &: \pi_4 \neq 0 \end{aligned}$$

con un contraste t.

[25%] Usando el output de la regresión (4)

$$t_4 = \frac{\hat{\pi}_4}{se\,(\hat{\pi}_4)} = \frac{0.524}{0.0614} = 8.534$$

que es significativo comparado con cualquier valor crítico de una $N\,(0,1)$, por lo que $test2$ es relevante.

Test, if possible, whether $test2$ *is exogenous.*

[25%] El contraste de $Cov\,(test2, v) = 0$ no se puede realizar porque la ecuación está exactamente identificada al existir un sólo instrumento, $test2$, para el regresor endógeno, $test1$.

Which information is given by the estimation output about whether $test1$ *is endogenous or exogenous (assuming that* $test2$ *is exogenous)?*

[25%] Si $test1$ fuese exógena, entonces los estimadores MCO deberían ser consistentes, al igual que los estimadores MC2E, ya que el instrumento se supone exógeno. En este caso comparando las regresiones (2) y (6) en la Tabla 1 podemos ver que ciertos coeficientes, como el de *educ*, cambian sustancialmente, indicando que posiblemente algo vaya mal en la regresión MCO si damos por buena la regresión MC2E.

Note. With that general argument it would be enough for the full grade. It could be argued that $test1$ is a valid control variable, but it should be demonstrated/argued that $test1$ satisfies the conditions of a control variable (that is, once it is conditioned by $test1$, the expected value of error v does not change when it changes *educ*), although in reality there is no argument to justify this, the problem arises that the part of the error v that does not explain $test1$, may be correlated with *educ*, even if v is not.

For the demonstration that in particular the MCO of β_1 is consistent, the key would be to verify that when we replace the regression of the error v on the possible control variable $test1$,

$$v = \eta_0 + \eta_1 test1 + s, \quad Cov\,(test1, s) = 0, \quad \eta_1 \neq 0$$

where $\eta_1 \neq 0$ because we have concluded that $test1$ is endogenous, in the regression with $test1$ and we obtain a model with error s,

$$\begin{aligned} log(wage) &= \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \alpha_1 test1 + \eta_0 + \eta_1 test1 + s \\ &= \beta_0 + \eta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + (\alpha_1 + \eta_1)\, test1 + s, \end{aligned}$$

it holds that

$$Cov\,(educ, s) = 0 \tag{5}$$

(and similarly for $exper$ and $exper^2$).

From this OLS regression obviously one can not expect to consistently estimate the true coefficient of $test1$, α_1, but $\alpha_1 + \eta_1 \neq \alpha_1$, but also this problem is transmitted to the estimation of the other coefficients, except if it is fulfilled (5), which it is not true because

$$\begin{aligned} Cov\,(educ, s) &= Cov\,(educ, v - \eta_1 test1) \\ &= Cov\,(educ, -\eta_1 test1) \quad \text{because} \quad Cov\,(educ, v) = 0 \\ &= -\eta_1 Cov\,(educ, test1) \\ &= -\eta_1 Cov\,(educ, \gamma_1 abil + e1) \\ &= -\eta_1 Cov\,(educ, \gamma_1 abil) \quad \text{because} \quad Cov\,(educ, e1) = 0 \\ &= -\eta_1 \gamma_1 Cov\,(educ, abil) \end{aligned}$$

which is different from zero because we hope that $educ$ is correlated with $abil$ and $\eta_1\gamma_1 \neq 0$.

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.*

[25%] For that we have to test in the model including the binary regressor $female$ and all the interactions of $female$ with the regressors, whether all those variables depending on $female$ are jointly significant, i.e. testing in

$$\begin{aligned} sleep &= \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid \\ &\quad + \beta_6 female + \beta_7 totwrk * female + \beta_8 educ * female + \beta_9 age * female \\ &\quad + \beta_{10} age^2 * female + \beta_{11} yngkid * female + u \end{aligned}$$

the hypotheses

$$\begin{aligned} H_0 &: \beta_6 = \cdots = \beta_{11} = 0 \\ H_1 &: H_0 \text{ is false.} \end{aligned}$$

[25%] For that we can conduct an F test under the assumption of homoscedasticity,

$$F = \frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} \frac{n-k-1}{q} = \frac{0.131 - 0.115}{1 - 0.131} \frac{706 - 11 - 1}{6} = 2.13$$

comparing the restricted model (1) with the unrestricted (3). The 5% critical value is given by the $\chi^2(6)/6$ distribution of the F test for large samples, i.e. $12.592/6 = 2.099$, so that the F statistic is significatively different from zero at the 5% level, and we can reject (marginally) H_0, concluding that the regression for females is different from that for males.

Note. Testing the hypotheses

$$\begin{aligned} H_0^* &: \beta_6 = 0 \\ H_1^* &: \beta_6 \neq 0 \end{aligned}$$

in a model with only $female$

$$\begin{aligned} sleep &= \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid \\ &\quad + \beta_6 female + u \end{aligned} \tag{6}$$

is not correct because (6) is not accouting for (full) separate regressions for women and men, but a restricted version of the general model where H_1^* is just a particular deviation of the hypothesis of equal regressions, once the restrictions $\beta_7 = \cdots = \beta_{11} = 0$ are imposed without justification.

[25%] To identify discrimination against women we can test

$$\begin{aligned} H_0 &: \beta_{11} = 0 \\ H_1 &: \beta_{11} < 0 \end{aligned}$$

where the alternative indicates that women sleep less on average than males when children are present, with a one-sided t-test

[25%]

$$t_{11} = \frac{\hat{\beta}_{11}}{se\left(\hat{\beta}_{11}\right)} = \frac{-178.7}{117.6} - 1.519\,6$$

which is not significative at the 5% level, for which the one-sided critical value from the $N(0,1)$ is -1.645, so that there is no enough evidence supporting discrimination.

(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 equal.*

[30%] The hypotheses to be tested are

$$\begin{aligned} H_0 &: \beta_9 = \beta_{10} = 0 \\ H_0 &: \beta_9 \neq 0 \text{ or } \beta_{10} \neq 0 \end{aligned}$$

with an F test under the assumption of homoscedasticity,

[30%]

$$F = \frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} \frac{n-k-1}{q} = \frac{0.131 - 0.126}{1 - 0.131} \frac{706 - 11 - 1}{2} = 2$$

comparing the restricted model (4) with the unrestricted (3). The 5% critical value is given by the $\chi^2(2)/2$ distribution of the F test for large samples, i.e. $5.99/2 = 2.99$, so that the F statistic is not significatively different from zero at the 5% level, so we can not reject H_0, concluding that there is not empirical evidence supporting that the effect of *age* over *sleep* is different by gender.

[40%] For women the effect of *age* is described by

$$(\beta_3 + \beta_9)\, age + (\beta_4 + \beta_{10})\, age^2$$

and, given that $\hat{\beta}_4 + \hat{\beta}_{10} > 0$, the minimum is estimated as

$$age^*_{female} = -\frac{\hat{\beta}_3 + \hat{\beta}_9}{2\left(\hat{\beta}_4 + \hat{\beta}_{10}\right)} = -\frac{7.157 - 37.51}{2 * (-0.0448 + 0.413)} = 41.22.$$

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

[75%] This effect is given by the coefficient β_2, so the confidence interval is

$$\hat{\beta}_2 \pm 1.96 se\left(\hat{\beta}_2\right)$$

i.e., using output (3) we obtain

$$-13.05 \pm 1.96 \cdot 7.767 \quad \text{or} \quad [-28.\,273,\ 2.\,173\,3]$$

[25%] meaning that this effect is not significatively different from zero at the 5% level.

(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.*

[30%] The two effects are

$$\begin{aligned} & E\left[sleep|\, age = 21, male, x\right] - E\left[sleep|\, age = 20, male, x\right] \\ = \;& \beta_3 (21) + \beta_4 (21)^2 - \left(\beta_3 (20) + \beta_4 (20)^2\right) \\ = \;& \beta_3 + 41\beta_4, \end{aligned}$$

and

$$E\left[sleep|\, educ-1, male, x\right] - E\left[sleep|\, educ, male, x\right] = \beta_2\left(educ-1\right) - \beta_2 educ = -\beta_2,$$

respectively, and equality between the two effects implies that

$$H_0 : \theta = 0$$

where

$$\theta = \left(\beta_3 + 41\beta_4\right) - \left(-\beta_2\right) = \beta_2 + \beta_3 + 41\beta_4 = 0$$

[40%] and replacing $\beta_2 = \theta - \left(\beta_3 + 41\beta_4\right)$ in the model we obtain

$$\begin{aligned} sleep &= \beta_0 + \beta_1 totwrk + \left\{\theta - \left(\beta_3 + 41\beta_4\right)\right\} educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid \\ &\quad + \beta_6 female + \beta_7 totwrk * female + \beta_8 educ * female + \beta_9 age * female \\ &\quad + \beta_{10} age^2 * female + \beta_{11} yngkid * female + u \end{aligned}$$

or equivalently,

$$\begin{aligned} sleep &= \beta_0 + \beta_1 totwrk + \theta educ + \beta_3\left(age - educ\right) + \beta_4\left(age^2 - 41educ\right) + \beta_5 yngkid \\ &\quad + \beta_6 female + \beta_7 totwrk * female + \beta_8 educ * female + \beta_9 age * female \\ &\quad + \beta_{10} age^2 * female + \beta_{11} yngkid * female + u \end{aligned}$$

so for testing H_0 against $H_1 : \theta \neq 0$ we use a t-test for the coefficient of *educ* in regression (5) of Table 2,
[30%]

$$t_\theta = \frac{\hat{\theta}}{se\left(\hat{\theta}\right)} = \frac{-7.731}{11.58} = -0.667\,62$$

which is not significative against the $N\left(0,1\right)$ critical value at any usual level, meaning that we cannot reject the null of equality between both effects.

3. *Consider a simple regression model*

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

and let Z_i be an binary instrument for X_i.

(a) *Show that the 2SLS estimator of β_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

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

[50%, 10% for the first expression of 2SLS] We know that

$$\begin{aligned} \hat{\beta}_1^{2SLS} &= \frac{\widehat{Cov}\left(Y,Z\right)}{\widehat{Cov}\left(X,Z\right)} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i Z_i - \bar{Y}\bar{Z}}{\frac{1}{n}\sum_{i=1}^{n} X_i Z_i - \bar{X}\bar{Z}} \\ &= \frac{\frac{1}{n}\sum_{i:Z_i=1} Y_i - \bar{Y}\frac{n_1}{n}}{\frac{1}{n}\sum_{i:Z_i=1} X_i - \bar{X}\frac{n_1}{n}} = \frac{\frac{n_1}{n}\bar{Y}_1 - \bar{Y}\frac{n_1}{n}}{\frac{n_1}{n}\bar{X}_1 - \bar{X}\frac{n_1}{n}} = \frac{\bar{Y}_1 - \bar{Y}}{\bar{X}_1 - \bar{X}} \end{aligned}$$

because

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\sum_{i:Z_i=1} 1 = \frac{n_1}{n}$$

where n_1 is the number of observations for which $Z_i = 1$, and

$$\bar{Y}_1 = \frac{1}{n_1}\sum_{i:Z_i=1} Y_i, \quad \bar{X}_1 = \frac{1}{n_1}\sum_{i:Z_i=1} X_i.$$

Next, denoting as n_0 is the number of observations for which $Z_i = 0$, $n = n_1 + n_0$, we can write

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

[50%] and similarly for $\bar{X}$, we obtain

$$\begin{aligned}
\hat{\beta}_1^{MC2E} &= \frac{\bar{Y}_1 - \bar{Y}}{\bar{X}_1 - \bar{X}} = \frac{\bar{Y}_1 - \frac{n_1}{n}\bar{Y}_1 - \frac{n_0}{n}\bar{Y}_0}{\bar{X}_1 - \frac{n_1}{n}\bar{X}_1 - \frac{n_0}{n}\bar{X}_0} = \frac{\bar{Y}_1\left(1 - \frac{n_1}{n}\right) - \frac{n_0}{n}\bar{Y}_0}{\bar{X}_1\left(1 - \frac{n_1}{n}\right) - \frac{n_0}{n}\bar{X}_0} \\
&= \frac{\bar{Y}_1\frac{n_0}{n} - \frac{n_0}{n}\bar{Y}_0}{\bar{X}_1\frac{n_0}{n} - \frac{n_0}{n}\bar{X}_0} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0}.
\end{aligned}$$

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 is likely to be related to parent's annual income. Does this mean that parental income is a good instrumental variable for PC? Why or why not.*

[30%] Parents income can be correlated with many causal factors included in u describing different aspects of previous education and access to learning opportunities that affect GPA, and also would be correlated to PC ownership, everything else equal, because of the availability of a larger budget.

[30%] This implies that u and PC would be correlated through income,

[40%] and therefore PC is endogenous in the equation. In sum, parental income would be correlated with PC (relevance) but also with u (so not exogenous) so it would not be a valid instrument.

(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.*

[40%] We should construct a binary instrumental variable Z_i setting $Z_i = 1$ if the student received the grant and 0 if not. We expect that Z_i should be correlated with PC_i because receiving the grant gives incentives to buy a computer, everything else equal, even if not everybody receiving the grant bought this (or other student might have bought a PC without receiving the grant).

[30%] Then Z_i should be also independent with respect any factor in u_i because it was randomly assigned, and then it is exogenous.

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?*

[30%]

$$\hat{\beta}_1^{2SLS} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0} = \frac{3.05 - 2.75}{0.90 - 0.75} = 2.$$

(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* β_1*?*

[30%] In this case Z_i as defined before would be (negatively) correlated with some factors in u_i related to parents income, as Z_i is assigned differently in terms of this income.

[30%] However if we construct a binary variable $W_i = 1$ if students qualified for the grant, $= 0$ otherwise, and we include this in the regression with PC_i,

$$GPA_i = \beta_0 + \beta_1 PC_i + \beta_2 W_i + v_i$$

[40%] then Z_i now becomes uncorrelated with the remaining factors included in the new error term v_i because Z_i was assigned independently of them (conditionally on W_i) and would be a valid instrument.