EXTRAORDINARY EXAM. ECONOMETRICS

Answer each question in a different booklet in two hours and a half.
All questions (a) , (b) , etc. have the same grading.

1. A model for estimating the effects of smoking on annual income (perhaps through work absences due to illness or effects on productivity) is

$$\log(income) = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 age + \beta_4 age^2 + u_1 \tag{1}$$

where *cigs* is the number of cigarettes smoked per day, on average, and *educ* (years of schooling) and *age* (in years) are supposed exogenous.

To reflect the fact that tobacco consumption could be determined simultaneously with the income, a cigarette demand function is specified,

$$cigs = \gamma_0 + \gamma_1 \log(income) + \gamma_2 educ + \gamma_3 age + \gamma_4 age^2 + \gamma_5 \log(cigpric) + \gamma_6 restaurn + u_2, \tag{2}$$

where *cigpric* is the price of a pack of cigarettes (in cents) and *restaurn* is a binary variable equal to one if the person lives in a state whose restaurants restrict the consumption of tobacco, and it is assumed that these are exogenous variables to the individual.

(a) How do you interpret β_1 and γ_1? Which signs do you expect for γ_5 and γ_6? Under which assumptions would equations (1) and (2) be identified, if any?

(b) Execute all hypothesis tests that are feasible to check the two conditions for the identification of (1) and (2) given the information in Table 1. Make explicit any further assumption you use.

(c) Compare the estimation of the income equation by OLS and by 2SLS. What conclusions can you draw? Provide an estimate of the expected variation in income from increasing tobacco consumption by 1 cigarette per day. Could you make a valid test with the given information to determine if the variation is negative?

Table 1: Regression table

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent var.:	log(*income*)	cigs	cigs	log(*income*)	hatu2	hatu2
cigs	0.00173					
	(0.00143)					
educ	0.0604	-0.450	-0.472	0.0397	-0.000301	-1.15e-10
	(0.00745)	(0.156)	(0.156)	(0.0115)	(0.0103)	(0.0103)
age	0.0577	0.823	0.824	0.0938	-0.000269	-3.60e-10
	(0.00920)	(0.136)	(0.137)	(0.0172)	(0.0115)	(0.0116)
age^2	-0.000631	-0.00959	-0.00958	-0.00105	0.00000214	3.17e-12
	(0.0000984)	(0.00144)	(0.00146)	(0.000197)	(0.000123)	(0.000124)
log(*cigpric*)		-0.351			0.439	
		(6.027)			(0.392)	
restaurn		-2.736			-0.0145	
		(1.001)			(0.0675)	
hatcigs				-0.0421		
				(0.0188)		
Constant	7.795	1.580	-0.332	7.781	-1.784	1.05e-08
	(0.208)	(25.19)	(3.226)	(0.208)	(1.668)	(0.259)
Observations	807	807	807	807	807	807
R^2	0.165	0.071	0.044	0.169	0.002	0.001

Standard errors in parentheses. All regressions are fitted by OLS.
hatcigs are the predictions of cigs from the regression in column (2).
hatu2 are the residuals of the 2SLS regression of the demand function, eq. (2), using lcigpric and restaurn as instruments for cigs.

2. We want to estimate a linear probability model to study the determinants that lead to the subscription of a pension plan,

$$pp = \beta_0 + \beta_1 \log(income) + \beta_2 age + \beta_3 age^2 + \beta_4 male + \beta_5 married + \beta_6 male * married + u, \quad (3)$$

where $pp = 1$ if the individual has subscribed a plan and 0 otherwise, *income* is the annual income, *age* is age in years, *married* is a binary variable indicating if the individual is married and *male* is a binary variable indicating if the individual is a male. Model (3) has been estimated with a sample of 9275 individuals and the results are summarized in Table 2, together with those of other related models.

(a) Interpret the coefficient β_1 in (3) and provide a 95% confidence interval for the effect on pp of a 10% increment in income.

(b) Determine if the probability of a single man subscribing to a pension plan is different from that of a single woman using the model (3) . What is the estimated difference between this probability for a man and a woman, both married?

(c) Explain why with the information provided it is not possible to carry out a valid hypothesis test about whether the probability of subscribing a pension plan depends on the gender and marital status of the individual.

(d) Finally, it is decided to estimate different models for men and for women, including explanatory variables $\log(income)$, age, age^2 and *married*. Provide the estimated equation for men using the output from column (3).

Table 2: Regression table

	(1)	(2)	(3)
Dependent var.:	pp	pp	pp
log(*income*)	0.229	0.217	0.231
	(0.00838)	(0.00765)	(0.00948)
age	0.0137	0.0142	0.0158
	(0.00347)	(0.00346)	(0.00394)
age^2	-0.000160	-0.000165	-0.000181
	(0.0000398)	(0.0000397)	(0.0000448)
male	-0.0151		0.197
	(0.0143)		(0.177)
married	-0.0360		-0.0374
	(0.0116)		(0.0118)
male*married	-0.00875		-0.00145
	(0.0244)		(0.0258)
$\log(income) * male$			-0.0112
			(0.0203)
age*male			-0.00761
			(0.00867)
age^2*male			0.0000747
			(0.000102)
Constant	-0.774	-0.770	-0.830
	(0.0723)	(0.0718)	(0.0824)
Observations	9275	9275	9275
R^2	0.084	0.083	0.084

Robust standard errors in parentheses
All regressions are fitted by OLS

3. Suppose that the true model that you want to estimate to explain the results of an exam is

$$Testscore_i = \beta_0 + \beta_1 STR_i + \beta_2 Effort_i + u_i$$

where STR is a continuous variable that increases when the class size gets larger. *Effort* is a continuous variable that increases when the student puts more effort in class. Finally, the error component u_i incorporates all other factors that are assumed to be uncorrelated with the other explanatory variables. Assume $\beta_1 < 0$ and $\beta_2 > 0$, $\text{Cov}(STR, Effort) > 0$, $\text{Cov}(STR, u) = 0$, $\text{Cov}(Effort, u) = 0$.

However, the econometrician estimates the following specification

$$Testscore_i = \gamma_0 + \gamma_1 STR_i + e_i. \tag{4}$$

(a) Discuss whether the OLS estimation of γ_1 in (4) is biased or not for β_1, and if so, which is the sign of the bias.

(b) Imagine that instead of *Effort* we observe an indicator of class attendance, CA_i for which it is known that $\text{Cov}(CA, Effort) > 0$, so that it holds that

$$Effort_i = \alpha_0 + \alpha_1 CA_i + v_i, \quad \text{Cov}(CA, v) = 0, \quad \alpha_1 > 0.$$

It is also known that

$$\text{Cov}(STR, v) = 0.$$

Interpret this condition.

(c) What would be the expected sign of δ_3 in the following regression model?

$$Testscore_i = \delta_0 + \delta_1 STR_i + \delta_3 CA_i + w_i$$

Do you think that the OLS estimate of δ_1 would be biased or unbiased for β_1? Why?

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.