STOCKHOLMS UNIVERSITET
Statistiska institutionen
Oskar Gustafsson

HOME EXAM, ECONOMETRICS I
2021-04-30

Time for examination: 9.00-14.00, The exam should be submitted electronically at the latest 15.00. You will find clear instructions regarding the submission in a separate file.

Allowed tools: Pocket calculator, computer, course books and lecture notes.
Note that the exam should be written individually. All types of collaborations and/or help from others are strictly forbidden.

For questions regarding the content of the exam, email to: oskar.gustafsson@stat.su.se
For questions regarding the submission, email to: expedition@stat.su.se

The exam consists of 4 independent problems. Well motivated and clear solutions are required for full scoring on a problem. Don't forget to state any necessary assumptions or conditions where needed.
Passing rate: 50% of overall total, which is 100 points. For detailed grading criteria, see the course description. Answers may be given in English or Swedish.

Good luck!

Problem 1. (18 points)
Indicate which alternative that is correct. Answering more than one alternative result in 0 points on the sub-question. No motivation is required.

1. Simultaneous equation models are typically estimated with:

   (a) Tobit regression models

   (b) Logistic regression

   (c) 2SLS/IV regression

   (d) Weighted least squares

2. If the error term is heteroscedastic, OLS estimates of the parameters are not:

   (a) normally distributed

   (b) unbiased

   (c) consistent

   (d) possible to calculate

3. What is generally **not** true for a valid instrumental variable:

   (a) they are uncorrelated with the endogenous regressor(s)

   (b) they are significant in the reduced regression

   (c) they are uncorrelated with the error term in the structural equation

   (d) they are correlated with the other exogenous regressor(s)

4. The probit model:

   (a) is estimated by maximum likelihood

   (b) takes on values between -1 and 1

   (c) is homoscedastic

   (d) gives identical parameter estimates as the logit model

5. Which of the following statements is false

   (a) OLS estimates are consistent when $n \to \infty$ when we have omitted variables correlated with the error term and the independent variables.

   (b) The R^2 can only take on values between 0 and 1.

   (c) If we have a heteroskedastic error term we cannot rely on our usual t-tests and F-tests even in large samples.

   (d) The estimated variance for the slope coefficient in *insrumental variable regression* is always greater than or equal to that of OLS. I.e. $\hat{var}(\hat{\beta}_{IV}) \geq \hat{var}(\hat{\beta}_{OLS})$ always holds.

6. If we have a data set for which the dependent variable is unobserved for values such that $y < 0$. I.e. y can take on values lower than 0, but we cannot observe them. Which model should be used to obtain appropriate parameter estimates?

   (a) The Tobit model

   (b) OLS

   (c) Censored regression

   (d) The Heckman error correction model

Problem 2. (30 points)

Consider the following regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 \log x_{2i} + u_i, \qquad \text{where } u_i \sim N(0, \sigma_i^2),$$

where the variance of the error term depend on x_2 through the relation: $\sigma_i^2 = \sigma^2 \times \log x_2$

Table 1:

	Dependent variable: y		
	coef	std. error	t-value
intercept	5.0	0.3	?
x_1	1.4	0.56	?
x_1^2	-0.125	0.1	?
$\log x_2$	6.2	1.1	?
$\hat{\sigma}^2 = 0.25$	$n = 300$		

Note: The standard errors are estimated using OLS.

Also assume that the first four Gauss-Markov assumptions are fulfilled (i.e. all except the one regarding homoscedasticity).

a) Calculate the missing t-values in Table 1. Which coefficients are statistically significant at the 1%-level (using a two-sided test)?

b) What is the effect of a one unit increase in x_1 on y? Interpret the result.

c) What is the effect of a one unit increase in x_2 on y? Interpret the result.

d) Given the estimated parameters and standard errors of the model above. Assume that we want to predict the value of a new observation with $x_{1j} = 4$ and $x_{2j} = 30$. Calculate the predicted value $\hat{y}_j$ together with a 95% prediction interval.

e) Let's say that we change the dependent variable to instead be $\tilde{y} = 2y + 3$, how will that affect our coefficients? And t-values?

f) Indicate whether each of these statements are true or false (no motivation needed)

1. The slope coefficients are consistently estimated using OLS, and are unbiased.
2. OLS is the *best linear unbiased estimator, (BLUE)*.
3. If the sample size is large enough we can trust our t-tests, f-tests and confidence intervals.

g) Which method would you use to estimate the above model? Clearly motivate your choice, what are the advantages and/or disadvantages of your selected method compared to OLS? (Write maximum half a page)

Problem 3. (27 points)

Consider the dataset *WAGE2* from the R-package *wooldrige* described in Table 2 below. A student has access to n=300 randomly selected observations from the original data set, where some of the original variables have been removed. She is interested in modelling the wage level as a function of some individual characteristics. In particular, she is interested in the effect of *company tenure* on the *wage*. Is it the case that the workers gets rewarded for their loyalty?

Table 2: Data description.

Variable name	Description	Variable name	Description
wph	wage per hour	IQ	IQ score
educ	years of education	age	age in years
exper	years of working experience	tenure	years with current employer
urban	=1 if live in an urban area	married	= 1 if married
black	=1 if black	south	= 1 if live in south

The first thing the student does is to run the following regression:

$$\log wage = \beta_0 + \beta_1 tenure + u. \tag{1}$$

This is referred to as (1) in Table 3 below.
The student remembers that she should probably include more independent variables in her analysis, and she therefore run another regression:

$$\log wage = \beta_0 + \beta_1 tenure + \beta_2 educ + \beta_3 IQ + \beta_4 exper + \beta_5 exper^2 + \beta_6 age + u. \tag{2}$$

which is refered to as (2) in Table 3. When she looks at the results, she notice that the coefficients for $exper$, $exper^2$ and age are not significant. She therefore remove them and estimate the third model:

$$\log wage = \beta_0 + \beta_1 tenure + \beta_2 educ + \beta_3 IQ + u. \tag{3}$$

Finally she realizes that she has access to some more background variables that possibly could be important, so she estimates a fourth regression model:

$$\begin{aligned}\log wage =& \beta_0 + \beta_1 tenure + \beta_2 educ + \beta_3 IQ + \beta_4 exper + \beta_5 exper^2 + \beta_6 age + \\ &+ \beta_7 south + \beta_8 urban + \beta_9 black + \beta_{10} married + u.\end{aligned} \tag{4}$$

Table 3:

	Dependent variable:			
	lwph			
	(1)	(2)	(3)	(4)
tenure	0.017***	0.013***	0.015***	0.013***
	(0.005)	(0.005)	(0.005)	(0.005)
educ		0.036**	0.031**	0.035**
		(0.014)	(0.012)	(0.014)
IQ		0.003*	0.003*	0.003
		(0.002)	(0.002)	(0.002)
exper		0.007		−0.004
		(0.024)		(0.023)
exper2		−0.0001		0.0003
		(0.001)		(0.001)
age		0.005		0.004
		(0.010)		(0.010)
south				−0.043
				(0.054)
urban				0.198***
				(0.051)
black				−0.084
				(0.081)
married				0.260***
				(0.072)
Constant	2.898***	1.861***	2.142***	1.683***
	(0.044)	(0.367)	(0.186)	(0.372)
Observations	300	300	300	300
R^2	0.037	0.097	0.092	0.177
Adjusted R^2	0.034	0.078	0.083	0.148
Residual Std. Error	0.433	0.423	0.422	0.407
F Statistic	11.466***	5.224***	10.053***	6.213***
Note:				*p<0.1; **p<0.05; ***p<0.01

Questions:
a) Interpret the coefficient for tenure in model (1). Do you consider it large or small? Motivate.

b) What is the *ceteris paribus* wage difference between two individuals with 4 years and 7 years of employment at a firm respectively according to model (4)? Interpret the result.

c) Do you think it is a good idea to extend model (1) by including more independent variables? What potential issues does this solve or cause? Motivate.

d) Help the student to calculate an F-test for whether to include *exper*, $exper^2$ and *age* in the model, i.e. test the joint hypothesis $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ in model (2). Do you reject the null-hypothesis at the 5%-level? What does the result mean?
[Hint: use model (3) as your restricted model]

e) Calculate the t-statistic for *urban* in regression (4). Is it significant? Interpret the practical meaning of it.

f) Test the hypothesis: $H_0 : \beta_{10} = 0.15$ versus $H_1 : \beta_{10} > 0.15$ in model (4). Interpret the result.

Problem 4. (25 points)

In a paper by Angrist and Evans (1998) they study the impact of an extra child on women labor supply. In this exam, our primary interest is in the probability of participating in the labor force ($inLf$).
The data is described in the following table

Table 4: Data description.

Variable name	Description	Variable name	Description
inLF (dependent variable)	=1 if in the labor force	age	age of the mother
mkids (main interest)	=1 if more than 2 kids	afam	=1 if African-American
gender1	=1 if first child is a boy	hispanic	=1 if Hispanic
gender2	=1 if second child is a boy	other	=1 if from other ethnicity
boyboy (instrument)	=1 if two first are boys	girlgirl (instrument)	=1 if two first are girls

The equation that we are really interested in is: $inLF = \beta_0 + \beta_1 mkids + u$, in particular we are interested in the ceteris paribus effect of $mkids$. However, there is likely to be correlation between $mkids$ and the error term even after including various independent variables in the regression model. As a solution to this problem Angrist and Evans (1998) suggests to use the information of whether the first two children have the same gender as instrumental variables.

a) What is the idea behind using these instruments? Do you think it is a good idea? What do you think of the model assumptions?

The reduced form, the 2SLS and the OLS (ignoring the potential correlation with the error term) regressions were estimated and the results are shown in Table 5. Use this table for the remaining questions.

b) Is the relevance condition of the instruments fulfilled? Motivate.

c) Can we test for exogeneity of the instruments? Motivate your answer.

d) Are the 2SLS and OLS estimates for $mkids$ significantly different? (both statistical- and practical difference). Motivate.

e) Why do we include the exogenous independent variables in the reduced regression?

Table 5:

	Dependent variable:		
	workp		mkids
	2SLS	*OLS*	*reduced regression*
	(1)	(2)	(3)
mkids	-0.119^{***}	-0.129^{***}	
	(0.028)	(0.002)	
girlgirl			0.079^{***}
			(0.003)
boyboy			0.059^{***}
			(0.003)
age	0.013^{***}	0.013^{***}	0.015^{***}
	(0.001)	(0.0003)	(0.0003)
gender1	0.001	0.001	-0.001
	(0.002)	(0.002)	(0.003)
gender2		-0.005^{***}	
		(0.002)	
afam	0.210^{***}	0.211^{***}	0.100^{***}
	(0.005)	(0.004)	(0.004)
hispanic	0.001	0.003	0.151^{***}
	(0.006)	(0.004)	(0.004)
other	0.032^{***}	0.032^{***}	0.028^{***}
	(0.005)	(0.005)	(0.005)
Constant	0.172^{***}	0.173^{***}	-0.139^{***}
	(0.009)	(0.009)	(0.009)
Observations	254,654	254,654	254,654
R^2	0.028	0.028	0.024
Adjusted R^2	0.028	0.028	0.024
Residual Std. Error	0.492	0.492	0.480
F Statistic		$1{,}058.384^{***}$	910.160^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01