Econ 7010 Final Exam

Ryan T. Godwin

- Time allowed: 2 hours (7:30 pm - 9:30 pm).
- Upload your answers to UM Learn between 9:30 pm and 9:40 pm. 9:40 pm is the final deadline.
- 4 questions total. Answer all questions.
- The number of marks allocated to each question is in [red].
- 120 marks, 120 minutes.
- **Do not collaborate with anyone on this exam.**

---

1. [**40 marks total**] Suppose that the true population model is:

$$\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

(a) [8] Let all of the usual assumptions hold. Suppose that the model that you actually specify and estimate by LS is $\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{u}$ (because X_2 is unobservable, for example). Show that, in general, $\boldsymbol{b}_1$ is biased.

(b) [8] The usual assumptions imply that:

$$\text{plim}\left(\frac{X_1'X_1}{n}\right) = Q_{X_1}$$

and

$$\text{plim}\left(\frac{X_1'\epsilon}{n}\right) = \boldsymbol{0}$$

However, in this case X_2 is unobservable and related to X_1, so that $\text{plim}\left(\frac{X_1'\epsilon}{n}\right) \neq \boldsymbol{0}$.

In this case, show that $\boldsymbol{b}_1$ is inconsistent, in general.

(c) [2] Under what special circumstance would the LS estimator still be unbiased and consistent?

(d) [4] A solution to the problem presented above is to use Instrumental Variables (IV) estimation instead of LS. The simple IV estimator for the problem above is $\boldsymbol{b}_{IV} = (Z'X_1)^{-1}Z'\boldsymbol{y}$. Briefly explain how the instrument is used to "fix" the problem.

(e) [6] Prove that the simple IV estimator is consistent.

(f) [8] Derive the formula for the simple IV estimator using the two-stage-least-squares (2SLS) interpretation of IV: (1) Regress X_1 on Z, get the LS fitted values. (2) Estimate the model $\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{u}$ by OLS.

(g) [4] Suppose that we want to compare the asymptotic variance of $\boldsymbol{b}$ with $\boldsymbol{b}_{IV}$. Explain why we need to consider the asymptotic distributions of $\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})$ and $\sqrt{n}(\boldsymbol{b}_{IV}-\boldsymbol{\beta})$, instead of the distributions of just $\boldsymbol{b}$ and $\boldsymbol{b}_{IV}$.

2. [**30 marks total**] Use some of the following 6 estimated models:

Table 1: Estimation results for question 2

	Dependent variable:					
	log(wage)					
	(1)	(2)	(3)	(4)	(5)	(6)
education	0.047***	0.056***	0.046***	0.058***	0.047***	0.044***
	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.007)
experience	0.014***	0.016***	0.014***	0.015***	0.015***	0.014***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
age	0.019***	0.019***	0.020***	0.020***	0.019***	0.019***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
female	−0.259**	0.082***	−0.275**		−0.189	−0.284**
	(0.125)	(0.030)	(0.121)		(0.120)	(0.125)
Manitoba	−0.099***	−0.102***	−0.064***	−0.066***	−0.103***	
	(0.032)	(0.032)	(0.022)	(0.022)	(0.031)	
Saskatchewan	0.129***	0.131***	0.103***	0.101***	0.127***	
	(0.030)	(0.030)	(0.021)	(0.022)	(0.030)	
female × education	0.019**		0.020***		0.018**	0.021**
	(0.008)		(0.008)		(0.008)	(0.008)
female × experience	0.002*		0.003**			0.003**
	(0.001)		(0.001)			(0.001)
female × Manitoba	0.066	0.065			0.068	
	(0.044)	(0.044)			(0.044)	
female × Saskatchewan	−0.052	−0.061			−0.053	
	(0.043)	(0.043)			(0.043)	
Constant	1.716***	1.555***	1.724***	1.564***	1.685***	1.775***
	(0.087)	(0.065)	(0.086)	(0.065)	(0.086)	(0.087)
Observations	1,000	1,000	1,000	1,000	1,000	1,000
R^2	0.765	0.762	0.763	0.756	0.764	0.749
Adjusted R^2	0.762	0.760	0.761	0.754	0.762	0.747
Note:						*p<0.1; **p<0.05; ***p<0.01

The sample size is 1000. The variables in the data are:

- wage - yearly wage of the worker, measured in thousands of dollars
- experience - years of work experience
- age - the age of the worker in years
- female - a dummy variable equal to 1 if the individual is female, 0 otherwise
- Manitoba - a dummy variable equal to 1 if the worker lives in Manitoba, 0 otherwise
- Saskatchewan - a dummy variable equal to 1 if the worker lives in Saskatchewan, 0 otherwise

---

Table 2: Critical values for the F-test statistic.

q	5% critical value
1	3.84
2	3.00
3	2.60
4	2.37
5	2.21

Use a 5% significance level for all questions.

(a) [10] Do the wages of workers depend on province (location)? Use a hypothesis test.

(b) [6] Is there a different effect of education on wages for men vs. women? Use a hypothesis test.

(c) [8] What are the risks and benefits of using any of the models (2) - (6), instead of model (1)? Use the RLS estimator in your explanation.

(d) [4] Suppose that the error term is **not** Normally distributed. How would you go about testing the hypotheses in parts (a) and (b)?

(e) [2] Why should we use the F-test instead of the Wald test, if we can?

3. [**20 marks total**]

(a) [6] Show how a regression model that includes a polynomial in the variable x allows for a non-linear effect between x and $\boldsymbol{y}$.

(b) [4] Suppose that you are going to estimate a non-linear model by NLS. Explain why you will likely need a numerical algorithm to find the estimates.

(c) [4] Briefly describe what is meant by "tolerance" and "iterations" in the context of a numerical algorithm (such as the Newton-Raphson algorithm).

(d) [6] Briefly describe how the parameter estimates are calculated in each successive iteration of the Newton-Raphson algorithm.

4. [**30 marks total**] Consider the following population model where the variables are *averaged over groups*:

$$\log(\bar{income}_i) = \beta_1 \bar{education}_i + \boldsymbol{\beta_2} \bar{\boldsymbol{x}}_i + \boldsymbol{\epsilon}_i$$

$\bar{income}_i$ is the average income of workers in location i, $\bar{education}_i$ is the average education in location i, and $\bar{\boldsymbol{x}}_i$ is a vector of regressors containing economic and demographic "controls", all of which are also averaged by location. There are only two different locations: North and South. For each observation, there is also a dummy variable d which tells you whether the location is North or South:

$$d = 1 \text{ if location is in the North}$$
$$d = 0 \text{ if location is in the South}$$

In the North, all variables have been averaged over group sizes of 400. In the South, all variables have been averaged over a group size of 100.

(a) [4] For this model and data, which of the standard assumptions have been violated?

(b) [10] What are the consequences of the assumption(s) in (a) being violated, in terms of estimation and hypothesis testing?

(c) [16] Provide solutions to the problem(s) identified in part (b). Explain carefully how you might use the dummy variable, and/or information about the differences in group size.