

Econometrics
Final Examination
Suggested Answer

11/25/2004
Daiji Kawaguchi
College of International Studies

Name _____
Student # _____

Answer each question in the space provided. Your answer must be in English. You are allowed to use a dictionary. Please write legibly so that I can read your answer. This examination ends at 9:55.

Part 1 Basic Part (60pts in total)

1. Multiple Regression Model: Estimation

A linear multiple regression model is given as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

We obtain the random sample from the population and confirm that each of x_1, x_2 has variation within the sample.

a) State when the explanatory variable x_1 is called to be endogenous. (5pts)

$E(u | x_1) \neq 0$, or x_1 and u are correlated.

b) Suppose that x_1, x_2 are both exogenous and linearly independent but two variables are highly correlated. What problem does this cause for the estimation of β_1 . (5pts)

This situation is called as multicollinearity. $Var(\beta_1)$ tends to be large. In other words, we cannot estimate β_1 precisely.

c) Suppose that $\beta_2 > 0$ in the population. A researcher estimated the model without including x_2 , i.e.:

$$y = \beta_0 + \beta_1 x_1 + v.$$

What is the direction of bias of the OLS estimator of β_1 given $Cov(x_1, x_2) < 0$. (Hint: Call the OLS estimator of β_1 for the above model as $\tilde{\beta}_1$. Compare $E(\tilde{\beta}_1)$ and β_1 .) (5pts).

$$E(\tilde{\beta}_1) < \beta_1.$$

d) Continuation from the previous question. We continue to assume $\beta_2 > 0$ but do not assume $Cov(x_1, x_2) < 0$. State under what additional assumption, $\tilde{\beta}_1$ is an unbiased estimator. Prove the unbiasedness of the $\tilde{\beta}_1$ under your additional assumption. (10pts.)

If we assume $E(u | x_1, x_2) = 0$, and $E(x_2 | x_1) = 0$ or some other constant, we can show the unbiasedness of the $\tilde{\beta}_1$.

Proof.

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1) y_i}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} = \beta_1 + \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1) v_i}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}.$$

$$\begin{aligned}
E(\tilde{\beta}_1) &= E\left[E\left(\beta_1 + \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)v_i}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \mid x_1\right)\right] \\
&= \beta_1 + E\left[E\left(\frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)v_i}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \mid x_1\right)\right] \\
&= \beta_1 + E\left[E\left(\frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)v_i}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \mid x_1\right)\right] \\
&= \beta_1 + E\left[E\left(\frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} + u_i)}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \mid x_1\right)\right] \\
&= \beta_1 + E\frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)[E(x_{2i} + u_i \mid x_1)]}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \\
&= \beta_1 + E\frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)[E(x_{2i} + u_i \mid x_1)]}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \\
&= \beta_1 + E\frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)E(x_{2i} \mid x_1) + \sum_{i=1}^n (x_{1i} - \bar{x}_1)E(u_i \mid x_1)}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \\
&= \beta_1 + E\frac{0 + \sum_{i=1}^n (x_{1i} - \bar{x}_1)E[E(u_i \mid x_1, x_2)]}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \\
&= \beta_1 + E\frac{0 + \sum_{i=1}^n (x_{1i} - \bar{x}_1)E[0]}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \\
&= \beta_1
\end{aligned}$$

2. Multiple Regression Model: Inference

A researcher is interested in the structure of wage determination among workers in Tsukuba city. He hypothesized that the log of hourly wage is the function of the year of education, the year of job experience, its squared and a dummy variable that indicates female.

A linear multiple regression model is given as:

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \beta_4 \text{female} + u$$

and he, in addition, assumed $u \sim \text{Normal}(0, \sigma^2)$. He obtained a random sample of workers of Tsukuba residents from a national survey conducted in 2002 and estimated the parameters by OLS. The estimated results were

$$\log(\hat{\text{wage}}) = -3.59 + 0.12 \text{educ} + 0.06 \text{exper} - 0.001 \text{exper}^2 - 0.43 \text{female},$$

$n=520, R^2=0.37$

where standard errors are in parenthesis.

a) Test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_1 : \beta_1 \neq 0$ at 5% significance level using t-test. Use the attached statistical table to look up the critical value. (5pts)

$t=0.12/0.013=9.23$. This exceeds the critical value 1.96, thus we reject the null hypothesis.

b) Interpret the estimated coefficient for education, which is 0.12. What does this mean? (Hint: log is taken for dependent variable.) (5pts)

When the years of education increases by one year, the hourly rate of pay increases by about 12 percent.

c) Interpret the estimated coefficient for female, which is -0.43. (5pts)

Holding the years of education and years of job experience constant, female workers earn about 43 percent less than male workers.

d) Holding the years of education and sex of worker constant, at which year of experience, workers experience maximum value of $\log(\text{wage})$? In other words, what is the year of experience at the turn-around point? Suppose workers start to work at the age of 22. Does your answer make sense from your casual observation? (5pts)

Turn around point = $-(0.06/2*-0.001) = 30$. Those workers who have 30 years of job experience starts to experience hourly rate of pay decline. The hourly rate of pay starts to decline at age 52 and this makes sense.

e) When a researcher concludes that the negative estimated coefficient for female dummy implies the discrimination against female workers in the labor market, another researcher criticized that conclusion pointing out the possibility that female dummy can be endogenous. Why the female dummy could be endogenous in the above model. (5pts)

The error term may include the difference of productivity in the labor market between sexes and this makes the female dummy endogenous.

f) An economic theory of marriage predict that marriage makes male workers more productive while it makes female workers less productive in the labor market because male workers exert more effort in labor market relieved from household duties and female workers exert less effort in labor market due to additional household duties. A researcher attempts to test this economic theory by including the dummy variables that indicates whether a worker is married and its interaction term with female dummy. The results of OLS estimation with these dummy variables are:

$$\begin{aligned} \log(\widehat{wage}) = & -3.71 + 0.12 \text{educ} + 0.07 \text{exper} - 0.001 \text{exper}^2 - 0.11 \text{female} \\ & \quad \quad \quad (0.21) \quad (0.01) \quad (0.008) \quad (0.0001) \quad (0.10) \\ & + 0.25 \text{married} - 0.47 \text{female} * \text{married} \\ & \quad \quad \quad (0.10) \quad (0.12) \end{aligned}$$

$n=519, R^2=0.39.$

Test the null hypothesis that these two additional variables can be excluded from the model at 5% significance level. You should use F-test. Use the attached statistical table to look up the critical value.

(i.e. $H_0 : \beta_4 = 0, \beta_5 = 0$ where β_4 is the coefficient for the married dummy and β_5 is the coefficient for the interaction term of female and married dummy variables. Denominator degree of freedom can be approximated as infinity.) (5pts).

We use the formula based on R^2 .

$$\begin{aligned} F &= \frac{(SSE_r - SSE_{ur}) / q}{SSE_{ur} / (n - k - 1)} \\ &= \frac{(R_{ur}^2 - R_r^2) / q}{(1 - R_{ur}^2) / (n - k - 1)} \\ &= \frac{(0.39 - 0.37) / 2}{(1 - 0.39) / (519 - 6 - 1)} \\ &= 8.39 \end{aligned}$$

This exceeds the critical value of $F(2, \quad) = 3.00$. Thus we reject the null hypothesis.

g) Holding the years of education and experience constant, how much do married women earn less than single women? On the other hand, how much do married men earn more than single men? Answer using approximate percentage. Do this estimation results support the proposed economic theory? (5pts)

Married women earn less than single women by about 47 percent, while married men earn more than single men by about 25 percent. These results are consistent with the proposed economic theory.

2. Applied Part (20pts in total)

An economist regressed the log of birth weight of new born baby on the number of cigarettes smoked during the mother's pregnancy, log of family income. The results of estimation obtained by STATA were following. The t values for *cigs* are suppressed as XXX and the confidence interval for *cigs* are suppressed as YYY and ZZZ.

```
. reg lbwght cigs lfaminc
```

Source	SS	df	MS			
Model	1.29879229	2	.649396147	Number of obs =	1388	
Residual	49.1215413	1385	.035466817	F(2, 1385) =	18.31	
Total	50.4203336	1387	.036352079	Prob > F =	0.0000	
				R-squared =	0.0258	
				Adj R-squared =	0.0244	
				Root MSE =	.18833	

lbwght	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cigs	-.0040816	.0008582	XXX	0.000	YYY	ZZZ
lfaminc	.0162657	.0055833				
_cons	4.718594	.0182445				

a) Calculate the t statistics (XXX) from the given information. (5 pts)

$$-0.0040816/0.0008582 = -4.76$$

b) Calculate the 95 percent confidence interval (YYY and ZZZ) for the coefficient for *cigs*. Note that the sample size is large enough to assume that the estimated coefficient is normally distributed. Use the attached statistical table to look up the critical value. (5 pts)

$$YYY = -0.0040816 - 1.96 * 0.0008582 = -0.00576367$$

$$ZZZ = -0.0040816 + 1.96 * 0.0008582 = -0.00239953$$

c) Now we use the number of packs of cigarettes as an explanatory variable. What is the estimated coefficient (XXX), and t-statistics (YYY) ? Note that one pack of cigarettes contains 20 cigarettes. (10 pts)

```
. reg lbwght packs lfaminc
```

Source	SS	df	MS	Number of obs =	1388
Model	1.2987923	2	.64939615	F(2, 1385) =	18.31
Residual	49.1215413	1385	.035466817	Prob > F =	0.0000
				R-squared =	0.0258
				Adj R-squared =	0.0244
Total	50.4203336	1387	.036352079	Root MSE =	.18833

lbwght	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
packs	XXX		YYY	0.000	
lfaminc	.0162657	.0055833		0.004	
_cons	4.718594	.0182445		0.000	

The coefficient become 20 times larger than the original coefficient. XXX=20* -.0040816=-.081632
 The t-statistics should be preserved. YYY=-4.76

3. Advanced Part (20pts in total)

Consider a simple linear regression model:

$$y = \beta_0 + \beta_1 x + u$$

and we obtain random sample from the population and x has variation in the sample. We assume $E(u|x)=0$. The variance of u is constant; $Var(u | x) = \sigma^2$.

a) Write down the formula of OLS estimator of β_1 , which is $\hat{\beta}_1$. (5pts)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

b) Show that $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$ given the values of x_i for $i=1, 2, \dots, n$ under the random

sampling assumption, which is $Cov(u_i, u_j) = 0$, for all $i \neq j$. (10 pts)

$$\begin{aligned} Var(\hat{\beta}_1) &= Var\left(\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\ &= Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\ &= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} Var\left(\sum_{i=1}^n (x_i - \bar{x}) u_i\right) \\ &= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} Var\left((x_1 - \bar{x})u_1 + (x_2 - \bar{x})u_2 + \dots + (x_n - \bar{x})u_n\right) \\ &= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} \left[(x_1 - \bar{x})^2 Var(u_1) + (x_2 - \bar{x})^2 Var(u_2) + \dots + (x_n - \bar{x})^2 Var(u_n) \right. \\ &\quad \left. + 2(x_1 - \bar{x})(x_2 - \bar{x})Cov(u_1, u_2) + 2(x_1 - \bar{x})(x_3 - \bar{x})Cov(u_1, u_3) + \dots + 2(x_{n-1} - \bar{x})(x_n - \bar{x})Cov(u_{n-1}, u_n) \right] \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

c) Suppose we fail to obtain random sample from the population and $Cov(u_i, u_j) > 0$, for all $i \neq j$. Is $Var(\hat{\beta}_1)$ larger or smaller than the one calculated in b)? Explain your answer based on the results obtained in b). (5pts)

We cannot say whether the variance is smaller or larger than the one calculated in b). It depends on the sign of

$$2(x_1 - \bar{x})(x_2 - \bar{x})Cov(u_1, u_2) + 2(x_1 - \bar{x})(x_3 - \bar{x})Cov(u_1, u_3) + \dots + 2(x_{n-1} - \bar{x})(x_n - \bar{x})Cov(u_{n-1}, u_n)$$

$$= 2 \sum_{i=1}^{n-1} [(x_i - \bar{x}) \sum_{j=i+1}^n (x_j - \bar{x}) Cov(u_i, u_j)]$$