Final Exam

Economics 401

Fri., May 15, 2008

Show All Work. Only partial credit will be given for correct answers if we can not figure out how they were derived.

Points:

Problem 1:	30
Problem 2:	15
Problem 3:	20
Problem 4:	15
Problem 5:	20
Total:	100

Problem 1: Consider a job training program that gives training to poor workers on basic skills and how to find jobs. You have data on high school dropouts from Chicago from 2005-2007. In particular you have the following three variables:

h_{2007i} : Number of Weeks worked in 2007

T_i : A dummy variable for whether the person received job training in 2006 (it takes value 1 if person i took training, and value 0 otherwise)

h_{2005i} : Number of weeks worked in 2005

You first run the regression

$$h_{2007i} = \beta_0 + \beta_1 T_i + u_i$$

You get the results

Variable	Parameter	Standard Error
Intercept $\left(\widehat{\beta}_0\right)$	23.45	7.84
$T_i \left(\widehat{\beta}_1\right)$	-3.44	2.11

a) Intepret the point estimate of the slope coefficient $\left(\widehat{\beta}_1\right)$ from this regression in a descriptive way. What does it mean?

The coefficient of -3.44 means that people who received job training worked 3.44 fewer weeks in 2007 than workers that did not take job training.

b) Worrying about precision of the estimate how does that modify your interpretation of a) (that is how confident are you)?

The standard error of this coefficient is 2.11 meaning the 95% confidence interval is

$$(-3.44 - 1.96 \times 2.11, -3.44 + 1.96 \times 2.11) = (-7.58, 0.70).$$

I am confident that the difference in weeks worked between trained and untrained individuals lies somewhere between -7.58 and 0.70.

c) Explain the meaning of the point estimate of the intercept $\left(\widehat{\beta}_0\right)$ from a descriptive perspective.

We know that

$$E(h_{2007i} \mid T_i = 0) = \beta_0.$$

Thus the average number of weeks worked for workers who did not receive the training is estimated to be 23.45.

d) Now suppose you want to interpret the coefficient β_1 causally. What is the biggest problem with this interpretation?

The key assumption of causality is that T_i is uncorrelated with u_i which is the error term in the regression model. This would mean that whether people receive training or not is uncorrelated with unobservable variables that determine weeks worked. There are many reasons to think that training would be related, and it could be either positive or negative. For example, people who have good stable jobs will have high u_i and will not need job training. This would lead to a negative correlation. Alternatively, one might think lazy people will tend to work relatively little and would also be unlikely to enter job training. This would yield a positive correlation.

Now consider the new regression

$$h_{2007i} = \beta_0 + \beta_1 T_i + \beta_2 h_{2005i} + u_i$$

e) Does this help the problem you described in d)? What would you expect to happen to the point estimates and why?

It helps for either case. It would work particularly well for my first concern. If the problem is that people with steady jobs tend to work a lot and are unlikely to take training, controlling for weeks worked in 2005 should help the problem substantially.

Now suppose you get the regression results.

Variable	Parameter	Standard Error
Intercept ($\widehat{\beta_0}$)	2.15	1.73
$T_i(\widehat{\beta_1})$	8.44	1.88
$h_{2005i}(\widehat{\beta_2})$	0.96	0.33

f) Interpret what you learn about the effect of training $\left(\widehat{\beta_1}\right)$ I have left this question intentionally vague. Please give a complete answer of what you think we learn about the effectiveness of training from this.

In this case it looks like training has a positive effect on hours worked. If I just interpret this as a descriptive number it says that if we compare two people who worked the same amount in 2005, the person who received the training works about 8.44 more weeks in 2007. The confidence interval on this effect is

$$(8.44 - 1.96 \times 1.88, 8.44 + 1.96 \times 1.88) = (4.76, 12.12)$$

The lower end of the confidence interval is 4.76 which still seems pretty big. Thus it looks like the difference between those that got training and those that did not is large. If one interpreted the point estimate as a causal effect, one would say that training increases work hours by 8.44 weeks per year. This requires the assumption that training is uncorrelated with the error term. While controlling for hours worked in 2005 helps get rid of some of the omitted variable bias, I would think that there are still others to worry about, so I would be a bit cautios in interpreting this effect as causal.

Problem 2: Suppose you have data on height of men and women from France and England. Let

F_i : Dummy variable for a person from France

m_i : Dummy variable for male

H_i : Height of person i (in centimeters)

Consider the following two regressions:

$$H_i = \gamma_0 + \gamma_1 F_i + \gamma_2 m_i + \gamma_3 (F_i \times m_i) + \varepsilon_i$$

Derive $\gamma_0, \gamma_1, \gamma_2$, and γ_3 in terms of the underlying conditional expectations.

$$\begin{aligned} E(H_i \mid \text{French Male}) &= \gamma_0 + \gamma_1 + \gamma_2 + \gamma_3 \\ E(H_i \mid \text{English Male}) &= \gamma_0 + \gamma_2 \\ E(H_i \mid \text{French Female}) &= \gamma_0 + \gamma_1 \\ E(H_i \mid \text{English Female}) &= \gamma_0 \end{aligned}$$

So

$$\begin{aligned} \gamma_0 =& E(H_i \mid \text{English Female}) \\ \gamma_1 =& E(H_i \mid \text{French Female}) - E(H_i \mid \text{English Female}) \\ \gamma_2 =& E(H_i \mid \text{English Male}) - E(H_i \mid \text{English Female}) \\ \gamma_3 =& E(H_i \mid \text{French Male}) - E(H_i \mid \text{French Female}) \\ & - [E(H_i \mid \text{English Male}) - E(H_i \mid \text{English Female})] \end{aligned}$$

Problem 3: Suppose you have monthly data on stock prices and the unemployment rate in the U.S. over the last 10 years.

Consider the regression:

$$Stock_t = \beta_0 + \beta_1 unemp_t + v_t.$$

Assume there is not omitted variable bias so that $E(v_t \mid unemp_t) = 0$.

Suppose you estimate it in stata in the following 4 ways

- reg stock unemp
- reg stock unemp, robust
- arima stock unemp, ar(1 2) ma(1 2)
- newey stock unemp, l(5)

Discuss the advantages and disadvantages of the approaches. Which are biased, which give correct standard errors, which work best under the correct assumptions?

Methods (1),(2), and (4) will give the exact same point estimates, but different standard errors. Method (1) requires homoskedasticity and no serial correlation. Method (2) allows for heteroskedasticity but not serial correlation. Method (4) allows for both serial correlation and heteroskeasticity. Method (3) will be efficient when the model is correctly specified. That is if the model is truly an ARMA(2,2) the estimates will be correct and efficient, and the other models will not be. However, if those assumptions are violated the estimates will be wrong.

Problem 4: Think about estimating the causal effect of private school on wages.

You have the following 4 variables

w_i : log of wage at age 30

PS_i : years attended private high school

$Coll_i$: years of college attended

PE_i : Parent's earnings when respondant was 16

Consider the following three regressions:

$$\text{(A)} \qquad w_i = \beta_0 + \beta_1 PS_i + \beta_2 Coll_i + u_i$$
$$\text{(B)} \qquad w_i = \gamma_0 + \gamma_1 PS_i + \gamma_2 PE_i + v_i$$
$$\text{(C)} \qquad w_i = \delta_0 + \delta_1 PS_i + \delta_2 Coll_i + \delta_3 PE_i + \varepsilon_i$$

Which regression makes the most sense? Why do you prefer this to the other two?

I prefer model (B). The problem is that college is endogenous. That is one of the major advantages of attending a private school is that one would be able to get into college. Thus I would not want to control for it. I can not make the same argument about Parent's earnings so controlling for this is sensible.

Problem 5: Consider the following regression model

$$Y_i = \beta_0 + \beta_1 w_i + \beta_2 x_i + u_i$$

Suppose we know that

$$\begin{aligned} E\left(u_i\right) &= 0 \\ E(u_i \mid z_i) &= 0 \\ E(u_i \mid x_i) &= 0 \end{aligned}$$

where z_i is another variable you have in your data.

(A) What three equations would you use to derive estimates of β_0, β_1, and β_2?

I would write the three population equations as:

$$\begin{aligned} E(u_i) &= 0 \\ E(z_i u_i) &= 0 \\ E(x_i u_i) &= 0 \end{aligned}$$

Translating to sample analogoues yields

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} Y_i - \widehat{\beta_0} - \widehat{\beta_1} w_i - \widehat{\beta_2} x_i &= 0 \\ \frac{1}{N}\sum_{i=1}^{N} z_i(Y_i - \widehat{\beta_0} - \widehat{\beta_1} w_i - \widehat{\beta_2} x_i) &= 0 \\ \frac{1}{N}\sum_{i=1}^{N} x_i(Y_i - \widehat{\beta_0} - \widehat{\beta_1} w_i - \widehat{\beta_2} x_i) &= 0 \end{aligned}$$

(B) Now suppose you replace those equations with

$$\begin{aligned} E\left(u_i\right) &= 3 \\ E(u_i - 4z_i \mid z_i) &= 2 \\ E(u_i - 3x_i^2 | x_i) &= 0 \end{aligned}$$

what equations would you use in this case?

I would use the three population equations as:

$$\begin{aligned} E(u_i) &= 3 \\ E(z_i[u_i - 4z_i]) &= 2 \\ E(x_i[u_i - 3x_i^2]) &= 0 \end{aligned}$$

The sample analogues are:

$$\frac{1}{N}\sum_{i=1}^{N} Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 w_i - \widehat{\beta}_2 x_i = 3$$

$$\frac{1}{N}\sum_{i=1}^{N} z_i(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 w_i - \widehat{\beta}_2 x_i - 4z_i) = 2$$

$$\frac{1}{N}\sum_{i=1}^{N} x_i(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 w_i - \widehat{\beta}_2 x_i - 3x_i^2) = 0$$