

## Econometrics 2. Sample Questions

Winter 2010

1. Comment on the following statement: proxy variables are of no use because they are basically RHS variables with measurement error, and therefore they lead to attenuation bias to the coefficients.
2. In the Neal-Johnson paper the log earnings of young American men is regressed on age and a Black and a Hispanic dummy. Here are the results (appropriately estimated standard errors in parentheses):

$$\widehat{\log(w)} = c + \underset{(0.014)}{0.48} \textit{age} - \underset{(0.026)}{0.244} \textit{black} - \underset{(0.030)}{0.113} \textit{hispanic}$$

- (a) What is the meaning of the *black* coefficient?
- (b) Does the coefficient on the black variable measure the extent of labor market discrimination? Why or why not?

Neal and Johnson re-estimated their regression with comprehensive test scores (*AFQT*). The test scores were measured at around age 18. They contain many skill elements, including IQ. *AFQT* is normalized (mean=0, std=1) and is entered into the regression in a quadratic form.

- (c) Do you think *AFQT* can be a proxy for unobservables that cause trouble in the original regression? If not, why? If yes, state exactly what it would proxy.
- (d) Can it be a perfect proxy?

Here are the results (appropriately estimated standard errors in parentheses):

$$\widehat{\log(w)} = c + \underset{(0.013)}{0.40} \textit{age} - \underset{(0.027)}{0.072} \textit{black} + \underset{(0.030)}{0.005} \textit{hispanic} + \underset{(0.012)}{0.172} \textit{AFQT} - \underset{(0.011)}{0.013} \textit{AFQT}^2$$

- (e) How do you interpret the coefficients on the *AFQT* variables?
- (f) The *black* coefficient is different here then in the previous estimation. Why?
- (g) Does this coefficient on the black variable measure the extent of labor market discrimination? Why or why not? If not, is this one closer to it?
- (e) Neal and Johnson argue that education (whether the individual achieves college education or not) would not a good control variable, that's why they don't include it here. Why would education be endogenous?

3. You would like to know whether privatization increases firm productivity. You look at a cross-section of state-owned firms in, say, 1995, and regress their productivity change to the next period (say, 2000) on whether they were privatized in the meantime (a dummy).
  - (a) The parameter estimate on the privatized dummy is inconsistent if firms that get privatized have higher productivity growth regardless of privatization (and that is well forecasted by potential buyers in 1995). Why, and what is the direction of the bias?
  - (b) Someone tells you that you should use the fraction of higher educated workers in the firm in 1995 as a proxy variable for eliminating the bias. Is that a good idea? (Under what conditions is that a good idea, and do you think those conditions are satisfied here?)

4. Suppose that you would like to estimate whether elite high schools add more value to their students than other high schools. Your data contains a test score on student competence measured at the end of high school, gender, parental education, and whether the student graduated from an elite high school.
  - (a) Write down a regression model that may enable you to estimate the effect of elite schools on student's achievement.
  - (b) State the assumptions under which OLS consistently estimates the effect in this regression. Do you think they are satisfied in this case? Derive the plim of the OLS estimator as if there were a single explanatory variable, and argue whether the asymptotic bias is zero, negative or positive.
  - (c) Could you use elementary school grade point average as a proxy variable for unobservables? Would its inclusion lead to consistent estimation of the effect?
  - (d) Would elementary school grade point average be a valid instrument? Would its use lead to consistent estimation of the effect?
  - (e) Suppose that you know where students lived before they enrolled to high-school. Do you think that their distance to the closest elite high-school is a valid instrument?
5. The question is returns to education. Education is endogenous because of unobserved ability. Give an example for a proxy variable. (State the definition of a proxy and argue that your choice satisfies it). Does including it into the model make the OLS estimator of the parameter on education consistent for the returns to education?
6. True or false? Consider a simple regression model where all the classical assumptions hold, but we can measure the left-hand side variable with a classical error (i.e. the error is independent of everything of importance). Then the OLS estimator for the slope parameter is going to be consistent (i.e. consistent for the slope parameter of the regression with the true left-hand side variable).
7. True or false? Consider a simple regression model where all the classical assumptions hold, but we can measure the right-hand side variable with a classical error (i.e. the error is independent of everything of importance). Then the OLS estimator for the intercept parameter is going to be consistent (i.e. consistent for the intercept parameter of the regression with the true right-hand side variable).
8. Comment on the following statement: measurement error in the dependent variable of a linear regression does not make OLS inconsistent.
9. True or false? Classical measurement error in a right-hand side variable leads to a negative bias in the OLS estimator. (Derive it).
10. A regression of earnings on IQ gives you an  $R^2$  of 0.16. Assume that IQ approximates intelligence with a classical measurement error.
  - (a) What is the correlation between earnings and IQ?
  - (b) Would the true regression coefficient on intelligence be equal to, smaller or larger than the coefficient on IQ?
  - (c) Would the true correlation between intelligence and earnings be equal to, smaller or larger than the one between IQ and earnings?
  - (d) Instead of regressing earnings on IQ, assume that you run the reverse regression. Would the true regression coefficient on earnings (i.e. when intelligence is the LHS variable) be equal to, smaller or larger than your estimate?

- (e) Does your answer to part (d) imply that calculating the correlation coefficient from the  $R^2$  of the reverse regression is a better way than from the original regression?
11. Consider a simple regression model where all the classical assumptions hold, but we can measure the RHS variable with a classical error (i.e. the error is independent of everything of importance).
- (a) What are the properties of the OLS estimator for the slope parameter? How does its probability limit compare to the slope parameter of the regression on the error-free variables?
- (b) What are the properties of the OLS estimator for the intercept parameter? How does its probability limit compare to the intercept parameter of the regression on the error-free variables?
- (c) How does the  $R^2$  compare to the  $R^2$  of the regression on the error-free variables?
12. Consider a simple regression on a right-hand side variable that is measured with classical error. We would like to estimate  $\beta$  in

$$y_i = \alpha + \beta x_i^* + \varepsilon_i \quad E(\varepsilon_i | x_i^*) = 0$$

but instead of  $x^*$ , we can measure  $x_i = x_i^* + w_i$  such that  $E(w_i | x_i^*) = 0$ . The estimated equation is then

$$y_i = \alpha + \beta x_i + u_i$$

- (a) Derive the inconsistency of  $\hat{\beta}_{OLS}$ , the OLS estimator of  $\beta$ .
- (b) For a variable  $z_i$  to be valid instrument, what conditions does it need to satisfy and why?
- (c) Assume that conditions in (b) are satisfied for some variable  $z_i^*$ , but you can observe only an error-ridden measure of it,  $z_i$  such that  $z_i = z_i^* + \omega_i$ ,  $Cov(\omega_i | x_i^*) = Cov(\omega_i | \varepsilon_i) = 0$ . Is  $z_i$  a valid instrument?
13. You would like to estimate how GPA in a given term is affected by average daily time spent on studying (in hours) in that term. You collect data from your fellow students and estimate the following regression:

$$\widehat{GPA} = 3.0 + 0.1 \text{hours}$$

- (a) What is the interpretation of the constant?
- (b) In what direction is the estimated effect biased if people report their average hours with a classical error?
- (c) In what direction is the constant biased in that case?
- (d) What is the sign of the correlation between reporting error and true hours if people who study less than average tend to overreport, and people who study more than average tend to underreport their hours?
- (e) Under (d), would the bias to the estimated effect and the constant change relative to (b) and (c), and in what direction?

14. What is the difference between strict and contemporaneous exogeneity in a time-series regression? Which one is necessary for consistency of OLS? Is that also sufficient?
15. Consider the following stochastic process.

$$Y_t = 0.5\varepsilon_{t-1} + 0.5\varepsilon_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2).$$

- (a) What is the mean and variance of  $Y_t$ ?
  - (b) What are the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> order autocorrelation coefficients?
  - (c) Is the process (covariance) stationary?
16. Is the first difference of a white noise process ( $\Delta\varepsilon_t = \varepsilon_t - \varepsilon_{t-1}$  where  $\varepsilon_t \sim WN(0, \sigma^2)$ ) stationary? Is it white noise? Why or why not?
  17. Is a random walk process stationary? Is the first difference of a random walk process stationary?
  18. Comment on the following statement: If we estimate a distributed lag model by OLS, autocorrelation in the error term makes the parameter estimates inconsistent.
  19. Comment on the following statement: The Durbin-Watson test is useless if there is lagged dependent variable on the right-hand side.
  20. What can you infer from a Durbin-Watson statistic of 3?
  21. Comment on the following statement: The Durbin-Watson statistic is useless because it is based on assumptions that are rarely satisfied in time series regressions.
  22. Comment on the following statement: A Durbin-Watson statistic below 2 shows negative serial correlation because it is a statistic based on the AR(1) coefficient of the error term.
  23. Consider a time series process that follows a random walk with drift. Is the first difference of that time series stationary? Is it weakly dependent?
  24. Comment on the following statement: Coefficients of an autoregressive model can be estimated in an unbiased way by OLS, as long as all right-hand side variables are strictly exogenous.
  25. Comment on the following statement: A time-series regression on non-stationary variables cannot produce unbiased estimates. That's the reason why stationarity is important in time series regressions.
  26. What are the consequences of estimating regressions on trending data?
  27. If  $y_t$  is white noise with mean  $\mu$  and variance  $\sigma^2$ , what process does its difference ( $\Delta y_t = y_t - y_{t-1}$ ) follow? What is its mean, variance and first and second order autocorrelation?
  28. If  $y_t$  is an MA(1) process  $y_t = \phi\varepsilon_{t-1} + \varepsilon_t$  where  $\varepsilon_t \sim WN(0, \sigma^2)$ , what process does its difference ( $\Delta y_t = y_t - y_{t-1}$ ) follow? What is its mean, variance and first and second order autocorrelation?

29. You are interested in the causal effect of  $x$  on  $y$ . What can you infer about the causal effect from the following results (Newey-West  $t$ -values in parentheses), and what additional assumptions you may need for causal inference:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$$

(3.1)      (4.2)

if  $y$  and  $x$

- (a) are trending?
  - (b) are stationary?
  - (c) follow seasonal variation?
  - (d) follow a random walk?
  - (e) follow a random walk with drift?
  - (f) are white noise?
30. Are weakly dependent time series also stationary?
31. Are stationary time series also weakly dependent?
32. You want to regress  $y_t$  on  $x_t$ . Each follows a linear (deterministic) trend. One friend says you should first detrend each series and run regression on detrended variables. Another friend says you should run  $y$  on  $x$  as they are but include a deterministic trend in your regression. Who is right?
33. You want to regress  $y_t$  on  $x_t$ . Each follows a (deterministic) seasonal variation. One friend says you should first take out the seasonal variation by running each variable on seasonal dummies, and then run the regression on the residuals of these regressions. Another friend says you should run  $y$  on  $x$  as they are but include seasonal dummies in your regression. Who is right?
34. Show that the random walk process is not stationary.
35. Show that an MA(1) process is (covariance) stationary. Is it weakly dependent?
36. Consider a time series process that follows a deterministic linear trend plus a zero-mean white noise. Is the first difference of that time series stationary? (Derive all of its first and second moments.) Is it weakly dependent?
37. The following regression model is estimated on a sample of stationary variables ( $t = 1, \dots, 50$ ):

$$y_t = \beta_0 + \beta_1 x_t + u_t,$$

$x$  is strictly exogenous, and the Durbin Watson test implies positive serial correlation.

- (a) are (a1) the point estimates and (a2) the conventional standard error estimates (the ones computer packages produce by default ) unbiased and consistent? Why?
  - (b) Would you estimate something in a different way? What and how?
  - (c) What are the conditions for the Durbin-Watson test?
  - (d) What test would you carry out if the conditions in (c) are not satisfied?
38. The following regression model is estimated on a sample of size 50 of stationary variables:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + u_t,$$

and the Breusch-Godfrey Serial Correlation LM test statistic (with 3 lagged variables) rejects its null hypothesis. Are the point estimates and conventional standard error estimates unbiased and consistent?

39. Consider the following regression and a sample of stationary variables:

$$y_t = \beta_0 + \beta_1 x_t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad |\rho| < 1, \varepsilon_t \sim WN(0, \sigma^2).$$

- (a) What are the properties of OLS estimate  $\beta_1$ ?  
 (b) and its standard error?

40. Consider the following regression and a sample of stationary variables:

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad |\rho| < 1, \varepsilon_t \sim WN(0, \sigma^2).$$

- (a) What are the properties of OLS estimate  $\beta_1$ ?  
 (b) and its standard error?

41. A 1<sup>st</sup> year CEU MA student wants to analyze how the program affects his own sleeping habits. For the entire Fall semester, he keeps a diary about how much he slept, together with some covariates. When the semester is over, he runs the following regression:

$$sleep_t = \beta_0 + \beta_1 psdue_t + \beta_2 wend_t + \beta_3 party_t + \beta_4 t + u_t$$

where  $sleep_t$  is sleeping time in hours on the night of day  $t$ .  $t = 1 \dots T$ , from the first to the last day of the semester.  $psdue_t$  is a dummy that is 1 if there was a problem set due the next day;  $wend_t$  is a dummy indicating whether next days was a week-end day (and so one could sleep until late), while  $party_t$  is a dummy indicating whether there was a party on the given night.

- (a) What is the meaning of  $\beta_0$ ? Of  $\beta_2$ ?  
 (b) What is the sign of  $\beta_4$  if people tend to sleep less and less as final exams approach?  
 (c) If parties tend to be on week-ends, would leaving out  $party_t$  from the regression make the OLS estimator for  $\beta_2$  larger, smaller or not change?  
 (d) Suppose that one tends to make up for unusually little sleep on one night by sleeping more on the next night, and vice versa. What does that imply for the serial correlation in  $u_t$  (positive, negative, or zero)?  
 (e) What would be the value of the Durbin-Watson statistic based on your answer in (d)?  
 (f) What are the assumptions behind the Durbin-Watson test? How would you test for serial correlation in the error term if those assumptions are not likely to be satisfied? State the null and alternative hypotheses of the test, and name the test (for extra points, you may give more details about the test).
42. Suppose you have estimated a linear regression on a sample of stationary time-series variables by OLS, and the Breusch-Godfrey serial correlation LM test rejects its null hypothesis (but everything else conforms the usual assumptions). What are the properties of the point estimates and the t-tests on them if
- (a) you estimated White standard errors?  
 (b) you estimated Newey-West standard errors?  
 (c) you used simple standard errors that regression programs (incl. EViews) give by default?  
 (d) Do you need to do something in a different way in order to get efficient point estimates? If yes, what? If not, why not?

43. The following demand equation was estimated on a seasonally adjusted monthly time-series of mineral water consumption ( $n = 60$ ). The estimated parameters are in the equation, and their appropriately estimated standard error below in parentheses.

$$\log(Q_t) = \underset{(0.20)}{1.5} - \underset{(0.10)}{0.75} \log(P_t) + \underset{(0.20)}{0.25} \log(P_t^*) + u_t,$$

where  $Q_t$  is mineral water sold in million gallons,  $P_t$  is price, and  $P_t^*$  is the price of soda beverage like Coca Cola (both in real terms). Assume that all variables (their logs) are stationary, and variation in prices were exogenous to demand if seasonally adjusted.

- Test whether mineral water and other soda beverages are substitutes.
  - Test whether the demand is price-elastic.
  - Can you test whether a simultaneous 1 per cent increase of the price of mineral water and other soda beverages would lead to a decrease in mineral water consumption? If yes, what is the result? If no, describe a test procedure step by step.
  - Would there be problems if we used seasonally unadjusted series? If yes, what and why?
44. Is the sum of two stationary AR(1) processes also an AR(1) process?  
Hint: write

$$\begin{aligned} Y_t &= c_y + \rho_y Y_{t-1} + \varepsilon_t, & \varepsilon_t &\sim WN(0, \sigma_y^2) \\ X_t &= c_x + \rho_x X_{t-1} + \omega_t, & \omega_t &\sim WN(0, \sigma_x^2) \end{aligned}$$

Define  $Z_t = X_t + Y_t$ . Can you write  $Z_t = c_z + \rho_z Z_{t-1} + \nu_t$  such that  $\nu_t$  is white noise?

45. What process does the difference of a White Noise follow? Is it stationary? How about the difference of its difference?
46. The question is the effect of child-related benefits (government transfers) on fertility. The data is yearly time series from a particular country. Consider first the regression of the following form:

$$y_t = \alpha + \beta x_{t-1} + u_t$$

where  $y_t$  is a measure of fertility in a country (say, number of births per women in a certain age range), and  $x_t$  is a measure of child-related government transfers in the country (total government budget allocated to such transfers in real terms, divided by the number of children). Assume moreover that both  $y_t$  and  $x_t$  follows a stochastic trend (random walk with drift).

- Why do we include a lagged right-hand side variable instead of a contemporaneous one?
- Under what assumptions is the OLS estimator for  $\beta$  in the equation above consistent for the causal effect of interest? Are those assumptions are satisfied here?
- Under what assumptions is the OLS estimator for  $\beta$  consistent for the causal effect of interest in the following equation?

$$\Delta y_t = \alpha + \beta \Delta x_{t-1} + \Delta u_t$$

where  $\Delta z_t = z_t - z_{t-1}$  for any variable  $z_t$ ? Are those assumptions are satisfied here?

- What problems does serial correlation in  $\Delta u_t$  cause to the OLS estimation, and how would you treat them?

- (e) Assume that once a while policymakers react to observed decline in fertility (a decline that is due to some outside factors) by increasing child-related benefits. Assume moreover that such a reaction takes one year to materialize. Write down the implied covariance between  $\Delta u$  and  $\Delta x$  in the appropriate lag structure and determine its sign.
- (f) What are the consequences of that correlation for the consistency of the OLS estimator of  $\beta$  if  $\Delta u_t$  follows a white noise process?

47. (a) If random effects and fixed effects estimates are very different, which would you prefer?  
 (b) If first difference and fixed effects estimates are very different, which would you prefer?  
 (c) How about (b) if  $T = 2$ ?
48. You would like to estimate the effect of police size on city-level crime. You have a panel data of many cities through a few years. Pooled OLS results of log crime rate on log police size per population gives a positive estimate.  
 (a) Pooled OLS estimates are arguably inconsistent for the causal effect of police size on crime? Why, and what is the direction of the bias?  
 (b) Some suggest that a fixed-effects model can solve the problem. What do you think?
49. In a linear regression on panel data, when would you estimate "cluster standard errors" and why?
50. Goldin and Katz (2002, JPE) argue that an important reason for college participation of American women increased in the early 1970's was the increased use of contraceptive pills by young single women. According to their argument, "the pill" allowed single young women to delay marriage and invest more into their career. Increased use of "the pill" by single young women was due to the fact that state laws started to allow them only after the late 1960's, 6 states introducing such laws in 1969, 16 other states in 1971, and the rest in 1974 or after. Goldin and Katz used panel data on U.S. states by year. The following linear regression is a good representation of one of their models:

$$M_{sy} = \beta' X_{sy} + \gamma P_{sy} + \alpha_s + \delta_y + u_{sy}$$

where  $s$  is state,  $y$  is year.  $M_{sy}$  is the fraction of women who were married by age 23;  $X_{sy}$  are some control variables;  $P_{sy}$  is a dummy that is one if state  $s$  in year  $y$  allowed young single women to take "the pill";  $\alpha_s$  are state fixed effects, and  $\delta_y$  are year fixed effects. They estimated  $\gamma$  to be negative and significant.

- (a) What is the reason for including state fixed effects?  
 (b) What is the reason for including year fixed effects?  
 (c) What is the meaning of  $\gamma$ ?  
 (d) What was the reason for applying the Fixed Effects method and not Random Effects?  
 (e) What was the reason for applying the Fixed Effects method and not First Differences?  
 (f) What assumptions Fixed Effects estimation needs about  $u_{sy}$  in order to consistently estimate  $\gamma$ ?
51. State-level data from the U.S. is used in order to estimate the deterrence effect of executions on potential murders. The data is a two-period panel, each period referring to a three-year interval.  $m_{it}$ , the state-level murder rate in state  $i$  in period  $t$ , is regressed on the number of executions in the given period,  $x_{it}$ .  
 (a) Under what conditions is the pooled OLS estimator consistent for the deterrence effect?  
 (b) Write down the appropriate panel regression model that includes both state and time fixed-effects.  
 (c) What is the interpretation of the state fixed effects?  
 (d) What is the interpretation of the time fixed effects?  
 (e) Write down the corresponding first-differenced (FD) model.  
 (f) Does the FD model contain state or time fixed-effects? If yes (to any), what is their (its) interpretation?

- (g) Under what conditions is the FD estimator consistent for the deterrence effect?
  - (h) Compare conditions (d) and (g). Which are more likely to be satisfied?
52. In their controversial study, Donohue and Levitt (QJE, 2001) estimate the effect of legalizing abortion on the level of crime years later. Assume for simplicity that crime is committed by 20-year old people only (i.e. neither younger nor older people commit crimes). Then their argument can be summarized as follows: legalized abortion allowed mothers to choose not to have unwanted children who, if born, would have become criminals with a higher likelihood 20 years later. Therefore, legalizing abortion can lead to lower levels of crime 20 years later. In order to check their argument they analyzed the evolution of crime rates across U.S. states and used the fact that some states legalized abortion earlier than others.
- (a) Write down the appropriate panel regression model that can estimate the relationship in question. Include state fixed effects as well as year dummies.
  - (b) What is the interpretation of the state fixed effects?
  - (c) What is the interpretation of the year dummies?
  - (d) The fixed effects estimator is equivalent to an OLS estimator on transformed data. What is the transformation and what does the transformed regression look like?
  - (d) When is the fixed effects estimator preferred to pooled OLS? (Give a formal answer and give an argument in the context of this study.)

53. In order to estimate the effect of  $x$  on  $y$ , you make use of an instrumental variable  $z$ .
- What conditions should  $z$  satisfy in order to have a consistent estimate of the effect of  $x$  on  $y$ ?
  - In your first stage regression, the t-statistic of the coefficient on  $z$  is  $t = 1.9$ . Is that a problem for the instrumental variable estimator for the effect of  $x$  on  $y$ ?
54. In a simple regression with a potentially endogenous right-hand side variable, you have two candidate instruments. One is strong but may be very weakly correlated with the error term, while the other is a lot weaker but it is known to be uncorrelated to the error term. Which instrument would you choose and why? What if the second instrument is so weak that it is practically uncorrelated with the right-hand side variable?
55. Consider the following simple regression model (on *iid* variables):

$$y_i = \alpha + \beta x_i + u_i, \quad \text{Cov}(x_i, u_i) = \gamma V(x_i), \gamma > 0$$

Derive the probability limit (*p lim*) of the OLS estimator for  $\beta$ .  
Suppose that you find a variable  $z$  such that

$$\text{Cov}(z_i, u_i) = 0.$$

Consider the following estimator for  $\beta$  (it is called the Instrumental Variables, or IV estimator):

$$\hat{\beta}_{IV} = \frac{\frac{1}{n} \sum (y_i - \bar{y})(z_i - \bar{z})}{\frac{1}{n} \sum (x_i - \bar{x})(z_i - \bar{z})}.$$

- Derive the probability limit (*p lim*) of  $\hat{\beta}_{IV}$ . Is it consistent for  $\beta$ ?
  - What if  $\text{Cov}(x_i, u_i) > \text{Cov}(z_i, u_i) > 0$ ? Is the IV estimator consistent?
  - If not, can you tell whether the IV or the OLS estimator is more biased (asymptotically)?
56. Consider the following simple regression model:

$$y_i = \alpha^* + \beta^* x_i^* + u_i, \quad \text{Cov}(x_i^*, u_i) = 0$$

but where we don't observe  $x^*$  only its error-ridden measure:

$$x_i = x_i^* + \varepsilon_i, \quad E(\varepsilon_i) = E(x_i \varepsilon_i) = E(y_i \varepsilon_i) = 0$$

so the estimable equation is

$$y_i = \alpha + \beta x_i + v_i.$$

Is the OLS estimator for  $\beta$  consistent for  $\beta^*$ ? Find its probability limit.  
Now suppose that you find a variable  $z$  such that

$$\begin{aligned} \text{Cov}(z_i, u_i) &= \text{Cov}(z_i, \varepsilon_i) = 0 \\ \text{Cov}(z_i, x_i) &\neq 0. \end{aligned}$$

Consider the IV estimator for  $\beta$ :

$$\hat{\beta}_{IV} = \frac{\frac{1}{n} \sum (y_i - \bar{y})(z_i - \bar{z})}{\frac{1}{n} \sum (x_i - \bar{x})(z_i - \bar{z})}.$$

Is the IV estimator of  $\beta$  consistent for  $\beta^*$ ? If not, derive its asymptotic bias.

What if  $Cov(z_i, u_i) \neq 0$ ? Is the IV estimator consistent? for  $\beta^*$  If not, can you tell when the IV estimator is less biased (asymptotically) than the OLS estimator?

57. Evaluate the following statement. If you have an endogenous right-hand side (RHS) variable, it is better to use an instrumental variable (IV) estimator even if the IV is not valid, as long as its correlation with the error term is smaller than correlation between the error term and the endogenous RHS variable.
58. When is an instrument weak and why is that a problem?
59. We are usually worried about ability bias when estimating returns to education. Are the following candidates valid instruments? Why or why not?
  - (a) Distance of parents' home to college.
  - (b) Parental education.
  - (c) Regional differences in compulsory schooling age.
60. You would like to estimate the effect of education on earnings. A regression of log earnings on education (in years) gives an estimate of 0.07. Suppose that the only thing you can't measure is "ability" that makes people get more education and also higher earnings (conditional on education).
  - (a) What is the thought experiment that would measure the causal effect education on earnings?
  - (b) Suppose now that you have to measure the effect on observational data. You regress log earnings on education (and measured stuff). Is the OLS coefficient on education consistent for the causal effect? If not, in what direction are they biased?
  - (c) What assumptions should a valid IV for education satisfy in this case?
  - (d) Do you think distance to school can serve as a valid instrument? Why or why not?
  - (e) How do your answers change to questions (b) to (e) if ability that makes people get more education is independent of the ability that makes them earn more (conditional on education)?
61. You would like to measure the return to education measured in earnings (holding pre-education abilities constant) on a cross-section of individuals, by a linear regression where the LHS variable is log earnings.
  - (a) What is the likely sign of the asymptotic bias of the OLS coefficient on education in a simple regression and why?
  - (b) What if you control for the IQ score of the individual at age 6?
  - (c) Do you think IQ can be a valid instrument? Why or why not?
  - (d) Do you think quarter of birth can be a valid instrument? Why or why not?
  - (e) How do your answers change if education is measured with a classical error?
62. Consider the following regression on a sample of countries:

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

where  $y$  is economic growth and  $x$  is the size of government budget relative to GDP. Suppose that we expect a negative effect of  $x$  on  $y$  *ceteris paribus*. Suppose that there is some unmeasured variable  $w$ , a measure of corruption that affects  $x$  positively and  $y$  negatively (at any level of  $x$ ).

(a) Is the OLS estimator for  $\beta_1$  consistent? If not, what do you think the direction of the (asymptotic) bias will be and why?

(b) Someone proposes using electoral cycles for instrumenting for the government budget. In election years, governments tend to increase spending in order to increase their chances to get re-elected. Formally, the IV would be a dummy, 1 if the given country is in election year, 0 otherwise. What are the formal requirements of the IV to be valid?

(c) Do you think those assumptions are met?

(d) What does it mean for your IV to be weak? What are the consequences of having a weak IV?

63. You would like to estimate the effect of police size on city-level crime on a cross-section of cities. A regression of crime rate on police size (per population) gives an estimate of 0. Suppose that people make their city increase police size if crime increases for some other reason.

(a) Does that lead to a bias the OLS estimates for the causal effect of police size on crime? If yes, in what direction?

(b) Does the estimated 0 coefficient mean that police size has no effect on crime?

(c) What assumptions should a valid IV satisfy in this case?

(d) Do you think electoral cycles can serve as a valid instrument (assuming incumbent city governments spend more on police in election years, *ceteris paribus*)? Why or why not?

(e) How do your answers change to questions (a) to (d) if crime rate is measured with a classical error?

(f) How do your answers change to questions (a) to (d) if crime rate is measured with an error that is negatively correlated with crime rate but uncorrelated with police size?

64. What are the problems with the linear probability model? Are they always problems?
65. You have estimated a linear probability model in order to see whether large multinational firms are more likely to sponsor classical music concerts than other firms. In the model, the LHS variable is 1 if the firm sponsors such events and 0 if not, and the RHS variables are the multinational dummy, the large firm dummy, and their interaction. Someone tells you that your model is not the right one because linear models are no good for estimating probabilities. Would you take the advice and estimate a different model, and if yes, what?
66. Write down a linear probability model for the difference in the probability of migrating from one's country for people with different levels of education. Interpret the coefficients. Can predicted probabilities from this model be negative or greater than one?
67. You estimate a linear probability model for unemployment and you get an  $R^2$  of 0.07. What does that mean?
68. Comment on the following statement. Since the  $R^2$  is not a good measure of fit for probability models, one should use the percent correctly predicted instead.
69. Comment on the following statement. In linear probability models, the  $R^2$  is not a good measure of fit because heteroskedasticity is present in such models.
70. You estimate a logit model for unemployment and you get a PCP (percent correctly predicted) of 92%. The unemployment rate in your sample is 8%. How good is the fit of your model? Would you use some other measure of fit? Why or why not?
71. One of your colleagues ran a probability model (e.g. a probit) and tells you that she got a PCP (percent correctly predicted) of 90%. She asks you what that precisely means. What would you say to her? Would you suggest some other measure of fit? Why or why not?
72. What is the difference between a probit and a logit model? On what basis can you choose which one to use, and why does it matter?
73. Consider a probit model for the effect of previous labor market experience on becoming unemployed (without other control variables). Assume that labor market experience is distributed uniformly in the population from 0 to 40 years. Assume moreover, that the probit coefficient on experience is estimated to be negative, and the predicted unemployment is in the (0, 0.4) interval, and the unconditional unemployment rate in the sample is 0.10 (both the actual and the predicted one).
  - (a) Can you tell whether the partial effect of experience is negative?
  - (b) How does the partial effect of experience change with the level of experience?
  - (c) Can you calculate the value of the conventionally calculated PCP (percent correctly predicted)?
74. Your friend wants to see whether intelligence or motivation (at age 16) is more important for chances to successfully complete high school (at age 18). She estimates a probit model with IQ and a motivation test score on the right-hand side (both standardized). She finds that the probit coefficient on motivation is roughly the same as on IQ. But she is not sure whether comparing these coefficients themselves give an answer to her question or she has to do something with them, and if yes, what. What would be your advise?

75. The question is how education, age and some family circumstances are related to female labor supply. The dataset is an individual cross-sectional sample of women between age 25 and 60, 57 per cent of whom work. Labor supply is modeled as the probability of being employed.
- (a) Write down a latent variable model for a probit model on the probability of employment, with education (in years), age, and age squared on the right-hand side. Assume exogeneity of all right-hand side variables for parts (a) through (e).
- (b) Give an interpretation of the latent variable and the unobserved component.
- (c) What is the partial effect of age on the employment probability?
- (d) The estimates of the probit model are the following (absolute value of t-statistics, also called z-statistics, are in parentheses):

$$-3.83 + 0.11edu + 0.14age - 0.002age^2$$

(3.0)      (4.9)      (2.4)      (2.5)

Can you tell whether the partial effect of age is positive, negative, or changes sign (and if yes, at what age)? Or would you need other information for that (and what information)?

- (e) What is partial effect of education on the employment probability? What is the average partial effect of education if the average of the normal density of the predicted index is 0.35? What does this imply for increasing education from high school (12 years) to college (16 years)?
- (f) If you include the number of young children (younger than 6) in the probit equation, the estimates change to this:

$$0.32 + 0.12edu - 0.006age - 0.0003age^2 - 0.85youngkids$$

(0.7)      (5.7)      (0.1)      (1.0)      (7.7)

and the  $p$ -value of the joint test of the coeff on age and age squared ( $h_0$ : both are zero) is 0.0000. Can you tell now whether the partial effect of age is positive, negative, or changes sign (and if yes, at what age)? Or would you need other information for that (and what information)?

- (g) Is there a qualitative difference between the effect of age in the two specifications? Why or why not?
76. We want to model the *probability of employment among all men between 30 and 60*, conditional on age, race, education (in grades completed), and marital status (1 if married, 0 otherwise). The employment rate in the sample is 85%. The parameter estimates from a *probit* model are the following:

$$-2.5 + 0.11age - 0.0015age^2 - 0.31black + 0.1edu + 0.52married$$

- (a) What is the partial effect of education on the probability of employment for 30 years old unmarried white men with 12 grades of education? ( $\hat{\beta}'x = 0.65$ )
- (b) What is the partial effect of age on the probability of employment for 30 years old unmarried white men with 12 grades of education? ( $\hat{\beta}'x = 0.65$ )
- (c) What is the partial effect of education on the probability of employment at the sample mean? ( $\hat{\beta}'\bar{x} = 1.1$ )
- (d) What is the partial effect of education on the probability of employment at the 90<sup>th</sup> percentile of the estimated probability? ( $\hat{\beta}'x = 1.65$ )

Additional information to the next questions: the estimated probability is between 50% and

97%.

(e) Is the partial effect larger at the top of the predicted probability distribution than at the mean? Why?

(f) Is the partial effect larger at the bottom of the predicted probability distribution than at the mean? Why?

77. What is a tobit model and when is it useful?
78. Comment on the following statement: the Tobit coefficients have no meaningful interpretation in themselves.
79. The question is the effect of income on the demand for fine wine, and many households in the data buy no fine wine at all. Compare the results from an OLS (on the whole sample) and a tobit. (Assume that fine wine is a normal good.)
- Which coefficient is going to be larger?
  - Which partial effect (on expenditures) is going to be larger?
80. We want to estimate the effect of yearly income (in dollars) on yearly entertainment expenditures (in dollars) for single 30 to 60 year-old men. 20% of people reported zero expenditure on entertainment. OLS results for the whole sample are the following (appropriately estimated standard errors in parentheses):

$$\widehat{ENTERT}_i = \underset{(420)}{920} - \underset{(8)}{15}AGE_i + \underset{(0.003)}{0.016}INCOME_i$$

OLS results for the nonzero subsample are:

$$\widehat{ENTERT}_i = \underset{(470)}{1140} - \underset{(10)}{15}AGE_i + \underset{(0.003)}{0.014}INCOME_i$$

Tobit estimates for the whole sample are

$$\begin{array}{r} \underset{(460)}{710} - \underset{(9)}{20}AGE_i + \underset{(0.003)}{0.021}INCOME_i \\ \text{Error St.Dev.} = 1600 \end{array}$$

The median person is 45 years old with 30 000 income (so that  $\Phi(\hat{\beta}'_{Tobit}x_i/\hat{\sigma}) = 0.6$ ). For somebody with twice the median earnings,  $\Phi(\hat{\beta}'_{Tobit}x_i/\hat{\sigma}) = 0.75$ .

- What are the assumptions behind the Tobit model for corner solution outcomes?
  - If those assumptions are satisfied, what is the partial effect of income on expected expenditures on entertainment for the median person?
  - For someone with twice as much income?
  - Is entertainment a normal good?
  - Do people spend less on entertainment as they get older (at a 5% significance level)?
  - What is the interpretation of the Tobit coefficients?
  - What's wrong with the OLS on nonzeros model?
81. You want to see how initial firm size affects the life-span of an enterprise (i.e. how long the firm lives before it shuts down or gets bought by some other firm). You follow a sample of firms established in one year and stop the observation after 5 years, by which time about half of the firms survived. A friend suggests that you should estimate a tobit model.
- Is that a good idea, and if yes, under what assumptions?
  - Does the estimated coefficient on firm size measure the effect you are after? Or do you need some other information, and if yes, what?

82. You want to see how bonuses affect job-tenure duration (the time employees spend at the same firm). You follow a sample of people who started working at their first job in one year, but stop the observation after 5 years. By then some people left their first employer while others stayed with them. A friend suggests that you estimate a tobit model.
- Is that a good idea, and if yes, under what assumptions?
  - Does the estimated coefficient on bonuses measure the effect you are after? Or do you need to do some more work to get that?
83. The question is the effect of a job-search help program on the duration of unemployment. The sample consists of unemployed people who became unemployed at the same time. Some of them participated in the program (and thus received extra help), while the others did not. Selection into the program was completely random. Observation was stopped at a certain point in time, at which point some people found a job while others were still unemployed. Write down a tobit model for estimating the effect of the program on unemployment duration. Would the estimated tobit parameter tell you the effect of the program? Would an estimated OLS parameter on the uncensored observations tell you the effect of the program?
84. The question is the effect of expectations on the fraction of savings one invests into stock-market assets such as mutual funds, individual stocks, etc. You have a measure for the individual's subjective expectations about stock market returns, and you observe the fraction of her/his savings in stock-market based assets.
- Write down a linear model for the effect of stock market expectations on the fraction of stock-market based assets. Be simple: the only right-hand side variable should be expectations. Assume for now that the linear specification is correct. Is the OLS estimator consistent for the causal effect of expectations if expectations are higher for those who follow the stock market because they have stocks? Derive the plim and sign the bias if nonzero.
  - Suppose that you have measure on how optimistic someone is about the weather (e.g. whether the person thinks it is more likely that the next day will be sunny than its objective likelihood). Do you think this measure can be a valid instrument for stockmarket expectations? Why or why not? State the formal condition(s) and argue substantively.
  - If the fraction in savings cannot be less than 0 and more than 1, is the linear model the best one? Write down a better model.
  - Is the partial effect of expectations on the fraction of stocks in the model you outlined in (c) equal to the estimated coefficient (here assume exogeneity)?
85. The question is the effect of income on expenditures on movies. About half of the households in the sample spent zero amount on movies in the observed period.
- Write down a model for the effect of income on expenditures that takes into account the corner solution of zero expenditures.
  - Express the effect of income on expenditures in terms of the parameters of the model you described.

86. Consider a tobit model with censoring from below at 0:

$$\begin{aligned}
 y_i^* &= \beta' x_i + u_i \quad u_i | x_i \sim N(0, \sigma^2) \\
 y_i &= y_i^* \text{ if } y_i^* > 0, \text{ and } 0 \text{ if } y_i^* \leq 0
 \end{aligned}$$

Derive the conditional expectations of  $E(y_i | x_i, y_i > 0)$  and  $E(y_i | x_i)$ .

87. You want to see how married women's labor supply is affected by family income that comes from other sources (*NWIFEINC*), measured in thousand dollars per year. You first estimate

a probit model, where the LHS variable is 1 if the woman works, and 0 otherwise (*WORKS*). You also control for education (years completed). Your sample consists of  $n$  women. The probit estimates are:

$$-1.13 + 0.15 EDUC_i - 0.02 NWINC_i$$

(0.26)      (0.02)      (0.004)

Let  $\phi_i$  denote the standard normal density of the estimated index  $(\hat{\beta}'x)$  for the  $i$ 'th woman in the sample.

- (a) What is the estimated average partial effect of having \$10,000 more other income on the probability of working if  $\frac{1}{n} \sum \phi_i = 0.3$ ?
- (b) What is the estimated partial effect of having \$10,000 more other income on working for the most likely working 10% of women (for whom  $\phi_i = 0.4$ )?
- (c) What is the estimated partial effect of having \$10,000 more other income on working for the least likely working 10% of women (for whom  $\phi_i = 0.2$ )?
- (d) Based on the probit estimates, if working women work 25 hours a week on average, what is the estimated increase in total weekly hours supplied by women if the government gives an extra \$1,000 to 10 million families? (Hint: you have to assume something about the hours lost because of some people opting out from the labor force. Average hours are one candidate for that. You may comment on the bias you expect from using average hours.)

After all that you realize that you observe hours worked as well. You therefore estimate a tobit model with the same RHS variables, and you get a  $-0.5$  coefficient estimate on *NWIFEINC*. The average of the standard normal distribution of the estimated tobit index is  $\frac{1}{n} \sum \Phi\left(\frac{\hat{\beta}'x_i}{\hat{\sigma}}\right) = 0.8$

- (e) What is the estimated average partial effect of having \$10,000 more other income on hours per week?
  - (f) Based on the tobit estimates, what is the estimated increase in total weekly hours supplied by women if the government gives an extra \$1,000 to 10 million families?
  - (g) Which answer would you prefer, (d) or (f), and why?
88. Is it true that sample selection causes no inconsistency if selection is random? If selection is a deterministic function of right-hand side variables? Why or why not?