

Econometrics 1. Sample Questions

Fall 2007

1. Suppose you have estimated a linear regression on an *iid* sample by OLS, and the White test rejects its null hypothesis. What are the properties of the point estimates and the t-tests on them if
 - (a) you estimated White standard errors?
 - (b) you used the standard errors regression programs (incl. EViews) give by default?
 - (c) Do you need to do something in a different way in order to get efficient point estimates? If yes, what? If not, why not?
2. The White test for conditional heteroskedasticity can be represented as a regression of the squared OLS residuals on what variables? Formulated in terms of that regression, what is the null hypothesis of the White test?
3. Is the White standard error estimator consistent when the error term of a linear regression estimated on an *iid* sample is homoskedastic?
4. Does the confidence interval of a simple regression parameter (OLS estimator) depend on the variance of the right-hand side variable (suppose that all classical assumptions hold)?
5. If you run a regression on variables of a not too large *iid* sample, the error term is homoskedastic and the RHS variables are exogenous, what may be the consequence (if any) of two strongly correlated RHS variables? Do you have to do something about that, and if yes, what?
6. Comment on the following statement: multicollinearity is analogous to having too small a sample.
7. Consider the following regression on a cross-sectional sample of countries:

$$growth_i = \beta_0 + \beta_1 budget_i + u_i,$$

where *growth* is economic growth and *budget* is the size of government budget relative to GDP. Suppose that we would like to test whether there is a negative effect of the budget on growth, *ceteris paribus*. Suppose that corruption affects the size of the budget in a negative way, but it also affects economic growth (conditional on the budget), negatively. Is the OLS estimator for β_1 consistent? If not, what do you think the direction of the (asymptotic) bias will be, and why?

8. Suppose that you estimated a simple regression on an *iid* sample, and you got the following results:

$$\hat{y}_i = 2 + x_i.$$

If in the sample, $V(y_i) = 4V(x_i)$,

- (a) what is the (sample) correlation between the two variables?
- (b) What is the R^2 of the regression?

9. Consider the following regression on an *iid* sample:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + \beta_4 zx + u$$

Assuming exogenous right-hand-side variables, derive the partial effect x on the expected value of y .

10. You have estimated the following regression (appropriately estimated standard errors in parentheses):

$$\hat{y} = \underset{(0.1)}{0.5} + \underset{(0.2)}{0.8}x$$

- (a) If x is exogenous, what is the partial (causal) effect of x on the expected value of y based on the point estimate?
 (b) What if x is endogenous so that $\text{Corr}(x, u) > 0$? Would OLS estimate the partial effect in a consistent way? If not, what will be the direction of the (asymptotic) bias?
 (c) If $\text{Corr}(x, u) > 0$, will the OLS estimator for the intercept be consistent? If not, what will be the direction of the (asymptotic) bias?
 (d) What if x is exogenous but u is heteroskedastic so that $\text{Corr}(u^2, x) > 0$? Would OLS estimate the partial effect in a consistent way? If not, what will be the direction of the (asymptotic) bias?
11. Consider the following regression on an *iid* sample:

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

Suppose that there is some unmeasured variable w that affects y but is uncorrelated of x . What will be the properties of the OLS estimators x will be exogenous and therefore the OLS estimator for β_1 will be biased and inconsistent.

- (a) Would OLS estimate the partial effect of x on y in a consistent way? If not, what will be the direction of the (asymptotic) bias?
 (b) Would OLS estimate the expected value of y given $x = 0$ in a consistent way? If not, what will be the direction of the (asymptotic) bias?
12. You have estimated the following regression (appropriately estimated standard errors in parentheses):

$$\hat{y} = \underset{(0.1)}{0.5} + \underset{(0.2)}{0.8}x - \underset{(0.01)}{0.05}x^2$$

- (a) If x is exogenous, what is the partial (causal) effect of x on the expected value of y based on the point estimates?
 (b) Is the partial effect positive for all x ? For some x values?
 (c) Is the partial effect negative for all x ? For some x values?
 (d) Is the partial effect zero for all x ? For some x values?
13. If one were to predict y for different values of x from a simple regression model (where all classical assumptions hold), would the prediction error be the same regardless of the value of x ?

14. The following demand equation was estimated on a cross-section of one good sold on different markets ($n = 100$). The estimated parameters are in the equation, and their appropriately estimated standard error below in parentheses.

$$\log(Q_i) = \underset{(0.20)}{1.5} - \underset{(0.10)}{0.75} \log(P_i) + \underset{(0.20)}{0.6} \log(Y_i) + u_i,$$

where Q_i is quantity sold, P_i is price, and Y_i is disposable income (both in real terms) of the consumers. Assume that all variables (their logs) are stationary, and variation in price and income were exogenous to demand.

- (a) Test whether the good is inferior.
- (b) Test whether the demand is price-elastic.
- (c) Would your answers be different if prices and income were nonstationary?

15. Consider the following equation:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 m_i + \beta_3 f_i + u_i,$$

where y is earnings, x is education, m is the mother's education, and f is the father's education (all measured in grades completed). Explain a simple t -test procedure (that does not require more information than what's produced by default by EViews), by which you can test the hypothesis that the mother's and the father's education have the same effect on earnings.

16. Consider the following regressions estimated on a sample of size 52. Suppose that all the classical assumptions hold (including non-stochastic RHS variables and a normally and independently distributed homoskedastic error term).

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i, & R^2 &= 0.31 \\ y_i &= \gamma_0 + \gamma_1 x_i + v_i, & R^2 &= 0.26 \end{aligned}$$

Can you use the above information to test whether $\beta_2 = 0$?

17. The following production function was estimated on a cross-sectional sample of firms:

$$\log(Y_i) = \beta_0 + \beta_1 \log(L_i) + \beta_2 \log(K_i) + u_i,$$

where Y_i is output, L_i is labor input, and K_i is capital input. Assume that all classical assumptions hold. Explain two different methods for testing whether there are constant returns to scale. State the adequate null hypotheses and the testing procedures step by step.

18. 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 β .
Suppose that you find a variable z such that

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

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

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

- (a) Derive the probability limit (*p* lim) of $\hat{\beta}_{IV}$. Is it consistent for β ?
 (b) What if $\text{Cov}(x_i, u_i) > \text{Cov}(z_i, u_i) > 0$? Is the IV estimator consistent?
 (c) If not, can you tell whether the IV or the OLS estimator is more biased (asymptotically)?
19. The following regression was estimated on an *iid* sample of employees who are all 20 to 60 years old (standard errors in parentheses):

$$\log(w_i) = \underset{(0.5)}{5} - \underset{(0.02)}{0.10} \text{male}_i + \underset{(0.02)}{0.07} \text{edu}_i + \underset{(0.005)}{0.02} \text{male}_i \times \text{edu}_i + u_i,$$

where w is wage, $\text{male} = 1$ if male and 0 if female, and edu is the highest educational attainment in years completed. IN the relevant population people have between 8 and 16 years of education.

Answer the following questions:

- a) What is the meaning of the *male* coefficient?
 b) What is the meaning of the *edu* coefficient?
 c) What is the meaning of the *male* \times *edu* coefficient?
 d) According to the point estimates, what is the (approximate) male-female earnings difference between people with 12 years of education?
 e) Can you test the hypothesis that men with 16 years of education earn at least 20% more than women with 16 years of education? If no, what additional information would you need? If yes, can you reject it?
 f) Can you test the hypothesis that men with 12 years of education earn the same as women with 16 years of education? If no, what additional information would you need? If yes, can you reject it?
20. Suppose that you can reject a hypothesis at 5% significance level and at 1% significance level as well. Which one would you report (emphasize) in your study?

21. Consider the following regression results on large a cross-sectional sample of employees:

$$\log(\text{wage}_i) = \underset{(0.03)}{9.6} + \underset{(0.02)}{0.25}\text{woman} + \underset{(0.01)}{0.15}\text{edu} - \underset{(0.01)}{0.03}\text{woman} \times \text{edu} + \underset{(0.01)}{0.016}\text{exper} - \underset{(0.0001)}{0.0002}\text{exper}^2$$

where *woman* is a dummy, *edu* is completed education in years, and *exper* is labor market experience in years.

Answer the following questions based on the point estimates.

- What is the meaning of the intercept?
- What is the meaning of the coefficient on education?
- What is the gender difference in earnings among people with 12 years of education and 0 labor market experience?
- Does the gender difference increase, decrease or not change with education?
- Does the gender difference increase, decrease or not change with experience?
- What is the return to education for men (possibly a function of other variables)?
- What is the return to education for women (possibly a function of other variables)?
- What is the return to experience for men (possibly a function of other variables)?
- What is the wage difference between a man with 10 and one with 20 years of experience?

More general questions:

- Why do we have log earnings on the LHS?
- If there is ability bias, do you think the real returns to education are smaller, larger, or the same as the estimates? Why?
- Does the estimated gender difference measure labor market discrimination against women?
- If, like in a transition economy, the labor market experience of older generations devaluated, do you think the real returns to experience (that young employees could expect for themselves in the future) are smaller, larger, or the same as the estimates? Why?

22. We would like to measure the effect of a training program for the long-term unemployed. The effect is defined as the change in the employment probability of a participant half a year after completion of the program relative to what that probability would be if she did not participate. The data at hand is a cross-sectional sample of participants and non-participants, and participation is granted if the applicant passes an entry exam test, but we do not know the result of that test. Assume that the unmeasured qualities that affect exam scores are very similar to those that increase marginal product if employed. If Y_i is employment half a year after the program ($Y = 1$ if employed, 0 if not), and D_i denotes participation ($D = 1$ if participated, 0 if not)
- Would a regression $Y_i = \alpha + \beta D_i + u_i$ consistently estimate the effect of the program or would the estimate be upward or downward biased (asymptotically)?
 - What if you included other observed personal characteristics in the regression?
 - What if participation is randomly assigned like in a medical experiment?
 - Comment on the following statement: the $Y_i = \alpha + \beta D_i + u_i$ model is of limited value even if D is exogenous, because the left-hand side variable is binary and therefore the estimated probability can be negative.

23. Consider the following regression results on large a cross-sectional sample of employees in Hungary, all 25 to 60 years old and with college education:

$$\log(w_i) = \underset{(0.01)}{12.0} - \underset{(0.01)}{0.15}f - \underset{(0.01)}{0.30}g - \underset{(0.01)}{0.03}f \times g + \underset{(0.001)}{0.02}(a - 25) - \underset{(0.00005)}{0.0002}(a - 25)^2$$

$f = 1$ if female, 0 if male
 $g = 1$ if works in the government sector (public admin, education, health), 0 otherwise
 $f \times g =$ interaction between the two dummies
 $a =$ age in years

Answer the following questions based on the estimates of the model (and the population the sample represents).

- What is the meaning of the intercept?
 - What is the meaning of the coefficient on g ?
 - What is the meaning of the coefficient on $f \times g$?
 - Is the average gender difference different in the government and the non-government sector? Is that difference less than 5%?
 - Does the gender difference increase, decrease or not change with age?
 - Do the results provide evidence for gender discrimination in the government sector?
 - Do older women earn more than younger ones?
 - Based on the results, can you test whether age differences are different in the government sector than elsewhere? If yes, state the null and alternative hypotheses and carry out the test. If not, would other estimates of the above model help? Or would you need to estimate a modified model, and if yes what model?
24. Consider the following null hypothesis:

$$H_0 : \beta_1 = 0 \quad \text{and} \quad \beta_2 = 0$$

How does the alternative hypothesis look like?

25. We would like to see if there is discrimination against African Americans on the U.S. mortgage loan market. Our sample consists of loan applications. We estimate the following model

$$approve_i = \beta_0 + \beta_1 black_i + u_i$$

where $approve_i = 1$ if the loan application of person i was approved and 0 otherwise; $black_i = 1$ if the applicant is African American, and 0 otherwise.

- What is the expected sign of β_1 if there is discrimination against African Americans?
- What is the meaning of β_0 (if any)?
- Linear probability models are often problematic because the predicted probability may go below zero or above one. Can that happen here?
- How would you estimate the standard errors?
- Is OLS efficient? Do you know about a more efficient estimator? If yes, how does that look like?
- Suppose that African Americans have lower income (and lower expected income in the future) and approval is more likely for higher-income people, regardless of race. Is the OLS estimate of β_1 consistent for the degree of discrimination? If not, is the bias positive?

26. 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). 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?)
27. Comment on the following statement: If the error term is heteroskedastic, the simple standard errors (the ones EViews gives by default) are inconsistent because the OLS estimator is not asymptotically normal.
28. Consider the following regression results on a large cross-sectional sample of college students in the U.S. (White standard errors in parentheses):

$$\log(\widehat{gpa}_i) = 1.00 + 0.05 \textit{female}_i - 0.12 \textit{athlete}_i + 0.06 \textit{female}_i \times \textit{athlete}_i$$

(0.01) (0.01) (0.02) (0.05)

Answer the following questions based on the estimates of the model (if there is not enough information, write down what else you would need):

- (a) What is the meaning of the intercept?
- (b) What is the meaning of the coefficient on *female*?
- (c) What is the meaning of the coefficient on *female* × *athlete*?
- (d) What is the estimated difference between the GPA of non-athlete male and non-athlete female students? Is that difference significant?
- (e) What is the estimated difference between the GPA of athlete male and athlete female students? Is that difference significant?
- (f) The White test produces a p-value of 0.11. Does that change your answer to parts (d) and (e)?

A re-estimated model includes the student's percentile ranking in his/her high school as additional right-hand side variable, in a quadratic form (*hsperc*; percentile ranking goes from 1 to 100; lower percentile ranking means better achievement in high school). The results are the following:

$$\log(\widehat{gpa}_i) = 1.16 + 0.02 \textit{female}_i - 0.04 \textit{athlete}_i + 0.04 \textit{female}_i \times \textit{athlete}_i - 0.015 \textit{hsperc}_i + 0.0001 \textit{hsperc}_i^2$$

(0.01) (0.01) (0.02) (0.04) (0.001) (0.00001)

- (g) Why did the coefficient on *athlete* become smaller in absolute value?
- (h) What is the effect of the student's percentile ranking in high school on his/her college GPA? Is the effect positive, negative, increasing, decreasing or what?
- (i) Can you test whether achievement in high school matters more for men than for women? If yes, how? If no, what kind of a modified regression would you need?
29. Consider two tests for the same null and alternative hypotheses. What does it mean for one test to be more powerful than the other?

30. Our question is whether airline companies use their market power to charge higher prices in the U.S. The data consists of average prices on the most popular routes (e.g. Boston-Chicago) for year 2000. OLS estimates of our regression are the following (White standard errors in parentheses):

$$\widehat{lprice}_i = \underset{(0.10)}{4.4} + \underset{(0.02)}{0.4} d_i - \underset{(0.01)}{0.06} p_i + \underset{(0.3)}{0.8} mkts_i - \underset{(0.2)}{0.4} mkts_i^2$$

where $lprice$ is log of the average price on the given route, d is distance of the route (in thousand miles, range is 0.1 to 3), p is number of average passengers per day (in thousands, range is 0.01 to 8), and $mkts$ is the market share of the biggest airline carrier on the given route (range is 0.1 to 1).

- (a) Based on the point estimates and assuming exogeneity, what is the partial effect of the market share of the largest carrier on prices?
- (b) Is your answer to (a) consistent with the hypothesis that firms use their market power to charge higher prices?
- (c) How would you test whether market power is used the same way on more popular and less popular routes? (Write down the model and the hypotheses, and describe the test procedure.)
- (d) What is the meaning of the coefficient on p ?
- (e) Are more longer routes as expensive as shorter ones? (Carry out the test.) Is the relationship linear? What would be the economic interpretation of a linear relationship?
- (f) Why did we estimate White standard errors? Was that the right thing to do?
31. A firm offered a training program to its sales employees last year, and now the management wants to know how effective the program was. You are a consultant, and you have some data on participants and non-participants of the program. All the data are from half a year after the program. sal_i is monthly sales for sales-person i , and D_i is the participation dummy (1 if i participated and 0 otherwise). You also know individual i 's education (ed_i), tenure with the firm i.e. number of years she/he has spent there (ten_i), and her/his overall labor market experience (ex_i). You specify the following regression

$$\log(sal_i) = \beta_0 + \beta_1 D_i + \beta_2 ed_i + \beta_3 ten_i + \beta_4 ex_i + u_i$$

- (a) What is the interpretation of β_1 (without assuming exogeneity)?
- (b) Does OLS consistently estimate the effect of the program under random assignment (i.e. if randomly chosen sales employees participated in the program)? If not, what is the direction of the (asymptotic) bias?
- (c) How would you answer (b) if you didn't control for education, tenure and experience?
- (d) Do you expect $\beta_2 = \beta_3 = \beta_4 = 0$ under random assignment?
- (e) Does OLS consistently estimate the effect of the program if, instead of random assignment, the firm made less able sales persons participate in the program? If not, what is the direction of the (asymptotic) bias?
- (f) It is possible that monthly sales are in part due to factors that the employee cannot influence. State the assumptions under which OLS would still yield consistent estimates for the effect of the program.
- (g) Are the assumptions in (f) likely to be satisfied under random assignment? Under the assignment in part (e)?
32. What does it mean exactly that an estimated coefficient $\hat{\beta}$ is "significant at 5%"?