

EC421: Advanced Econometric Methods - Spring 2021; Professor Vogelsang
Test 1, Monday March 8, 2021.

This test has two parts totaling 100 points. This is a take home test. You must submit your answers to the D2L dropbox by 11:59PM Eastern Time on Monday March 8, 2021. The test is OPEN BOOK, OPEN NOTES. Useful formulas and tables of distributions can be found on the course formula sheet and critical values documents posted to D2L. You MAY NOT share the test questions with any other individuals. You MUST work on the exam by yourself. The test is open book/open notes. The internet is NOT to be used. Calculators are permitted. Either print the answer sheet posted on D2L or use your own blank paper to write up your answers. You may handwrite your answers, type them, or both. Show all of your work as partial credit will be given. Please scan your answer sheets and submit using the D2L dropbox using **pdf** format if possible.

Part I: Short Answer (25 points, 5 points each)

In **WORDS**, briefly describe and discuss the following. Try to limit your answer to no more than three or four sentences. You may use formulas if they are helpful, but a formula without an explanation will receive no credit!

a) What are the implications of heteroskedasticity for OLS (assuming MLR.1, MLR.2, MLR.3 and MLR.4 hold)?

b) What are the implications for generalized least squares in a transformed regression that uses the WRONG model for h_i (assuming MLR.1, MLR.2, MLR.3 and MLR.4 hold). For example, the true model has $h_i = educ_i^2$ but the transformed regression is based on $h_i = educ_i$ where $educ_i$ is years of education.

c) Can the Gauss-Markov Theorem be used to rank the variances of the OLS estimator and the IV estimator when MLR.4 holds (assuming MLR.1, MLR.2, MLR.3 and MLR.5 also hold)? Explain why or why not.

d) Can the Gauss-Markov Theorem be used to rank the variances of the OLS estimator and the IV estimator when MLR.4 fails? Explain why or why not.

e) Are there conditions under which a variable, z_1 , can serve as both a valid instrument for an included regressor, x_1 , AND a valid proxy for an omitted variable, x_2 , that is correlated with the included regressor? In math notation, in the regression $y_i = \beta_0 + \beta_1 x_{i1} + u_i$ where u_i contains an omitted variable x_{i2} that is correlated with x_{i1} , are there conditions under which z_{i1} can be a valid instrument for x_{i1} AND a valid proxy for the omitted variable, x_{i2} ? Explain why or why not.

Part II: 75 points

This part of the test asks a series of questions about the relationship between years of education, $educ$, and the natural log of hourly wages, $\log(wage)$. Additional variables available to the empirical researcher are years of experience on the job, $exper$, and years of education of the person's mother, $motheduc$, and years of education of the person's father, $fatheduc$.

A random sample of 428 individuals was used to estimate three regression models of increasing complexity:

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + u_i \quad (1)$$

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + u_i \quad (2)$$

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + u_i \quad (3)$$

OLS estimates of the three regressions are as follows. Numbers in parentheses () are OLS standard errors. Numbers in brackets [] are heteroskedasticity robust standard errors.

$\begin{aligned} \widehat{\log(wage_i)} &= -.185 + .1086educ_i \\ &\quad (.185) \quad (.014) \\ &\quad [.171] \quad [.013] \end{aligned}$	$R^2 = .118$ $SSR = 197.0$
$\begin{aligned} \widehat{\log(wage_i)} &= -.400 + .1095educ_i + .0157exper_i \\ &\quad (.190) \quad (.014) \quad (.004) \\ &\quad [.183] \quad [.013] \quad [.004] \end{aligned}$	$R^2 = .148$ $SSR = 190.2$
$\begin{aligned} \widehat{\log(wage_i)} &= -.522 + .1075educ_i + .0416exper_i - .0008exper_i^2 \\ &\quad (.199) \quad (.014) \quad (.013) \quad (.0004) \\ &\quad [.202] \quad [.013] \quad [.015] \quad [.0004] \end{aligned}$	$R^2 = .157$ $SSR = 188.3$

- a) Notice that as the model becomes less complex (going from (3) to (2) to (1)), the SSR increases and the R^2 decreases. Provide a theoretical explanation for these patterns. (4 points)

- b) Using the estimated model (3), what is the impact, on average, of one additional year of education on wages for a person with 6 years of education ($educ = 6$)? What is the impact for a person with 12 years of education? What is the impact for a person with 16 years of education? In answering these questions, clearly state what is being held constant (you may assume MLR.4 holds). (3 points)

- c) Using the estimated regression model (3), what is the impact, on average, of one additional year of experience on wages for a person with 1 year of experience ($exper = 1$)? [**Hint:** modify the log-level result on page 3 of the formula sheet for the case where an x^2 term is in the model: $\log(y) = \beta_0 + \beta_1 x + \beta_2 x^2$.] What is the impact for a person with 10 years of experience ($exper = 10$)? What is the impact for a person with 20 years of experience ($exper = 20$)? What is happening to the impact of an additional year of experience on wages as experience increases? Does this make sense economically? Why or why not? (5 points)
- d) Assuming MLR.5 holds, test the hypothesis that an additional year of education is associated with a 5% increase in wages. Use regression model (3). Clearly state your null and alternative hypotheses in terms of β_1 . Clearly state your rejection rule. Carry out your test at the 5% significance level. (5 points)
- e) In a meeting to determine tuition increases, a college university president says that education is the 'only' factor that matters for wages and claims that years of experience on the job has no effect on wages. Test this hypothesis using regression model (3). Clearly state your null and alternative hypotheses in terms of the slope parameters. Clearly state your rejection rule. Carry out your test at the 5% significance level. Continue to assume MLR.5 holds. (5 points)

To assess whether regression model (3) satisfies MLR.5, both the Breusch-Pagan and White tests for heteroskedasticity were computed for regression (3). The F -statistic version of the **Breusch-Pagan** test yielded **3.98** and the F -statistic version of the **White** test yielded **1.79**.

- f) Let \hat{u}_i be the OLS residuals from regression model (3). Write down the regression model that was used to compute the Breusch-Pagan test F -statistic. Carry out the Breusch-Pagan test at the 5% significance level. (4 points)
- g) Write down the regression model that was used to compute the White test F -statistic. Carry out the White test at the 5% significance level. (4 points)
- h) Based on the heteroskedasticity tests, is the use of OLS standard errors in parts (d) and (e) appropriate? Why or why not? Does your answer to part (d) change in any substantial way if you use the robust standard error? Explain. (4 points)

Suppose we think that in regression model (3) the error has heteroskedasticity of the form

$$\text{var}(u_i | \text{educ}_i, \text{exper}_i, \text{exper}_i^2) = \sigma^2 h_i,$$

where

$$h_i = .885 - .055 \text{exper}_i + .0012 \text{exper}_i^2.$$

- i) Under the assumption this model of h_i is correct, what is the transformed version of regression model (3) that removes the heteroskedasticity and can be used to obtain the generalized least squares (GLS) estimators of the β parameters? What would you need to check about all the values of h_i before trying to estimate this transformed regression? (4 points)

Using h_i from part (i), OLS estimation of the transformed model (GLS) resulted in the fitted model:

$$\begin{array}{cccccc} \widehat{\log(wage_i)} & = & -.533 & +.112educ_i & +.0357exper_i & -.0006exper_i^2 & R^2 = .156 \\ & & (.199) & (.014) & (.013) & (.0004) & SSR = 188.49 \\ & & [.194] & [.013] & [.014] & [.0004] \end{array}$$

The Breusch-Pagan statistic for testing the null hypothesis of no heteroskedasticity in the transformed model resulted in a p -value of 0.9505.

- j) Is there evidence that the assumed form of h_i is correct and that the GLS transformation has removed heteroskedasticity from regression model (3)? Explain. (3 points)
- k) If you were doing this empirical analysis as part of your job, which estimated parameters would you report to your supervisor, OLS, GLS or both? Why? Carefully outline the logic and justification behind your decision. (4 points)

One concern with regression model (3) is the potential for MLR.4 to fail because of omitted variables such as *ability*. Suppose we plan to use one or both of the parental education variables, *motheduc* and *fatheduc* to solve the MLR.4 problem. The remaining questions explore how these variables could be used to solve the MLR.4 problem.

- l) To be specific, suppose that the error in regression model (3) depends on *ability* and *ability* is correlated with *educ*. What would this imply about the statistical properties of the OLS estimators of regression model (3)? Why? (3 points)
- m) An empirical researcher suggests that *motheduc* and *fatheduc* be used as proxies for *ability*. For *motheduc* and *fatheduc* to be good proxies for *ability*, how must they be related to *educ* and *ability*? (4 points)

The researcher estimated the following regression. Only robust standard errors are reported.

$$\begin{array}{ccccccccc} \widehat{\log(wage_i)} & = & -.470 & +.120educ_i & +.0412exper_i & -.0008exper_i^2 & -.016motheduc_i & -.005fatheduc_i & (4) \\ & & [.194] & [.013] & [.014] & [.0004] & [.012] & [.011] \\ & & R^2 = .163 & SSR = 186.90 \end{array}$$

- n) Is there any statistical evidence that wages of a person are related to *motheduc* and *fatheduc* given we have controlled for *educ*, *exper*, and *exper*²? Be precise in your answer. Is the estimated slope on *educ* substantially different from what OLS estimation of regression model (3) gave? Are you surprised? Why or why not? (5 points)
- o) Another empirical researcher suggests that *motheduc* and *fatheduc* be used as instruments for *educ*. For *motheduc* and *fatheduc* to be valid instruments, how must they be related to *educ* and to *ability*? (4 points)

This researcher estimates regression model (3) using 2SLS. In the first stage, the researcher uses *exper*, *exper*², *motheduc* and *fatheduc* to construct a proxy for *educ*:

$$\widehat{educ}_i = 8.367 + .0853exper_i - .0019exper_i^2 + .186motheduc_i + .185fatheduc_i \quad R^2 = .262$$

[.280] [.026] [.0009] [.026] [.024]

In the second stage, the researcher uses \widehat{educ}_i as a proxy for *educ*_{*i*} and obtains the IV estimated model

$$\widehat{\log(wage_i)} = .0481 + .061\widehat{educ}_i + .0442exper_i - .0009exper_i^2 \quad R^2 = .146$$

[.427] [.033] [.015] [.0004]

(5)

Only robust standard errors are reported in both cases.

- p) Notice that the IV estimated slope of β_1 (coefficient on \widehat{educ}) is substantially smaller than what OLS estimation of regression model (3) gave. Provide an explanation for this. (4 points)
- q) Is there any indication that *motheduc* and *fatheduc* are weak instruments? Be very precise in your answer and back up your conclusions with empirical evidence. (6 points)
- r) If you were doing this empirical analysis as part of your job, which estimated parameters would you report to your supervisor, regression model (4), the IV estimated model (5), or both? Justify your answer with evidence and logical arguments. (4 points)