Stockholm University

Department of Economics

Course name:	**Empirical Methods in Economics 2**
Course code:	**EC2404**
Type of exam:	**Retake**
Examiner:	**Daniel Knutsson**
Number of credits:	**7,5 credits**
Date of exam:	**24-04-2018**
Examination time:	**3 hours (09:00-12:00)**

Write your exam identification number on each answer sheet (the number stated in the upper right hand corner on your exam cover).

Explain notions/concepts and symbols. If you think that a question is vaguely formulated, specify the conditions used for solving it. Only legible exams will be marked. **No aids are allowed.**

The exam consists of 5 questions. The first two contain multiple choice questions which are to be answered on separate sheets. Each question is worth 20 points, 100 points in total. Note that Question 5 is the credit question.

For the grade E 45 points are required, for D 50 points, C 60 points, B 75 points and A 90 points.

Your results will be made available on your Ladok account **(**www.student.ladok.se) within 15 working days from the date of the examination.

Good luck!

Multiple choice 1. (20 p)

Please tick (kryssa för) the correct answer. Only one answer is correct.

1. The IV estimate is (almost) always interpreted as a Local Average Treatment Effect (LATE). This means that the causal effect is defined for:

   A. Everyone in the sample
   B. Always-takers, those that always take treatment
   C. Never-takers, those that never take treatment
   D. Compliers, those that take treatment only when the instrument encourage them to do so

2. Using fixed effects in a panel data setting means that,

   A. You can control for all omitted variables bias that is constant within groups
   B. You can control for all random measurement error that vary over time within groups
   C. You can solve problems with reversed causality within groups
   D. You can control for all omitted variables bias that vary over time within groups

3. Suppose I randomly assign treatment (X) within a population and run a bivariate regression using X on an outcome Y. This implies that the coefficient on X is

   A. Biased as there are no control variables to support the conditional mean assumption
   B. The causal effect of X on Y
   C. Biased, as randomization did not take place within subgroups
   D. Biased, as Y was not randomly assigned

4. In an IV setting, I can show that my instrument is unrelated to pre-determined variables using regression models. Which argument can these results strengthen?

   A. If the predetermined variables are relevant
   B. If the instrument is relevant
   C. If the instrument is exogenous
   D. If the predetermined variable should be included in the first stage

5. You estimate a bivariate regression model: $Y_i = \beta_0 + \beta_1 X_i + u_i$. However, you do not include the variable W_i as a control variable. W_i is positively related to both X_i and Y_i. Given this information, the omitted variables bias in $\beta_1$due to omitting W_i is:

   A. It does not create bias.
   B. Negative.
   C. Positive.
   D. Insufficient information to assess from the information above.

Multiple choice 2. (20 p)

Please tick (kryssa för) the correct answer. Only one answer is correct.

1. You are interested in estimating a population relationship, $Y_i = \beta_0 + \beta_1 X_i + u_i$, where X_i is randomly assigned. However, you are not sure about the functional form. You include X^2 in the regression model.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

You find that $\widehat{\beta_1} = 3$ and $\widehat{\beta_2} = -1$ when you estimate the regression, both highly significant. What is the estimated effect of going from 1 to 2 on the X scale? (Hint: plug in the values once you have calculated the change in coefficients)

A. 4
B. 2
C. -2
D. 0

2. In the same setting as above, you try another functional form using the natural logarithm of X: $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$. You find that the estimated $\widehat{\beta_1} = 0.15$. How do you interpret this coefficient?

A. A unit change in X increase Y by 15 percent
B. A one hundred percent change in X increase Y by 0.15 units
C. A unit change in X increase Y by 0.15 percent
D. A one hundred percent change in X increase Y by 15 percent

3. Which of the following statements is **false** when it comes to heteroscedasticity?

A. Heteroscedasticity tend to make standard errors larger than if the error was homoscedastic.
B. Heteroscedasticity should (almost) always be accounted for in regressions using heteroscedasticity robust standard errors.
C. Heteroscedasticity means that the error term exhibits serial correlation.
D. Heteroscedasticity means that the conditional variance of the error term is not constant.

4. Which of the following statements is **correct** regarding measurement errors?

A. Random measurement error in the outcome variable leads to inconsistent and biased estimates.
B. Random measurement error in the treatment variable leads to attenuation bias.
C. Random measurement error in the treatment variable only increase standard errors.
D. Random measurement error is not a problem in any variable as long as it is strictly random.

5. Consider the multiple regression model with two regressors X_1 and X_2, both are correlated with the dependent variable. $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. When will omitting X_2 from the regression cause omitted variables bias in β_1? (4 p)

A. If X_1 and X_2 are correlated
B. Always
C. If X_2 is correlated with the error term, u_i
D. If X_2 is a dummy variable

Panel Data (20 points)

You are now interested in if school resources (for example smaller classes) are important for educational outcomes. You have access to data on school resources and educational outcomes from all municipalities in Sweden for the years 1990, 1993, 1996 and 1999.

1. Write down the FE regression model with time effects and the Pooled OLS regression model. Without further information, which model do you prefer and why? Motivate your answer. (5 p)

2. In 1998 around half of municipalities were forced to increase resources due to a new law. Write down the difference-in-differences (DiD) model and interpret the main coefficient of interest. (5 p)

3. You estimate a more flexible difference-in-difference model as following:

$$Education_{it} = \beta_0 + \beta_1 T_i + \gamma_1 (T_i * Yr_{1990}) + \gamma_2 (T_i * Yr_{1993}) + \gamma_3 (T_i * Yr_{1999}) + \lambda_t + \alpha_i + u_{it}$$

Where the subscripts (i) index each municipality and (t) index each year. T_i is a binary treatment variable indicating if a municipality received more resources due to the new law (was in the treatment group). $(T_i * Yr_t)$ are treatment interaction variables for each year (omitting one).

Describe which coefficient estimate that give you the treatment effect. Which is the most important assumption for a causal interpretation in this setting? How can you, based on information in the estimated model, argue that this assumption holds and that you have estimated a causal effect? Draw a figure and show an example where this assumption does **not** hold. (10 p)

Instrumental Variables (20 p)

Suppose you are interested in how protests (rallies or demonstrations) affect policy (what politicians actually decide). You find out that a few years ago, right wing activists had organized demonstrations all over the country on the same day to demand more right wing policy. Specifically, you want to know if protest size (how many people that actually showed up), at that specific day but at different locations, affected how local politicians later voted. The equation of interest is hence:

$$Right\ Wing\ Policy_m = \beta_0 + \beta_1 Protest\ size_m + u_m$$

The *m* subscript refer to municipality. "Right Wing Policy" is the voting behavior of local politicians and is measured after the demonstrations. "Protest size", is how many people that protested in each municipality on that day. You have an instrument for protest size, if it rained on this specific protest day, $Z_m = Rainfall_m$.

1. Think about instrument relevance and exogeneity and discuss why rainfall could be a valid instrument in this specific context. Which of these assumptions is it possible to formally test and in which equation can you do that? (5 p)

2. Write down the first stage equation, the reduced form and show how the IV estimate can be described by the coefficients from these equations. (5 p)

3. You are told that there is a problem with your instrument. In places where there is less often rainfall, people usually demonstrate more and politicians have more right wing attitudes. You can control for past rainfall (the probability of rainfall) which is the omitted variable. Given this information, explicitly state the conditional mean independence assumption, under which the instrument is conditionally exogenous. (5 p)

4. Your reduced form estimate of *Rainfall* on *Right Wing Policy* one year after the demonstrations is negative and significant. You also test a specification using Right Wing Policy one year **before** the demonstrations as the outcome and find a small and insignificant estimate. What does these two findings suggest? Motivate your answer. (5 p)

Credit question (20 p)

(This question is only for those of you that have not handed in all assignments)

Consider the two IV papers that we have used in class:

Angrist J & Evans W (1998) "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *The American Economic Review*.

Card D (1995) "Using geographic variation in college proximity to estimate the returns to education", NBER working paper no. 4483.

1. Describe each papers identification strategy (how they attempt to estimate a causal effect). Write out the important equations, interpret the parameters of interest and explicitly state under which assumptions the IV strategy is valid. (5 p)

2. Discuss the credibility of each paper estimating a causal effect (discuss exogeneity). What are the strengths/weaknesses of each instrument? Motivate your answer. (10 p)

3. Suppose you only have only one instrument, how could you strengthen your claim that the instrument is exogenous using regression models? (Remember that there is no formal test) (5 p)

Try not to write more than 1.5 pages in total.