Department of Economics

Course name:	**Empirical Methods in Economics 2**
Course code:	**EC2404**
Semester:	**Autumn 2016**
Type of exam:	**Retake**
Examiner:	**Peter Skogman Thoursie**
Number of credits:	**7,5 credits**
Date of exam:	**Sunday 11 December 2016**
Examination time:	**3 hours (9:00-12:00)**

Write your identification number on each answer sheet. Only use printed answer sheets for your answers: Multiple-choice answer sheets for the multiple-choice questions and general answer sheets for all other questions. Do not answer more than one question on each answer sheet.

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, worth 4 points each. Questions 3-5 are worth 20 points each.

The maximum total point is 100. 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 "My Studies" account **(**www.mitt.su.se) on December 30 at the latest.

Good luck!

Question 1 – Multiple choice (20 points, 4 points each)

1) When you have an omitted variable problem, the assumption that $E(u_i \mid X_i) = 0$ is violated. This implies that
A) the sum of the residuals is no longer zero.
B) there is another estimator called weighted least squares, which is BLUE.
C) the sum of the residuals times any of the explanatory variables is no longer zero.
D) the OLS estimator is no longer consistent.

2) The following OLS assumption is most likely violated by omitted variables bias:
A) $E(u_i \mid X_i) = 0$
B) (X_i, Y_i) i=1,..., n are i.i.d draws from their joint distribution
C) there are no outliers for X_i, u_i
D) there is heteroskedasticity

3) Consider the multiple regression model with two regressors X_1 and X_2, where both variables are determinants of the dependent variable. When omitting X_2 from the regression, then there will be omitted variable bias for $\widehat{\beta_1}$
A) if X_1 and X_2 are correlated
B) always
C) if X_2 is measured in percentages
D) if X_2 is a dummy variable

4) In the multiple regression model, the *t*-statistic for testing that the slope is significantly different from zero is calculated
A) by dividing the estimate by its standard error.
B) from the square root of the *F*-statistic.
C) by multiplying the *p*-value by 1.96.
D) using the adjusted R^2 and the confidence interval.

5) All of the following are true, with the exception of one condition:
A) a high R^2 or $\bar{R}^2$ does not mean that the regressors are a true cause of the dependent variable.
B) a high R^2 or $\bar{R}^2$ does not mean that there is no omitted variable bias.
C) a high R^2 or $\bar{R}^2$ always means that an added variable is statistically significant.
D) a high R^2 or $\bar{R}^2$ does not necessarily mean that you have the most appropriate set of regressors.

Question 2 – Multiple choice (20 points, 4 points each)

1) Consider a regression with two variables, in which X_{1i} is the variable of interest and X_{2i} is the control variable. Conditional mean independence requires
A) $E(u_i|X_{1i}, X_{2i}) = E(u_i|X_{2i})$
B) $E(u_i|X_{1i}, X_{2i}) = E(u_i|X_{1i})$
C) $E(u_i|X_{1i}) = E(u_i|X_{2i})$
D) $E(u_i) = E(u_i|X_{2i})$

2) The interpretation of the slope coefficient in the model $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$ is as follows:
A) a 1% change in X is associated with a β_1 % change in *Y*.
B) a change in *X* by one unit is associated with a β_1 change in *Y*.
C) a change in *X* by one unit is associated with a 100 β_1 % change in *Y*.
D) a 1% change in *X* is associated with a change in *Y* of 0.01 β_1.

3) Including an interaction term between two independent variables, X_1 and X_2, allows for the following except:
A) the interaction term lets the effect on *Y* of a change in X_1 depend on the value of X_2.
B) the interaction term coefficient is the effect of a unit increase in X_1 and X_2 above and beyond the sum of the individual effects of a unit increase in the two variables alone.
C) the interaction term coefficient is the effect of a unit increase in $\sqrt{(X_1 \times X_2)}$.
D) the interaction term lets the effect on Y of a change in X_2 depend on the value of X_1.

4) For the polynomial regression model,
A) you need new estimation techniques since the OLS assumptions do not apply any longer.
B) the techniques for estimation and inference developed for multiple regression can be applied.
C) you can still use OLS estimation techniques, but the *t*-statistics do not have an asymptotic normal distribution.
D) the critical values from the normal distribution have to be changed to 1.96^2, 1.96^3, etc.

5) In the regression model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$, where *X* is a continuous variable and *D* is a binary variable, β_3
A) indicates the slope of the regression when *D*=1.
B) has a standard error that is not normally distributed even in large samples since *D* is not a normally distributed variable.
C) indicates the difference in the slopes of the two regressions.
D) has no meaning since $(X_i \times D_i) = 0$ when $D_i = 0$.

Question 3 – Difference-in-differences (20 points)

Say that you evaluate the effect of a labor market training program that took place in the beginning 2004. You have access to yearly average outcomes for treated ($g = 1$) and untreated ($g = 0$) during the period 2000-2005. The average outcome, Y_{gt}, for the two groups are the following:

- $Y_{0t} = 100$ in all years for the control group ($g = 0$)
- $Y_{1,2000} = Y_{1,2001} = Y_{1,2002} = Y_{1,2003} = 100$ i.e., the outcome is 100 for the treatment group in all years up to 2003. $Y_{1,2004} = 150$ and $Y_{1,2005} = 250$.

Let $T_g = 1$ for the treated group and 0 for the control group and $After_t = 1$ during years 2004 and 2005 (i.e., the after period) and zero otherwise.

If you estimate the following equation with OLS:

$$Y_{gt} = \beta_0 + \beta_1 T_g + \beta_2 After_t + \gamma After_t \times T_g + u_{gt}$$

(i) What would be your estimate of γ? (5 points)

Now estimate the following model with yearly "treatment" effects using OLS:

$$Y_{gt} = \beta_0 + \beta_1 T_g + \lambda_{2001} d2001_t + \lambda_{2002} d2002_t + \lambda_{2003} d2003_t + \lambda_{2004} d2004_t$$

$$+\lambda_{2005} d2005_t + \delta_{2001} d2001_t \times T_g + +\delta_{2002} d2002_t \times T_g + \delta_{2003} d2003_t \times T_g$$

$$+\delta_{2004} d2004_t \times T_g + +\delta_{2005} d2005_t \times T_g + u_{gt}$$

where $d2001_t$ is a dummy variable taking the value 1 in year 2001 and zero otherwise, and so on.

(ii) What would be your estimates of δ_{2004} and δ_{2005}? (7 points)

(iii) Would you claim that the estimates of δ_{2004} and δ_{2005} are causal effects? Motivate! (8 points)

Question 4 – IV (20 points)

Say that we are interested in the causal effect of a teacher training program in a developing country. Teacher training takes place at village level and in 1st grade. Each village has only one class in 1st grade and class size is the same in each village.

The equation of interest at village level is

$$Y_j = \beta_0 + \beta_1 D_j + u_j$$

where Y is the average pupil test score (ranging from 0-100) in village j and D_j is a dummy variable taking the value 1 if teachers in the village receive the training program, 0 otherwise.

You had the possibility to randomize the teacher program to 200 villages where half of them (i.e., 100 villages) randomly received the opportunity to have their teachers participating in the training program. However, 20 villages of those who were given this opportunity did not use it. Villages randomised to the control could not use the teacher training program.

(i) Write down the first stage equation and explicitly state the estimate (i.e., the number) of the main slope coefficient of this regression. Interpret this coefficient. (4 points)

(ii) Say that the estimated effect of the instrument in the reduced form outcome equation is 8. What is the IV-estimate (i.e., the number) of the teacher training program? Interpret this coefficient. (6 points)

The teacher training program also affected the attendance rate of pupils such that when teachers become more educated this increased the propensity for pupils to show up at school. This in turn also affects pupils' performance at school. One interesting question is then to separate between the effect of the teacher training program itself and the effect of increased attendance rate on test scores. Someone suggested that you in the reduced form outcome equation control for whether the village has a high or low attendance rate among pupils ($HA_j = 1$ if village has high attendace rate, 0 otherwise). In other words, the suggestion is to estimate the following equation:

$$Y_j = \pi_0 + \pi_1 Z_j + \pi_2 HA_j + \varepsilon_j$$

where Z_j is the instrument taking value 1 if villages were randomized to the teacher program, and 0 otherwise.

(iii) Interpret the following assumption

$$E(u_j | Z_j, HA_j) = E(u_j | HA_j)$$

Do you think this strategy is a good idea? Does the estimate of β_1 has a causal interpretation? Motivate! (10 points)

Question 5 – credit question. Acemoglu & Angrist (2001) paper (20 points)

This is the abstract from the Acemoglu & Angrist (2001) paper:

"The Americans with Disabilities Act (ADA) requires employers to accommodate disabled workers and outlaws discrimination against the disabled in hiring, firing, and pay. Although the ADA was meant to increase the employment of the disabled, the net theoretical effects are ambiguous. For men of all working ages and women under 40, Current Population Survey data show a sharp drop in the employment of disabled workers after the ADA went into effect. Although the number of disabled individuals receiving disability transfers increased at the same time, the decline in employment of the disabled does not appear to be explained by increasing transfers alone, leaving the ADA as a likely cause. Consistent with this view, the effects of the ADA appear larger in medium-size firms, possibly because small firms were exempt from the ADA. The effects are also larger in states with more ADA-related discrimination charges."

Describe how they have econometrically reached to the main conclusion that the ADA seems to have a negative effect on the employment of disabled. Especially, the following issues must be included and explain intuitively as well as using equation notations:

(i) The main strategy used?
(ii) What is the key identifying assumption for estimating the causal effect of the ADA?
(iii) How do they econometrically investigate if this assumption is valid?

Write maximum 1 ½ A4-page for the answers!