Supplementary Written Examination in Econometrics (B2)

Spring 2017

2017-04-19 13.00-17.00 Venue: Bergsbrunnag. 15

Lars Forsberg, Department of Statistics, Uppsala University

Allowed means of assistance:

1. Pen or **pencil** (recommended) and eraser
2. **Calculators**,
   (a) 'programmable' calculator, e.g. calculator with graphing functions is OK.
   (b) Calculators with blue-tooth are not allowed.
   (c) Calculators with access to internet are not allowed.
   (d) Calcuators with which it is possible to send and recieve messages of any kind are not allowed.
3. **Physical (paper) dictionary** (no electronic dictionary allowed).
   (a) Dictionary must contain *no notes* of any kind.
   (b) Each student must have his/her own dictionary. It is not allowed for students to pass a dictionary between them.
4. **Ruler.**
5. Collection of formulae and Statistical Tables named '*Collection of Formulae and Statistical Tables for the B2-Econometrics and B3-Time Series Analysis courses and exams*', that the student brings to the exam location.
6. Please note that a collection of critical values for the Student´s t, Normal, Chi-square and F-distributions is given in the Appendix of the '*Collection of Formulae and Statistical Tables for the B2-Econometrics and B3-Time Series Analysis courses and exams*'.
7. Also note that the '*Test template*', that should be used when performing tests, is given in the '*Collection of Formulae and Statistical Tables for the B2-Econometrics and B3-Time Series Analysis courses and exams'.*

That is:

1. NO BOOK (except paper-dictionary) is allowed.
2. NO (student-written) notes are allowed.
3. NO other document than the one 'Collection of Formulae and Statistical Tables for Time Series Exam' is allowed.

Instructions: **Please note the following:**

1. Start with reading through the instructions!
2. Make sure you **follow** the instructions!
3. Start with reading through the exam.
4. You may write your solutions in Swedish or English.
5. Total score is **100** points
   (a) If you want the ECTS grades, please indicate that on the cover page!
   (b) For each task the maximum number of points is given within parenthesis, e.g. (16p in total).
   (c) For each subtask the number of points is given within parenthesis, e.g. (2p)
6. All solutions must be on separate sheets. No solutions on the questionnaire! (If so, they will be disregarded.)
7. Make sure your solutions are: easy to read and easy to understand, that is:
   (a) For each task that you solve, please start with a new sheet: after Task 1, start with a blank sheet for Task 2, etc.
   (b) Write the *task number* at the top of each page, in the

   ....**MIDDLE OF THE PAGE!!!**..

   Like:

   ....**TASK 1**...

   - if you write it in the upper left corner, the staple will cover it, and there is no for way for the examinator to know if the text of that sheet belongs to the previous sub-task or what it is. The Examinators will not make any 'qualified guesses' of what is being displayed on any given page. It is the responsibility of the student to make sure that every task and sub-task is easily identifiable.

(c) If you continue a sub-task on the next sheet of paper - indicate that at the top of the page - IN THE MIDDLE OF THE PAGE, like, for example:

....'Task 1B (cont.)'...

(d) Please separate each subtask A, B etc with a horizontal line across the sheet
____
if they are on the same sheet of paper - that way it will be easy for the examinator to actually see where one subtask ends and next begins.

(e) For examinator readability, it is highly recommended that you use a pencil, (and not a pen), which will allow you to erase and rewrite if you make a mistake. Crossed-over text and corrections using 'tipp-ex' will just cause blurriness and confusion to the examinator.

(f) For examinator readability: Write clearly, that is, letters, mathematical/statistical symbols and numbers should be easy recognizable!! Do not underestimate the correlation between readability and points scored, that is, when readability goes to zero, points scored also goes to zero, no matter your intentions or wheather *you* can read it or not.

(g) Also note that everything that you write will be taken at 'face value'. That is, for example, if you write β_1 the examinator will take that as a β_1 even though you may claim that it is given from the context it should be clear that you meant something else, like β_3. Thus, given this example, writing β_1, and that is not correct in that specific formula or statement, this will lead to subtraction of points, even if you will claim that it is just a typo, and that in another task or subtask, it is clear that you understand the issue.

(h) Please put the sheets in **order**, that is first Task 1, and then Task 2 etc...

8. Please keep the questionaire.

9. Do well!

Task 1

(12 points in total) Consider the following single linear regression

$$Y_i = \beta_1 + \beta_2 X_i + u_i.$$

A) (6p) Do the following:

1. Draw a Figure representing the Population Regression Function (PRF), draw the regression line, mark out what is displayed on the axes.
2. Mark out what distance is represented by β_1.
3. Mark out what distance is represented by β_2.
4. Mark out an arbitrary observation Y_i, given this observation, mark out the *conditional expected value* given the corresponding X_i, that is, mark out exactly where in the Figure this conditional expected value is 'located',
5. Write down a formula for the conditional expected value of Y.
6. Indicate in the Figure what distance that is represented by u_i.

B) (6p) Do the following:

1. In a SEPARATE FIGURE from the one in Sub-task A, draw a Figure representing the corresponding *Sample* Regression Function (SRF) for the model above, draw the sample regression line. Mark out what is displayed on the axes.
2. Mark out what distance is represented by $\widehat{\beta_1}$.
3. Mark out what distance is represented by $\widehat{\beta_2}$.
4. Mark out an arbitrary observation Y_i, and given this observation, mark out the *estimated conditional expected* value given the corresponding X_i, that is, mark out exactly where in the Figure this estimated conditional expected value is 'located'.
5. Write down a formula for the estimated conditional expected value of Y_i given that value of X_i.
6. Indicate in the Figure what distance that is represented by $\widehat{u}_i$.

For the sub-tasks below: interpret the β_2 coefficient - if nessecary - use a tranformation (scaling) of the coefficeint such that you can do the interpretation in terms of percentage points.

C) (4p) The log-lin model

$$\ln Y_i = \beta_1 + \beta_2 X_i + u_i.$$

D) (4p) The lin-log model

$$Y_i = \beta_1 + \beta_2 \ln X_i + u_i.$$

E) (4p) The log-log model, sometimes called the double-log model

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i.$$

Task 2

(34 points in total)

Note that, for each test you perform - for full score on the task - follow the test-template as outlined in the Collection of Formulae.

When it comes to education, it is well known that the number of years of education for a person, is highly correlated with that persons parents education. That is, the higher the education of the parents, the higher the education for the child. However, there are of course other variables that can help explain the level of (measured in years of) education. Find below some models for the number of years of education.

The variables are

- $Educ_i$ is the number of years of eduction for individual i
- $Feduc_i$ is the number of years of eduction for the father of individual i
- $Meduc_i$ is the number of years of eduction for the mother of individual i
- $Male_i$ is a dummy variable that takes on the value one if individual i is a male, zero otherwize
- iq_i is the intelligence quotient for individual i
- $Urban_i$ is a dummy variable that takes on the value one if individual i is a from an urban area, zero otherwize
- $Married_i$ is a dummy variable that takes on the value one if individual i is married, zero otherwize
- Age_i is the age of individual i.

Now, consider the following models.

$$Educ_i = \beta_1 + \beta_2 Feduc_i + \beta_3 Meduc_i + u_i \tag{1}$$

$$Educ_i = \beta_1 + \beta_2 Feduc_i + \beta_3 Meduc_i + \beta_4 male_i + u_i \tag{2}$$

$$Educ_i = \beta_1 + \beta_2 Feduc_i + \beta_3 Meduc_i + \beta_4 iq_i + u_i \tag{3}$$

$$Educ_i = \beta_1 + \beta_2 Feduc_i + \beta_3 Meduc_i + \beta_4 Male + \beta_5 iq_i + u_i \tag{4}$$

$$Educ_i = \beta_1 + \beta_2 Feduc_i + \beta_3 Meduc_i + \beta_4 Male_i + \beta_5 Urban_i + \beta_6 Married_i + \beta_7 Age_i + u_i \quad (5)$$

The estimation output for these models are given in figures 1 to 5. Use that information to solve the sub-tasks below. Please note that you may or may not make use of all output for solving the sub-tasks below, (and the sub-tasks in Task 3 - same output will be used in Task 3). It is part of the tasks to know *what* outputs to use for each sub-task.

A) (4p) For model (1) interpret the PRF-coefficient β_2.

B) (4p) Given the output for model (4) interpret the estimated coefficient in front of the variable iq_i.

C) (4p) For model (2) interpret the coefficient of determination.

D) (6p) Refer to model (1), perform a t-test of the hypothesis that the variable $Feduc_i$ explains nothing of the variation in the dependent variable against the alternative that it explains some of the variation in the dependent variable.

E) (6p) Using the output from model (2), perform a test whether the males have less than, or equal to half a year longer education against the alternative that they have at least half a year longer education.

F) (6p) Start with a relevant probabilistic statement and derive a 95% confidence interval for the coefficient in front of $Feduc_i$ for the model in (1). Comment on each and every step of the derivation. The end result should be such that it is possible to use in practice, i.e. next subtask.

G) (4p) Given your result in the previous sub-task, *calculate* and *interpret* the 95% confidence interval for the coefficient in front of $Feduc_i$ for the model in (1).

Dependent Variable: EDUC
Method: Least Squares
Date: 10/27/14 Time: 10:37
Sample (adjusted): 1 934
Included observations: 722 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.845360	0.304709	32.31072	0.0000
FEDUC	0.215635	0.027518	7.836229	0.0000
MEDUC	0.148691	0.032159	4.623569	0.0000

R-squared	0.206017	Mean dependent var	13.66343
Adjusted R-squared	0.203809	S.D. dependent var	2.236755
S.E. of regression	1.995846	Akaike info criterion	4.224160
Sum squared resid	2864.067	Schwarz criterion	4.243199
Log likelihood	-1521.922	Hannan-Quinn criter.	4.231509
F-statistic	93.28057	Durbin-Watson stat	1.504312
Prob(F-statistic)	0.000000		

Figure 1. Eviews output from estimation of model 1

Dependent Variable: EDUC
Method: Least Squares
Date: 10/27/14 Time: 10:36
Sample (adjusted): 1 934
Included observations: 722 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.838238	0.302312	32.54336	0.0000
FEDUC	0.207978	0.027387	7.594151	0.0000
MEDUC	0.134331	0.032164	4.176493	0.0000
MALE	0.536655	0.151911	3.532688	0.0004

R-squared	0.219582	Mean dependent var	13.66343
Adjusted R-squared	0.216321	S.D. dependent var	2.236755
S.E. of regression	1.980101	Akaike info criterion	4.209698
Sum squared resid	2815.135	Schwarz criterion	4.235083
Log likelihood	-1515.701	Hannan-Quinn criter.	4.219497
F-statistic	67.33992	Durbin-Watson stat	1.515938
Prob(F-statistic)	0.000000		

Figure 2. Eviews output from estimation of model 2

Dependent Variable: EDUC
Method: Least Squares
Date: 10/27/14 Time: 11:34
Sample (adjusted): 1 934
Included observations: 722 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.662092	0.478005	9.753235	0.0000
FEDUC	0.149931	0.025190	5.952021	0.0000
MEDUC	0.080041	0.029324	2.729560	0.0065
IQ	0.064621	0.004888	13.22047	0.0000

R-squared	0.361456	Mean dependent var	13.66343
Adjusted R-squared	0.358788	S.D. dependent var	2.236755
S.E. of regression	1.791097	Akaike info criterion	4.009058
Sum squared resid	2303.364	Schwarz criterion	4.034444
Log likelihood	-1443.270	Hannan-Quinn criter.	4.018857
F-statistic	135.4779	Durbin-Watson stat	1.605334
Prob(F-statistic)	0.000000		

Figure 3. Eviews output from estimation of model 3

Dependent Variable: EDUC
Method: Least Squares
Date: 10/27/14 Time: 11:35
Sample (adjusted): 1 934
Included observations: 722 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.773494	0.481555	9.912656	0.0000
FEDUC	0.147913	0.025180	5.874144	0.0000
MEDUC	0.075050	0.029420	2.550940	0.0109
MALE	0.243244	0.139125	1.748380	0.0808
IQ	0.063191	0.004949	12.76879	0.0000

R-squared	0.364167	Mean dependent var	13.66343
Adjusted R-squared	0.360620	S.D. dependent var	2.236755
S.E. of regression	1.788537	Akaike info criterion	4.007574
Sum squared resid	2293.586	Schwarz criterion	4.039306
Log likelihood	-1441.734	Hannan-Quinn criter.	4.019823
F-statistic	102.6637	Durbin-Watson stat	1.603014
Prob(F-statistic)	0.000000		

Figure 4. Eviews output from estimation of model 4

Dependent Variable: EDUC
Method: Least Squares
Date: 10/27/14 Time: 11:40
Sample (adjusted): 1 934
Included observations: 722 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.875595	0.879613	10.09034	0.0000
FEDUC	0.209746	0.027582	7.604438	0.0000
MEDUC	0.132622	0.032201	4.118598	0.0000
MALE	0.518962	0.157405	3.296984	0.0010
URBAN	0.057022	0.169136	0.337140	0.7361
MARRIED	-0.326596	0.236811	-1.379138	0.1683
AGE	0.037073	0.024264	1.527922	0.1270

R-squared	0.224000	Mean dependent var	13.66343
Adjusted R-squared	0.217489	S.D. dependent var	2.236755
S.E. of regression	1.978626	Akaike info criterion	4.212330
Sum squared resid	2799.197	Schwarz criterion	4.256754
Log likelihood	-1513.651	Hannan-Quinn criter.	4.229478
F-statistic	34.39872	Durbin-Watson stat	1.510784
Prob(F-statistic)	0.000000		

Figure 5. Eviews output from estimation of model 5

Task 3

(30 points in total)

Note that, for each test you perform - for full score on the task - follow the test-template as outlined in the Collection of Formulae.

For the entire Task 3, refer to the models and the output in Task 2

A) (6p) Perform a test of the model in (1).

B) (5p) A friend of yours that is studying economics, and has not taken yet taken any econometrics course, has formulated the model

$$Educ_i = \beta_1 + \beta_2 Feduc_i + \beta_3 Meduc_i + \beta_4 Male_i + \beta_5 Female_i + u_i$$

where

- Female$_i$ is a dummy variable that takes on the value one if individual i is a female, zero otherwize

She claims that there is something wrong with the estimation software that she is using, since she does not get any output from the model. However, there is nothing wrong with the software that she is using, it is something wrong with the model. How would you explain what is going on here, and why and in what way the model is not very well specified?

C) (5p) After your explanation, she now understands why the above model does not produce any output. For interpretation purposes, she wants a formulation of the model such that both the variable $Female_i$ and $Male_i$ are in the model. Suggest how this can be done such that the model does not suffer from the problem that her original model suffered from. Write down the model using a formula.

D) (4p) Another friend of yours has estimated the above models. And he is now comparing models (2) and (5) using the R^2 from the two models. Of the two, he has concluded that model (5) is the better model. Explain why this is not a proper way of doing the comparision (no matter the outcome of this particular comparision) between these two models.

E) (4p) Refer to sub-task D, and suggest another measure that would be better to use doing a first 'quick-and-dirty' comparision between the models (2) and (5). Explain why this measure is better in this situation, and interpret this measure for model (5) . (Do not perform any test in this sub-task)

F) (6p) Referring to model (5) , perform a test that answers the question whether the variables $urban_i$, $married_i$ and age_i should be included in model, that is, test if they contribute to the explanatory power of model (5) .

Task 4

(12 points in total)

Consider the single linear regression model where the model is

$$Y_i = \beta X_i + u_i, \tag{6}$$

$i = 1, ...n$, where $X_i > 0$ for all i is considered fixed in repeated sampling and

$$u_i \sim NID\left(0, \sigma^2\right) \tag{7}$$

(For the sake of notational simplicity there is no need to write out any (number) subscripts on β or X_i.) The Sample Regression Function is given by

$$Y_i = \widehat{\beta} X_i + \widehat{u}_i, \tag{8}$$

where $\widehat{\beta}$ is the OLS-estimator(s) of the $\beta-$parameter. Recall that the idea of the ordinary least squares (OLS) estimator is to minimize the sum of squared residuals, that is to minimize

$$\Sigma_{i=1}^{n} \widehat{u}_i^2$$

with respect to the parameter in the regression equation, that is, the OLS estimator for the slope coefficient is given by

$$\widehat{\beta}^{OLS} = \arg\min_{b} \Sigma_{i=1}^{n} \left(Y_i - bX_i\right)^2$$

A) (6p) Derive the *OLS estimator* of β in model (6), that is, find $\widehat{\beta}^{OLS}$. Carefully document and explain every step in the derivation, and state all assumptions that you need as you carry out each step of the derivation. Use the model as specified in (6) and (7).

B) (6p) Derive the *expected value* of the OLS-estimator you derived in subtask A, that is derive $E\left(\widehat{\beta}^{OLS}\right)$. Carefully document and explain every step in the derivation, and state all assumptions that you need as you carry out each step of the derivation. Use the model as specified in (6) and (7).