Written Examination in Econometrics (B2)

Fall 2016

2016-10-28 14.00-18.00

Fyrislundsgatan 80, room 1.

Lars Forsberg, Department of Statistics, Uppsala University

Allowed means of assistance:

1. Pen or **pencil** (recommended) and eraser
2. **Calculator**,
   (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 $\widehat{u}_i$.

Task 2

(34 points in total) Consider the following model for modeling 'Average hourly earnings' for a specifik industry

$$Ahe_i = \beta_1 + \beta_2 Age_i + \gamma_1 Female_i + \gamma_2 Bachelor_i + u_i, \quad (1)$$

where

- Ahe_i: Average[1] hourly earnings in dollars for individual i
- Age_i: Age in years for individual i
- $Female_i$: 1 if individual i is female, 0 if male (here we make the simplyfying assumption that all individuals can be considered either Female or male)
- $Bachelor_i$: 1 if individual i has a bachelor's degree; 0 if worker has a high school degree, (but not a bachelor's degree) - all individuals has either (only) a high school degree or a bachelor's degree.

Eviews output from estimating model (1) is given in Figure (2.1)

Ms Claire. L Inton is 27 years old and has a bachelor and is about to be employd by a company in this industry. She is now about to negotiate her salary with the CEO Mr Todd R. Ump and he says that 'given her profile' her hourly salary should be \$17.44. He claims that this number comes from 'a standard algorithm' that is being used within the industry. Ms Inton has talked to friends in the industry and she thinks this number is a bit low.

She gets hold of the 'algorithm' that Mr Ump has used, and it is the regression model above and the regression output below. She now contacts you as a statistician and ask you for your expertise.

A) (4p) To give her an initial idea of how 'good' the model is, explain to her how much of the variation in the Average hourly earnings the model explains, choose a relevant measure given the number of variables in the model. (No need to discuss if this is a 'good' model or not.)

B) (6p) Do a formal test of the model to see if it has any explanatory power at all. Choose the significance $\alpha = 0.05$. Make sure all the steps and calculations you make are easy to follow and understand. Fully document the test procedure as outlined in the test-templete.

[1] Average here refers to the average hourly earnings that individual i have obtaind over some period of time. It is NOT the average of all the individuals.

C) (5p) Now, do a point prediction of her salary given her profile (her values of the variables in the regression and the estimates of the parameters given in Figure 2.1), that is, do a point prediction of the Average hourly earnings she can expect given the model. (No need to do any prediction interval.)

D) (3p) You give her the result from the individual prediction and it is clear to her that this number is different from the one Mr Ump gave her, and she suspects this is due to discrimintation due to gender. She asks you to look further into the matter. Your task is: tell her how much less than a man with equal profile the estimated model tells her that she would get. (That is, interpret the relevant parameter estimate.)

E) (6p) She now calls Mr Ump and accuses him that he is discriminating her due to her gender. Mr Ump gets upset and says that the likelihood of that happening is lower than that he would ever run for President of the USA, which would *never* happen.

She calls you and asks you examine this from a statistical point of view. That is, perform a test of the null hypothesis that there is no discrimination due to gender against the against the alternative that women gets lower hourly earnings than men.

F) (6p) Now, *derive* a 95% confidence interval for the relevant parameter that captures potential discrimintation. State each and every assumption that you (have to) make in order to derive this interval.

G) (4p) Now, *calculate* and *interpret* a 95% confidence interval for the relevant parameter that captures potential discrimintation. Explicitly state the upper and lower limit of this interval.

Dependent Variable: AHE
Method: Least Squares
Date: 05/07/12 Time: 15:21
Sample: 1 7986
Included observations: 7986

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.883797	0.920292	2.046956	0.0407
AGE	0.439204	0.030529	14.38664	0.0000
FEMALE	-3.157864	0.180365	-17.50821	0.0000
BACHELOR	6.865150	0.178369	38.48856	0.0000

R-squared	0.189998	Mean dependent var	16.77115
Adjusted R-squared	0.189694	S.D. dependent var	8.758696
S.E. of regression	7.884317	Akaike info criterion	6.968129
Sum squared resid	496180.7	Schwarz criterion	6.971628
Log likelihood	-27819.74	Hannan-Quinn criter.	6.969327
F-statistic	624.0988	Durbin-Watson stat	1.892619
Prob(F-statistic)	0.000000		

Figure 2.1

Correlation				
	AHE	AGE	BACHELOR	FEMALE
AHE	1.000000	0.149176	0.369487	-0.135804
AGE	0.149176	1.000000	-0.001059	-0.025976
BACHELOR	0.369487	-0.001059	1.000000	0.116829
FEMALE	-0.135804	-0.025976	0.116829	1.000000

Figure 2.2

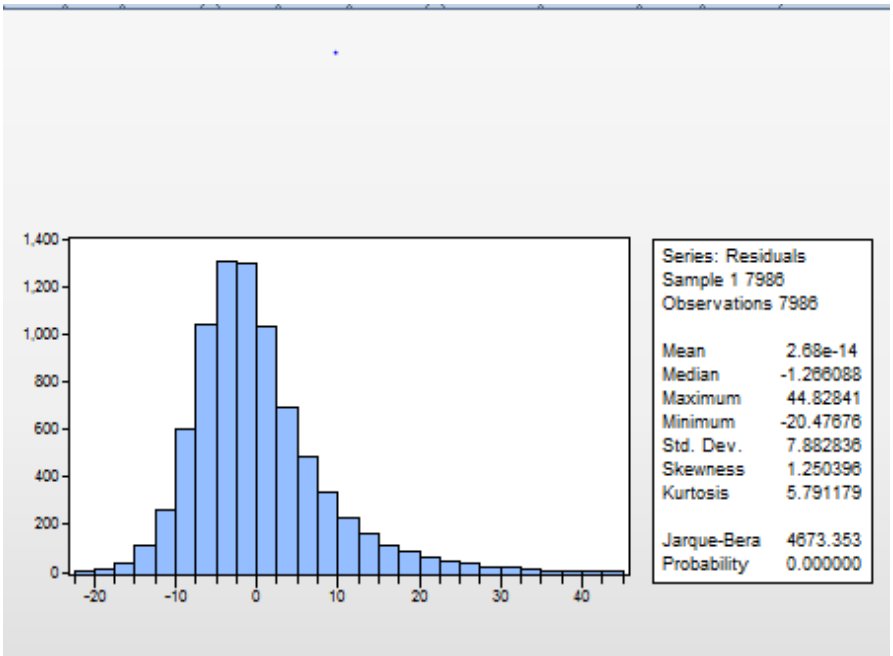
1,400
1,200
1,000
800
600
400
200
0
-20
-10
0
10
20
30
40
Series: Residuals
Sample 1 7986
Observations 7986
Mean 2.68e-14
Median -1.266088
Maximum 44.82841
Minimum -20.47676
Std. Dev. 7.882836
Skewness 1.250396
Kurtosis 5.791179
Jarque-Bera 4673.353
Probability 0.000000

Figure 2.3

Task 3

(36 points in total)

(This task is a continuation of task 2, Eviews output from task 2 might be needed in this task.)

Thanks to all your calculations Ms Ump reaches out to women working for Mr Ump and they form a class-action lawsuit against Mr Ump (who in fact has a so called Empire in this industry, with several high profile facicilities, and many employees).

Mr Umps lawyer, Mr Moe R. Onoe *now* claims that the model is flawed and that it cannot be used. The reason that they used it before was due to an evil conspiracy where the people that presented the model to Mr Ump had lied and decieved him that the model was OK. The statisticians that designed and estimated this model were corrupt and dis-honest and were in fact paid by Ms Lintons husband Bob, and the sole purpose of this endevaour was to set Mr Ump up for this class-action law suit. Mr Ump and his lawyers have now filed counter-lawsuits for mail order fraud, malicious prosecution, breach of trust, adviceing in bad faith, conspiracy to humiliate and slander.

Mr Onoe hires a statistician: Mr Carl R. Ook, that makes the following claims:

1. The model suffers from severe degree of multicollinearity and therefore all the inference is flawed. Also,

2. the error term is not normally distributed and therefore the inference is flawed.

A) (6p) Comment on the first claim, check the Eviews output(s) to see if there are any signs of multicollinearity. What especially do you look at/for? What would the signs of multicollinearity be? Are they there?
Also, given the test result from the test of no discrimination in the previous task - how, if at all, would you suspect that the result of that test to be different, if in fact, the model suffered from extremely high degree of multicollinearity?

B) (4p) Talking to Mr Ook it becomes clear to you that he has a very wierd idea about what multicollinearity is. (He worked as a cowboy in Texas before being hired by Mr Onoe, and he just does not understand what multicollinearity is in this context.)
Explain to him what it means just using words, (do not use any formulae in this sub-task, Mr Ook would not understand that).

C) (4p) To give Mr Ook an example of (exact) multicollinearity, reformulate the model (1), such that the new model in fact suffers from *perfect* or *exact* multicollinearity. Hint: add a variable - you may need to define the variable yourself.

D) (6p) Perform a test of normality of the error term (choose a test given the output you have). Choose the significance $\alpha = 0.05$. Make sure all the steps and calculations you make are easy to follow and understand. Fully document the test procedure as outlined in the test-templete.

E) (5p) Given the result of the test of normality, is there any validity to the claim that the error term is not normally distributed? If so, is the inference flawed and therefore not valid in this situation? Is there anything that 'saves' the inference (so that it still valid in this situation), if so, what is it, and how come it 'saves' the inference?

Realizing that Mr Ook is not a very good statistican, Mr Onoe hires Mr J. the 2^{nd} H. Utt, (who claims to be a statistician) and he says that the two variables gender and bachelor should *not* be included in the model. That is, gender and education (in terms of a Bachelor degree) taken together, has nothing whatsoever to do with the Average hourly earnings. (Mr Utt explains that, where he comes from, females do not get paid for their work anyway and education does not even exist, and he has been doing fine throughout his life in his particular line of work, albeit that line of work is somewhat different from this industry.)

So, Mr Utt has, with some external help, estimated another model

$$Ahe_i = \beta_1 + \beta_2 Age_i + e_i$$

(output in Figure 3.1).

F) (5p) Compare the explanatory powers of the two models using a relevant measure in this context. What would you tell Mr Utt concerning the difference in explanatory powers of the two models?

G) (6p) Mr Utt's father passed away in 1983, and the measure you mention brings up bad memories in his mind, something to do with the name of some character involved in the tragic passing of his father. He becomes very upset and claims that this 'somewhat arbitrary comparision of these two numbers' does not really prove anything in a staticially sound way.

So, now, using relevant output from task 2 and 3, perform a formal test of the null hypothesis that the two variables *female* and *bachelor* simultaneously does *not* explain any of the variation of the dependent variable. Choose the significance $\alpha = 0.05$. Make sure all the steps and calculations you make are easy to follow and understand. Fully document the test procedure as outlined in the test-templete.

Dependent Variable: AHE
Method: Least Squares
Date: 10/23/16 Time: 23:13
Sample: 1 7986
Included observations: 7986

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.324184	1.002230	3.316787	0.0009
AGE	0.451931	0.033526	13.48022	0.0000

R-squared	0.022254	Mean dependent var	16.77115
Adjusted R-squared	0.022131	S.D. dependent var	8.758696
S.E. of regression	8.661234	Akaike info criterion	7.155842
Sum squared resid	598935.5	Schwarz criterion	7.157591
Log likelihood	-28571.28	Hannan-Quinn criter.	7.156441
F-statistic	181.7164	Durbin-Watson stat	1.857141
Prob(F-statistic)	0.000000		

Figure 3.1

You hand over all these results to Ms Lintons lawyers and and they take it to court. The outcome is, at this point in time, still pending.

Task 4

(18 points in total)

Consider the following model

$$Y_i = \beta_2 X_{i,2} + u_i$$

A) (6p) Derive the OLS estimator for β_2. State any assumptions, if you make any, as you make them, that you need for this derivation. Any assumptions, that are not nessecary for this derivation, will result in reduction of points.

B) (6p) Derive $E\left(\widehat{\beta_2}\right)$ given the OLS estimator above. You do not have to restate any assumptions that you stated in subtask A. But over and above those, state any assumptions, if you make any, as you make them, that you need for this derivation. Any assumptions, that are not nessecary for this derivation, will result in reduction of points.

C) (6p) Allowing for heteroscedasticity and correlated error terms, derive $V\left(\widehat{\beta_2}\right)$ (start with the definition of the variance). You do not have to restate any assumptions that you stated in subtask A and B. But over and above those, state any assumptions, if you make any, as you make them, that you need for this derivation. Any assumptions, that are not nessecary for this derivation, will result in reduction of points.