Written Examination in Econometrics (B2)

Spring 2016

2016-03-23 08.00-12.00

Bergsbrunnagatan 15, room2.

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

(18 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 $\widehat{u}_i$.

C) (3p) Write down a log-lin model with an intercept and one and only one varying regressor. Is it linear in the parameters? Is it linear in the variables? Interpret the slope coefficient.

D) (3p) Write down a lin-log model with an intercept and one and only one varying regressor. Is it linear in the parameters? Is it linear in the variables? Interpret the slope coefficient.

Task 2

(29 points in total)

Consider the following models for Miles per Gallons of fuel consumption for cars

$$MPG = \beta_1 + \beta_2 Age + \beta_3 Weight + \beta_4 Acceleration + \beta_5 Displacement + \beta_6 Horsepower + u \quad (1)$$

$$MPG = \alpha_1 + \alpha_2 Age + \alpha_3 Weight + \alpha_4 Displacement + u \quad (2)$$

$$MPG = \gamma_1 + \gamma_2 Age + \gamma_3 Weight + u \quad (3)$$

Where the variables are

1. MPG: Miles per Gallon (gasoline)
2. Cylinders: number of cylinders
3. Displacement: (engine displacement) 'size of engine' in square inches (sv: cylinder volym)
4. horsepower: 'strength' of engine
5. weight: weight of car (in pounds)
6. Acceleration: speed of speedincrease
7. Age: age of car

For all subtasks in this task, find relevant Eviews output in the figures below.

A) (3p) Given the model in (1), interpret $\widehat{\beta_2}$.

B) (3p) Given the model in (1), interpret R^2.

C) (6p) Mr Berra R. Aggare likes big engines, and in an environmental debate with Ms Åsa Wroomson on twitter, he claims that the Displacement has nothing to do whatsoever with the miles per gallon consumption of the car (does not increase nor decrease), and he assigns you as a statistician to perform a test to test this claim. Use the output from model (2) to perform this test. Use 5% significance level. Document the test as outlined in the test template.

D) (4p) Now, Ms Wroomson is surprised by this result and she asks you why this is the case. Given all the output below, what would you say is the reason for this somewhat surprising result (it is surprising in the sense that one might think that the larger the engine, the higher the fuel consumption, and thus the lower the MPG). Comment on each and every sign of that there is a (statistical) problem.

E) (4p) Compare the explanatory power of models (1) and (3) using a relevant measure. For model (1), interpret this measure. (No need to perform any test in this subtask.)

F) (6p) Ms Wroomson claims that the there should be a tax on the age of the car. That is, the older the car, the higher tax (sv: fordonsskatt), since this will give incentives for people to buy newer cars. The reason behind this is the belief that older cars have higher fuel consumption, that is, lower miles per gallon, than newer cars, and that there are environmental gains to be made if the car fleet would, as a aggregate, consume less gasoline. Use output from model (3) to perform a test of the null hypothesis that the effect of age is equal to or bigger that zero, against the alternative that it is negative. Use 5% significance level. Document the test as outlined in the test template.

G) (3p) Refer to Figure 2.6, for the model (3) , what kind of a potential problem could be present here? If so, what would be the consequences in terms of the OLS estimation? Is OLS still unbiased? Is it effiecient? (No need to prove or derive anything.)
This graph might encourage someone to claim we cannot trust the t-tests anymore. In this situation, how would you deal with this problem? What kind of method or estimator would you use if you were to run the estimations yourself?

Dependent Variable: MPG
Method: Least Squares
Date: 03/20/16 Time: 14:59
Sample: 1 392
Included observations: 392

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	72.04207	2.727733	26.41097	0.0000
AGE	-0.754115	0.052612	-14.33357	0.0000
WEIGHT	-0.006874	0.000665	-10.33256	0.0000
ACCELERATION	0.090324	0.101907	0.886333	0.3760
DISPLACEMENT	0.002782	0.005462	0.509312	0.6108
HORSEPOWER	0.001020	0.013763	0.074121	0.9410

R-squared	0.808767	Mean dependent var	23.44592
Adjusted R-squared	0.806289	S.D. dependent var	7.805007
S.E. of regression	3.435184	Akaike info criterion	5.321206
Sum squared resid	4554.989	Schwarz criterion	5.381990
Log likelihood	-1036.956	Hannan-Quinn criter.	5.345296
F-statistic	326.4950	Durbin-Watson stat	1.224330
Prob(F-statistic)	0.000000		

Figure 2.1: Eviews output

Dependent Variable: MPG
Method: Least Squares
Date: 03/20/16 Time: 14:59
Sample: 1 392
Included observations: 392

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	73.56971	2.209504	33.29693	0.0000
AGE	-0.758043	0.051001	-14.86322	0.0000
WEIGHT	-0.006664	0.000571	-11.67003	0.0000
DISPLACEMENT	0.000283	0.004744	0.059754	0.9524

R-squared	0.808182	Mean dependent var	23.44592
Adjusted R-squared	0.806699	S.D. dependent var	7.805007
S.E. of regression	3.431551	Akaike info criterion	5.314053
Sum squared resid	4568.910	Schwarz criterion	5.354576
Log likelihood	-1037.554	Hannan-Quinn criter.	5.330114
F-statistic	544.9172	Durbin-Watson stat	1.230797
Prob(F-statistic)	0.000000		

Figure 2.2: Eviews ouput

Dependent Variable: MPG
Method: Least Squares
Date: 03/20/16 Time: 15:00
Sample: 1 392
Included observations: 392

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	73.50167	1.891077	38.86763	0.0000
AGE	-0.757318	0.049473	-15.30781	0.0000
WEIGHT	-0.006632	0.000215	-30.91071	0.0000

R-squared	0.808180	Mean dependent var	23.44592
Adjusted R-squared	0.807194	S.D. dependent var	7.805007
S.E. of regression	3.427153	Akaike info criterion	5.308960
Sum squared resid	4568.952	Schwarz criterion	5.339353
Log likelihood	-1037.556	Hannan-Quinn criter.	5.321006
F-statistic	819.4730	Durbin-Watson stat	1.230426
Prob(F-statistic)	0.000000		

Figure 2.3: Eviews output

Correlation	MPG	WEIGHT	HORSEPO...	DISPLACEM...	CYLINDERS	AGE	ACCELERA...
MPG	1.000000	-0.832244	-0.778427	-0.805127	-0.777618	-0.580541	0.423329
WEIGHT	-0.832244	1.000000	0.864538	0.932994	0.897527	0.309120	-0.416839
HORSEPO...	-0.778427	0.864538	1.000000	0.897257	0.842983	0.416361	-0.689196
DISPLACEM...	-0.805127	0.932994	0.897257	1.000000	0.950823	0.369855	-0.543800
CYLINDERS	-0.777618	0.897527	0.842983	0.950823	1.000000	0.345647	-0.504683
AGE	-0.580541	0.309120	0.416361	0.369855	0.345647	1.000000	-0.290316
ACCELERA...	0.423329	-0.416839	-0.689196	-0.543800	-0.504683	-0.290316	1.000000

Figure 2.4: Correlation matrix

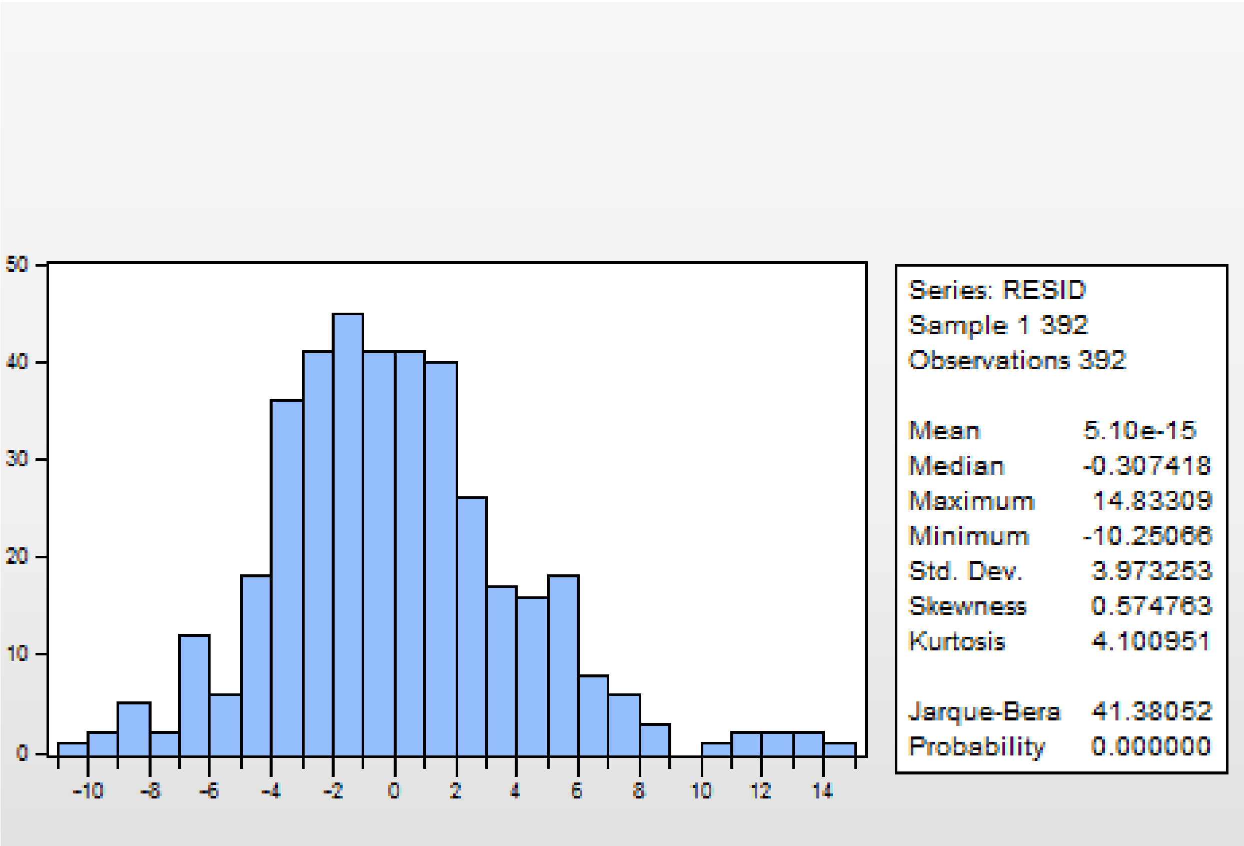

Figure 2.5: Residual analysis from Model 1

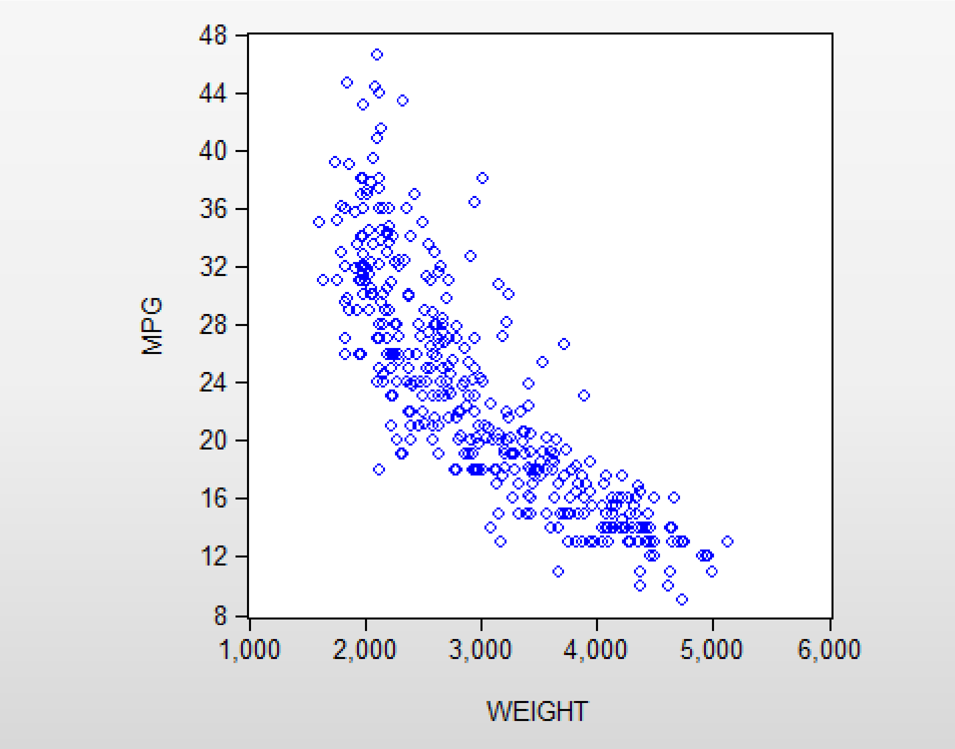

Figure 2.6: Scatterplot

Task 3

(31 points in total)

For models and output, refer to task 2.

A) (6p) An auto mechanic by the name of Mr M. Eka, says that, roughly speaking, this is how it works:

1. As technology develops over time, engines becomes more and more fuel efficient, that is, age of car is definitely a factor when it comes to MPG.
2. A large engine (displacement) gives the car a higher weight, but also, the larger the engine, the higher the number of horsepower (engine strenght) and thus the better the acceleration.
3. The heavier the car, the more fuel (all else being equal) it takes to actually 'move the car'.

This line of reasoning implies that it would be sufficient to have a model that includes the variables Age and weight in order to model MPG. Perfom a test to test the claim that, given that age and weight is in the model, the variables displacement, horsepower and acceleration does not contribute anything at all to the explanatory power of the model. Use 5% significance level. Document the test as outlined in the test template.

B) (6p) After listening to Mr Eka, and having taken a break from the car-debate, (spending the weekend to repaint her boat and to think long and hard about environmental issues), Ms Wroomson now claims that she has found the ultimate model for gasoline consumption. The only variables that are needed are age and weight. Perform a test of such a model. Use 5% significance level. Document the test as outlined in the test template.

C) (6p) Now, Mr G. Peaceoline, for reasons unknown, does not like four wheel drive cars very much at all (maybe due to the belief that four wheel drive cars consume more gasoline, all else being equal). Now, a car can have one of three wheel drive systems:

1. Rear wheel drive
2. Front wheel drive and
3. Four wheel drive

The variable 'wheel drive' is not in this dataset. But let's say we obtained a dataset where this variable is included. Starting with model (3) , how would you implement this variable? That is, the task is to formulate a model where the variable 'wheel drive' is included.

D) (6p) Now, finding himself in a somewhat uphill battle concerning the debate on twitter, Mr Aggare claims that the tests above are not valid since the normality assumption of the error terms is not fulfilled (for any of the models), but in particular not for model (1). Figure 2.5 displays some output to settle this dispute. Perform a test to test the null that the error term is normally distributed. Use 5% significance level. Document the test as outlined in the test template.

E) (4p) Now, filled with self confidence, Mr Aggare claims that, given the result of the test in the previous subtask, all the tests are useless and the results cannot to be trusted. Please explain to him why this is not nessecarily the case (you may use more that 140 characters). What, in fact, is it that we need to have following a normal distribution? It is possible to perform a test of that? If not, how can we know if it is fulfilled? In case the error term is not normally distrubuted, what do we need to rely upon for the inference to be valid?

F) (3p) Encouraged by model 3, Ms Wroomson suggests a law to force automobile producers to produce cars that are as new and as light as possible, say, weigh only 500 pounds. To motivate this law, she wants you to make a point prediction of the MPG of such a car, that is, a brand new car (age=0) and weight being 500 pounds.

Now, from an enginering point of view, model 3 might not be the perfect model. That aside, let's focus on the statistical part of it: the range of the variable weight in the sample is 1630-5140, the range of age is 10-30 years. Explain to her the 'danger' and the potential problem(s) that could be associated with such a point prediction. (No need to actually perform the point prediction.)

Task 4

(22 points in total)

Consider the following model

$$Y_i = \beta_2 X_2 + u$$

A) (6p) Derive the OLS estimator for β_2. State any assumptions that you make as you make them.

B) (6p) Under the assumption of $Var(u) = \sigma_i^2$ and $Cov(u_i, u_j) = 0$, derive the variance of the OLS estimator. State any assumptions that you make as you make them.

C) (10p) According to the literature, it is reasonable to assume that in a particular application, the regressor X_2 is in fact correlated with the error term. Now

1. Give an example of why this could be the case.
2. What happens with the OLS estimator in this situation? (No need for formuale or any derivations what so ever.)
3. Suggest a remedy for this situation, and explain the intuition for this remedy.
4. Say that you have two variables X_3 and X_4 that are uncorrelated with the error term, but, at least to some extent correlated with X_2. How could you use these variables to remedy this situation? Just explain what you would do, no need to prove that it would work.
5. Now, these two variables X_3 and X_4 are highly correlated with each other (0.99). Given the suggestion in the prevous sub-sub task, is this a problem? If not, why not, if so, why?