ECOM-G314 Econometrics 1
Example exam

Note that many of the question options are randomized, so the order of choices on your exam may not be the same as in this document.
For questions and corrections, contact heikki.korpela@helsinki.fi.

1. Consider the following linear regression model

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$$

where ε_t is a zero-mean error term, and assumptions (AS1*) - (AS4*) hold. In addition, there is a variable z_t such that $E(\varepsilon_t \mid z_t) = 0$.

The parameter vector $\beta = (\beta_1, \beta_2)'$ is estimated by ordinary least squares.

Which covariance matrix estimators of the OLS estimator b are consistent in each of the cases below, where additional information about the error term ε_t is given?

Which covariance matrix estimator of the OLS estimator b do you choose based on the properties of the error term (if mentioned) and the p-values of the diagnostic tests in each of the cases where test results are given? Use the 5% level of significance in all tests.

The following acronyms are used:
HO = Conventional covariance matrix estimator assuming homoskedasticity
HC = Heteroskedasticity consistent covariance matrix estimator
HAC = Heteroskedasticity and autocorrelation consistent covariance matrix estimator

In this question, you want to remember that HC and HAC are still consistent even if there is no autocorrelation and no heteroskedasticity. Those are "safe" to use even if they aren't strictly required. (The only thing you win by using HO errors is better precision if the errors are truly homoskedastic with no serial correlation, but the gains are usually small.) Other than that, this is a simple case of eliminating the covariance estimators which are non-consistent.

(a) $E(\varepsilon_t^2) = 1.8$, and $E(\varepsilon_t \varepsilon_{t-j}) = 0, j = 1, 2, 3, \ldots$

The variance of the error term is just a constant (it does not depend on x or z), and there is no serial correlation in error terms. HO, HC and HAC are all consistent.

(b) $E(\varepsilon_t^2) = 0.5 \exp(0.5 z_t), E(\varepsilon_t \varepsilon_{t-1}) = -0.4$, and $E(\varepsilon_t \varepsilon_{t-j}) = 0, j = 2, 3, \ldots$

The errors are heteroskedastic, as the variance depends on z_t (which we assume is non-degenerate, i.e., not just a constant). HO is not consistent because of both issues, and HC is not consistent because there is also autocorrelation of the first order. Only HAC is consistent. Note that it does not matter whether the heteroskedasticity depends on x or z.

(c) $E(\varepsilon_t^2) = 0.22z_t$, and $E(\varepsilon_t \varepsilon_{t-j}) = 0, j = 1, 2, \ldots$

The error term is heteroskedastic, as the variance depends on z_t. However, there is no serial correlation in error terms. Thus, HC and HAC are both consistent.

(d) $E(\varepsilon_t^2) = 0.31x_{t-1}^2$, and $E(\varepsilon_t \varepsilon_{t-j}) = 0, j = 1, 2, \ldots$

The error term is heteroskedastic, as the variance depends on x_{t-1}. Again, there is explicitly no serial correlation in *error terms*. Thus, HC and HAC are both consistent.

2. Consider the following linear regression model

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i,$$

where ε_i is a zero-mean error term with variance σ^2. Of the regressors, x_{i3} and x_{i4} satisfy $E(\varepsilon_i | x_{i3}) = E(\varepsilon_i | x_{i4}) = 0$, while x_{i2} is endogenous and $E(\varepsilon_i | x_{i5}) \neq 0$. In addition, the variables q_i and w_i are orthogonal to the error term, the variable r_i is such that $E(\varepsilon_i r_i) \neq 0$, and the variable p_i is such that $E(\varepsilon_i p_i) = 0$. The observations are a random sample from an independent joint distribution. Unless otherwise stated, in each case below, the instruments (or moment conditions) are relevant.

Which of the following statements are correct?

In this question, you want to pay attention to definitions. A number of things here mean the same, but are just worded differently:

- Estimator b is a consistent estimator of β if $b \underset{p}{\rightarrow} \beta$. (p. 34 in book)
- Regressor x_{ij} is endogenous if $E(\varepsilon_i x_{ij}) \neq 0$ (in our setting, you can regard $E(\varepsilon_i \,|\, x_{ij}) \neq 0$ as equivalent), and exogenous otherwise. (p. 147)
- Variable q_i is orthogonal to the error term iff $E(\varepsilon_i q_i) = 0$. (p. 66 in book)
- Given the above, we note that in the model there is the constant, two exogenous regressors (x_{i3}, x_{i4}) and two endogenous regressors (x_{i2}, x_{i5}).
- Additionally, the constant "regressor" 1 is exogenous.
- There are three valid instruments p_i, q_i, w_i, as they are also exogenous and relevant. (There is also one potential instrument r_i that is not valid as it is endogenous, but none of the statements below refer to it so we may ignore it.)
- A GMM model with the same number of moment conditions as parameters is exactly identified. A model with more moment conditions than parameters is over-identified (= it is identified *and* you can run over-identifying restrictions tests). (p. 165 in book)
- OLS is not consistent if there are endogenous regressors. (p. 146–147 in book)
- IV is consistent if the observations are iid, instruments are relevant and exogenous, and the instruments and dependent variables have nonzero finite fourth moments. (p. 151–152 in book) The question clearly focuses on the relevance and exogeneity requirements.

(a) The OLS estimator $b \underset{p}{\rightarrow} \beta = (\beta_1, \ldots, \beta_5)'$.

False. OLS is not consistent because there are endogenous regressors.

(b) The variables p_i and q_i together with the exogenous regressors as instruments exactly identify the parameter vector $\beta = (\beta_1, \ldots, \beta_5)'$.

True. With two instruments and two endogenous variables and the error term having mean zero (analogous to the constant term), the model has five moment conditions for five parameters. It is just-identified with IV/2SLS.

(c) The IV estimator with $z_i = (1, x_{i2}, x_{i4}, p_i, w_i)'$ as instruments consistently estimates the parameter vector $\beta = (\beta_1, \ldots, \beta_5)'$.

False. This is an easy place to go wrong, because you see two endogenous variables and two regressors. However, x_{i2} is explicitly stated as being endogenous, so it should not be used as an instrument.

(d) The IV estimator with p_i and q_i and the exogenous regressors as instruments, $\hat{\beta}_{IV} \underset{p}{\rightarrow} \beta = (\beta_1, \ldots, \beta_5)'$.

True. Two valid instruments for two endogenous variables.

(e) The value of the over-identifying restrictions test related to the two-stage least squares estimator of $\beta = (\beta_1, \ldots, \beta_5)'$ with $z_i = (1, x_{i2}, x_{i4}, p_i, q_i, w_i)'$ as instruments is positive.

True. There is one more variable defined as instrument than there are endogenous variables. Thus, the test can technically be performed, and will yield a positive test statistic (this is by construction, as the support of the F-distribution is non-negative). One of the instruments (x_{i2}) is not exogenous, but that doesn't mean you technically can't run the test.

(f) The parameter vector $\beta = (\beta_1, \ldots, \beta_5)'$ is consistently estimated by the GMM with moment conditions $E(\varepsilon_i) = E(\varepsilon_i x_{i3}) = E(\varepsilon_i x_{i4}) = E(\varepsilon_i p_i) = E(\varepsilon_i q_i) = E(\varepsilon_i w_i) = 0$.

True. This is just the generalization of IV when you have more instruments than endogenous variables. The moments specify instrument exogeneity.

(g) The GMM estimator with moment conditions $E(\varepsilon_i) = E(\varepsilon_i x_{i3}) = E(\varepsilon_i p_i) = E(\varepsilon_i q_i) = E(\varepsilon_i w_i) = 0$, $\hat{\beta}_{GMM} \underset{p}{\rightarrow} \beta = (\beta_1, \ldots, \beta_5)'$.

True. Note that the question gives a blanket statement that any moment conditions are relevant. Thus, while using exogenous x_{i4} directly might make sense here, it isn't required for identification.

(h) The parameter vector $\beta = (\beta_1, \dots, \beta_5)'$ is consistently estimated by the GMM with moment conditions $E(\varepsilon_i) = E(\varepsilon_i x_{i3}) = E(\varepsilon_i x_{i4}) = E(\varepsilon_i p_i) = E(\varepsilon_i w_i) = 0$.

True.

3. Select the correct alternative in each case below.

Consider the linear regression model

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i,$$

where ε_i is a zero-mean error-term with constant variance σ^2, and assumptions (AS1) - (AS4) hold. The model is estimated by OLS on 498 observations, and the point $(\beta_2, \beta_3) = (0.8, 0.5)$ belongs to the 95% joint confidence region of (β_2, β_3). Based on this, it can be inferred that

(a) $H_0 : \beta_3 = 0.5$ is not rejected at the 5% level of significance.

This cannot be inferred, because the null is for one parameter, while the confidence region is for two parameters (corresponding to a joint hypothesis).

(b) $H_0 : \beta_3 = 0.5$ is rejected at the 10% level of significance.

Same as a).

(c) $H_0 : (\beta_2, \beta_3) = (0.8, 0.5)$ is not rejected at the 5% level of significance.

This is true because of the duality of tests and confidence regions: a confidence region is *defined* as the set of parameters which, if set as a null hypothesis, are not rejected. (p. 25 in the book.)

(d) $H_0 : (\beta_2, \beta_3) = (0.8, 0.5)$ is rejected at the 10% level of significance.

This cannot be inferred from the information in the question. For OLS, the corresponding 90% level confidence set will be smaller than a 95% level one, but we do not know by how much. We are not even being told whether the point $(0.8, 0.5)$ might actually be the point estimate (if it is, then it is actually included in the confidence set at *any* significance level).

A simple example may be helpful here. Consider an example where there is only one regressor, the constant β_1, there are $n = 100$ observations, and the point estimate happens to be $\overline{y} = 0$ with a standard error of 1. The confidence interval is then $\overline{y} \pm \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}}$, where α is the level of significance. For $\alpha = 0.05$, $z_{\alpha/2} \approx 1.96$, yielding the interval $0 \pm \frac{1.96 \cdot 1}{10}$; for $\alpha = 0.1$, $z_{\alpha/2} \approx 1.64$, clearly yielding a more narrow interval. For $K > 1$ parameters, the confidence set is a region in R^K, but the principle is the same.

The confidence set has a duality with tests. Suppose we were to repeat our research setting (drawing a new sample and re-estimating) an extremely large number of times. Then the confidence set at significance level α for parameters θ includes all such values of the parameters which would be accepted at least $1 - \alpha$ number of times as a null hypothesis (or, equally, rejected at most α number of times). As the significance level

α is increased, the confidence level $1 - \alpha$ decreases, and we need to allow for a smaller number of parameters which might plausibly be consistent with the data.

(e) none of the above is correct.

This is clearly false because of c).

In this question, you needed to understand the duality of confidence regions and hypotheses, and that joint hypotheses are different from hypotheses about individual parameters/coefficients.

Note that the structure of the question implies that you need to pick the one item that is known to be correct with the available data. This does not mean all the other ones are incorrect, only that their correctness cannot be inferred.

Let A be an $r \times r$ invertible matrix, x is an $r \times 1$ vector, and y is a normally distributed scalar random variable with mean zero and variance σ^2. Moreover, plim $A_N = A$, $x_N \underset{p}{\rightarrow} x$, $y_N \underset{d}{\rightarrow} y$, and $E(x_N) = x$. Based on this information, which of the following statements is correct?

(a) x_N is a biased estimator of x.

False. An estimator b is an unbiased estimator of β if $Eb = \beta$.

(b) x_N is a consistent estimator of x.

True. An estimator b is a consistent estimator of β if $b \underset{p}{\rightarrow} \beta$.

(c) $E[\log(x_N)] = E[\log(x)]$.

False. This cannot be inferred; among other things, log is not linear. The simplest counter-example is probably where x_N gets arbitrarily close to zero as N grows, and has $x \equiv 0$ as the limiting distribution.

(d) $A_N x_N y_N \underset{d}{\rightarrow} Axyy'x'A'$.

False. The transpose $y'x'A'$ in the limiting distribution is in excess and yields the wrong distribution.

(e) $A_N y_N \underset{d}{\rightarrow} z$, where $z \sim \mathcal{N}(0, A\sigma^2)$.

False. z's covariance matrix is $A\sigma^2 A'$. (See any elementary probability/statistics material, or book p. 465.)

In this question, you needed to at least remember and understand the definitions of bias and consistency. The Slutsky lemmas are covered on slides 7 of the "Asymptotic properties of the OLS estimator" slides.

4. Consider the following linear regression model:

$$\text{logtraining}_i = \beta_1 + \beta_2 \text{grant}_i + \beta_3 \text{logsales}_i + \beta_4 \text{empl}_i + \varepsilon_i,$$

where logtraining_i is the log of hours of training per employee that firm i offers to its personnel, grant_i is a dummy variable equal to one if the firm received a job training grant from the government in 1988, and zero otherwise, logsales_i is the log of annual sales (in millions of euro) and empl_i is the number of employees of firm i. Finally, ε_i is assumed to be a zero-mean homoskedastic error term, and assumptions (AS1) - (AS4) are assumed to hold.

The model was estimated on a data set consisting of 405 firms by ordinary least squares. The estimation result is the following (conventional standard errors based on assuming homoskedasticity in parentheses):

$$\widehat{\text{logtraining}}_i = \underset{(43.41)}{46.67} + \underset{(0.07)}{0.12} \text{grant}_i + \underset{(0.04)}{0.07} \text{logsales}_i - \underset{(0.006)}{0.007} \text{empl}_i$$

The p-value of the White test equals 0.01.

It was suspected that empl_i is endogenous, and the model was also estimated by two-stage least squares (2SLS) using the hours of training in 1987, 1986 and 1985 as instruments. The F-statistic testing their joint significance in the reduced-form regression equals 5.22, with p-value 0.015. The value of the over-identifying restrictions test equals 6.81 with p-value 0.033. The p-value of the Durbin-Wu-Hausman test is 0.282.

Fill in the blanks below. Use the point as the decimal separator. The first six bullet points are related to the OLS estimation result. Each correct answer yields 1,25 points.

For this question, you mostly need to remember the rule of thumb: when a variable is specified in logs, discuss relative changes of that variable. When a variable is specified in levels (not in logs), discuss unit changes. When a variable is a dummy, discuss the effect of that dummy being true for an observation. The slides on "Interpretation of linear regression model" are additional reading, p. 7–8 in particular. Problem 1 in homework assignment 1 may also be helpful.

(a) Comparing two firms that have the same annual sales and the same number of employees, the one that has received the grant is expected to offer . . . hours/% less/more training.

We look at the effect of the grant regressor, ceteris paribus. The dependent variable is in logs, and the regressor is a dummy. Thus, the correct answer is 12% more.

(b) Comparing two firms that have not received the grant and have the same annual sales, the one with one more employee is expected to offer ... hours/% less/more training.

We look at the effect of the empl regressor, ceteris paribus. The dependent variable is in logs, and the regressor is in levels. Thus, the correct answer is 0.7% less.

(c) Comparing two firms that have received the grant and have the same number of employees, the firm with 1% lower annual sales is expected to offer ... hours less/more training.

We look at the effect of the log sales, ceteris paribus. The dependent variable is in logs, and the regressor is also in logs, but the question is about a relative negative change in sales. Thus, the correct answer is 0.07% less.

(d) According to a one-sided t-test test of $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 > 0$, β_2 is/is not statistically significantly different from zero at the 5% level of significance.

This is a straightforward calculation of comparing $(b_2 - 0)/se = 0.12/0.07$ to 1.64. Note that the test is one-sided! (If you wanted the exact distribution, you could check with `R` with `pt(0.12/0.07,lower.tail=F,df=405-5)`, but the estimates are already rounded up; the order of magnitude is what matters here.) The correct answer is "significant" (the test does reject the null, i.e., the coefficient is significant.)