Examiners' commentaries 2018

EC2020 Elements of econometrics

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2017–18. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the subject guide (2016). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

General remarks

Learning outcomes

At the end of the course, and having completed the Essential reading and activities, you should be able to:

- describe and apply the classical regression model and its application to cross-section data
- describe and apply the:
  - Gauss–Markov conditions and other assumptions required in the application of the classical regression model
  - reasons for expecting violations of these assumptions in certain circumstances
  - tests for violations
  - potential remedial measures, including, where appropriate, the use of instrumental variables
- recognise and apply the advantages of logit, probit and similar models over regression analysis when fitting binary choice models
- competently use regression, logit and probit analysis to quantify economic relationships using standard regression programmes (Stata and EViews) in simple applications
- describe and explain the principles underlying the use of maximum likelihood estimation
- apply regression analysis to time-series models using stationary time series, with awareness of some of the econometric problems specific to time series applications (for example, autocorrelation) and remedial measures
- recognise the difficulties that arise in the application of regression analysis to nonstationary time series, know how to test for unit roots, and know what is meant by cointegration.

Common mistakes committed by candidates

A large number of candidates are not able to clearly distinguish between sample variance and covariance, and population variance and covariance (this is happening year after year).

The use of $\text{Cov}(X, Y)$ and $\text{Var}(X)$ should be restricted to describing the population covariance and variances, respectively, with definitions:

$$\text{Cov}(X, Y) = \text{E}\left((X - \text{E}(X))(Y - \text{E}(Y))\right) = \text{E}(XY) - \text{E}(X)\,\text{E}(Y)$$

and:

$$\text{Var}(X) = \text{E}((X - \text{E}(X))^2) = \text{E}(X^2) - (\text{E}(X))^2$$

(you also may denote $\text{Cov}(X, Y) = \sigma_{XY}$ and $\text{Var}(X) = \sigma_X^2$). They are typically unknown, but fixed, quantities.

The sample covariance and variance are estimators of the population covariance and variance, respectively. They are defined as:

$$\text{Sample Cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

and:

$$\text{Sample Var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

(you also may use $\widehat{\sigma}_{XY}$ and $\widehat{\sigma}_X^2$). You can compute them given the data.

With a slight abuse of notation, we often divide by n instead, which is irrelevant if we let n be large. The division by $n - 1$ is a finite sample issue only (unbiasedness).

The sample covariance and variance show up in our definition of the OLS estimator of the slope in the simple linear regression model, not the population covariance and variance, as:

$$\widehat{\beta} = \frac{\sum\limits_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\text{Sample Cov}(X, Y)}{\text{Sample Var}(X)} \neq \frac{\text{Cov}(X, Y)}{\text{Var}(X)}.$$

Treating them as being the same results in incorrect analyses and candidates losing significant marks.

Candidates should realise that $\dfrac{1}{n}\sum\limits_{i=1}^{n} u_i$ is not the same as $\text{E}(u_i)$. So, while we typically assume $\text{E}(u_i) = 0$, this does not guarantee that $\dfrac{1}{n}\sum\limits_{i=1}^{n} u_i = 0$. Also, while we may be happy to assume $\text{E}(x_i u_i) = 0$ (uncorrelatedness between the errors and regressors), this does not guarantee that $\dfrac{1}{n}\sum\limits_{i=1}^{n} x_i u_i = 0$. Note that:

- both $\dfrac{1}{n}\sum\limits_{i=1}^{n} u_i$ and $\dfrac{1}{n}\sum\limits_{i=1}^{n} x_i u_i$ are random variables, which take the value 0 with probability 0 (continuous random variables)!
- $\text{E}(u_i) = 0$ and $\text{E}(x_i u_i) = 0$ are fixed, not stochastic!

The differences between sample and population moments need to come across clearly when looking at unbiasedness and making consistency arguments. In both cases, we first simplify our estimator (plug in the true model) to obtain:

$$\widehat{\beta} = \beta + \frac{\sum\limits_{i=1}^{n}(X_i - \bar{X})u_i}{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2} = \beta + \frac{\sum\limits_{i=1}^{n} x_i u_i}{\sum\limits_{i=1}^{n} x_i^2} \qquad \text{with } x_i = X_i - \bar{X}.$$

- For *unbiasedness*, clearly indicate that you want to show that $\mathrm{E}(\widehat{\beta}) = \beta$. Unbiasedness does not follow from $\sum\limits_{i=1}^{n} x_i u_i = 0$, instead it follows from $\mathrm{E}\left(\frac{\sum\limits_{i=1}^{n} x_i u_i}{\sum\limits_{i=1}^{n} x_i^2}\right) = 0$.

  If we treat x_i as fixed, $\mathrm{E}\left(\frac{\sum\limits_{i=1}^{n} x_i u_i}{\sum\limits_{i=1}^{n} x_i^2}\right) \equiv \mathrm{E}\left(\sum\limits_{i=1}^{n} d_i u_i\right) = \sum\limits_{i=1}^{n} d_i \, \mathrm{E}(u_i)$ and then unbiasedness follows as $\mathrm{E}(u_i) = 0$.
- For *consistency*, clearly indicate that you want to show that $\mathrm{plim}(\widehat{\beta}) = \beta$. Using the plim properties, we show:

$$\begin{aligned}
\mathrm{plim}\ \widehat{\beta} = \beta + \mathrm{plim}\left(\frac{\sum\limits_{i=1}^{n} x_i u_i}{\sum\limits_{i=1}^{n} x_i^2}\right) &= \beta + \frac{\mathrm{plim}\left(\frac{1}{n}\sum\limits_{i=1}^{n} x_i u_i\right)}{\mathrm{plim}\left(\frac{1}{n}\sum\limits_{i=1}^{n} x_i^2\right)} \\
&\equiv \beta + \frac{\mathrm{plim}\,(\text{Sample Cov}\,(x,u))}{\mathrm{plim}\,(\text{Sample Var}\,(x))} \\
&= \beta + \frac{\mathrm{Cov}\,(x,u)}{\mathrm{Var}\,(x)} \qquad \text{using the law of large numbers}
\end{aligned}$$

  where $\mathrm{Cov}(x,u) = 0$ and $\mathrm{Var}(x) > 0$, ensuring we get consistency.
- Remember, the law of large numbers ensures that sample averages converge to their population analogues.

Candidates struggled to give competent answers to the interpretation of empirical results. When interpreting an empirical result you should discuss the significance of the coefficients, magnitude and sign of the coefficients.

When conducting hypothesis tests, you should make sure that the Gauss–Markov conditions hold. The Gauss–Markov conditions have to be explicitly specified. Only writing that the Gauss–Markov conditions hold is not sufficient. As good practice, begin your examination by explicitly providing the Gauss–Markov conditions. You can then refer back to them thereafter. Moreover, ensure when conducting hypothesis testing that you clearly indicate the null and alternative hypotheses (in terms of the true parameters, say β_1), the test statistic (in terms of the parameter estimates, here $\widehat{\beta}_1$), its distribution (with degrees of freedom), the rejection rule (one-sided or-two sided) for a given significance level (typically 5%) with suitable critical values, and provide an interpretation of your result.

Just as last year, many candidates do not answer all parts of the question. Make sure you read the questions properly and provide all details that are requested. Not answering a question will automatically earn you a zero mark for that question.

Key steps to improvement

Essential reading for this course includes the subject guide and the following:

- Dougherty, C. *Introduction to econometrics.* (Oxford: Oxford University Press, 2016) 5th edition [ISBN 9780199676828]; http://oxfordtextbooks.co.uk/orc/dougherty5e/

Apart from the Essential readings you should do some supplementary reading. One very good book at the same level is:

- Gujarati, D.N. and D.C. Porter *Basic econometrics.* (McGraw–Hill, 2009, International edition) 5th edition [ISBN 9780071276252].

To understand the subject clearly it is important to supplement Dougherty's *Introduction to econometrics* (fifth edition) with the subject guide **EC2020 Elements of econometrics** (2016), especially Chapter 10 which covers maximum likelihood estimation. It is very important to carefully go through the subject guide. The subject guide contains solutions to the questions given in the main textbook and also some additional questions and solutions. Working through these will improve your understanding of the subject.

The chapter in the subject guide on maximum likelihood (Chapter 10) includes some additional theory which has not been covered in the main textbook. It is important to read the additional theory given in the subject guide to have a better understanding of the principles of maximum likelihood and tests based on the likelihood function.

Please check the VLE course page for resources for this subject such as a downloadable copy of the subject guide **EC2020 Elements of econometrics** (2016), PowerPoint slideshows that provide a graphical treatment of the topics covered in the textbook, datasets and statistical tables. Candidates should utilise datasets using standard regression programmes (STATA or EViews). This will help in the understanding of the subject.

Examination revision strategy

Many candidates are disappointed to find that their examination performance is poorer than they expected. This may be due to a number of reasons, but one particular failing is '**question spotting**', that is, confining your examination preparation to a few questions and/or topics which have come up in past papers for the course. This can have serious consequences.

We recognise that candidates might not cover all topics in the syllabus in the same depth, but you need to be aware that examiners are free to set questions on **any aspect** of the syllabus. This means that you need to study enough of the syllabus to enable you to answer the required number of examination questions.

The syllabus can be found in the Course information sheet available on the VLE. You should read the syllabus carefully and ensure that you cover sufficient material in preparation for the examination. Examiners will vary the topics and questions from year to year and may well set questions that have not appeared in past papers. Examination papers may legitimately include questions on any topic in the syllabus. So, although past papers can be helpful during your revision, you cannot assume that topics or specific questions that have come up in past examinations will occur again.

If you rely on a question-spotting strategy, it is likely you will find yourself in difficulties when you sit the examination. We strongly advise you not to adopt this strategy.

Examiners' commentaries 2018

EC2020 Elements of econometrics

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2016–17. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the subject guide (2016). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

Comments on specific questions – Zone A

Candidates should answer **EIGHT** of the following **TEN** questions: **ALL** of the questions in Section A (8 marks each) and **THREE** questions from Section B (20 marks each). **Candidates are strongly advised to divide their time accordingly.**

Section A

Answer all questions from this section.

Question 1

We are interested in investigating the factors governing the precision of regression coefficients. Consider the model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

with OLS parameter estimates $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$. Under the Gauss–Markov assumptions, we have

$$\text{Var}\left(\widehat{\beta}_2\right) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r_{X_2X_3}^2},$$

where σ_ε^2 is the variance of ε and $r_{X_2X_3}$ is the sample correlation between X_2 and X_3.

(a) **Provide four factors that help with obtaining more precise parameter estimates for, say, $\widehat{\beta}_2$.**

(4 marks)

(b) **In light of your answer to (a), discuss the concept of near multicollinearity. What consequences does its presence have when considering single and joint significance testing of our slope parameters?**

(4 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 3.3 (Properties of the multiple linear regression coefficients), 3.4 (Multicollinearity) and Chapter 3.5, Relationship between F statistic and t statistic).

Dougherty, C. Subject guide (2016): Chapter 3.

Approaching the question

(a) Candidates should note that the expression of the variance can be rewritten as:

$$\text{Var}(\widehat{\beta}_2) = \frac{\sigma_\varepsilon^2}{n \times MSD(X_2) \times (1 - r^2_{X_2X_3})}$$

where $MSD(X_2) \equiv n^{-1}\sum(X_{2i} - \bar{X}_2)^2$. The four factors that affect the precision then are: n, $MSD(X_2)$, $r^2_{X_2X_3}$ and σ^2_ε. Therefore, to obtain more precise parameter estimates of β_2 it is desirable to have: (i) small error variance, (ii) large sample size, (iii) large sample variability of the regressor X_2, and (iv) small correlation among the regressors X_2 and X_3.

Some candidates gave a discussion of the Gauss–Markov assumptions which is not the answer.

(b) The issue of near multicollinearity is associated with the setting where $r^2_{X_2X_3}$ is close to 1, yielding very imprecise parameter estimates (large variance). Concept and consequences of multicollinearity are standard bookwork.

Question 2

Consider the linear regression model

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + u_t, \quad t = 1, \ldots, T$$

where the errors u_t are distributed independently of the regressors X_t and $|\beta_2| < 1$. You suspect that the, mean zero, errors exhibit autocorrelation.

(a) **Explain what we mean by the concept of autocorrelation.**

(2 marks)

(b) **Assume that u_t follows an AR(1) process.**

i. **Discuss, for the given model, the consequences for the ordinary least squares estimator. Support your answers with suitable arguments.**

(3 marks)

ii. **Discuss how you would detect the presence of autocorrelation in the errors in this model. Clearly indicate the null and alternative hypothesis, the test statistic, and rejection rule.**

(3 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 12.3 (Fitting a model subject to AR(1) autocorrelation, and Chapter 12.2 (Detection of autocorrelation).

Dougherty, C. Subject guide (2016): Chapter 12.

Approaching the question

(a) Candidates should clearly indicate what autocorrelation is (standard bookwork) and indicate that in the presence of the lagged endogenous variable Y_{t-1} this yields inconsistency as $\text{Cov}(Y_{t-1}, u_t) \neq 0$.

(b) i. Specifically, since ε_t is uncorrelated with X_{t-1}, Y_{t-2} and u_{t-1} and since the errors u_t are distributed independently of the regressors.

$$\begin{aligned}\text{Cov}(Y_{t-1}, u_t) &= \text{Cov}(\beta_0 + \beta_1 X_{t-1} + \beta_2 Y_{t-2} + u_{t-1}, \rho u_{t-1} + \varepsilon_t) \\ &= \beta_2 \rho \text{Cov}(Y_{t-2}, u_{t-1}) + \rho \text{Cov}(u_{t-1}, u_{t-1}).\end{aligned}$$

Using the fact that $\text{Cov}(Y_{t-2}, u_{t-1}) = \text{Cov}(Y_{t-1}, u_t)$ by covariance stationarity we get:

$$\text{Cov}(Y_{t-1}, u_t) = (1 - \beta_2 \rho)^{-1} \rho \sigma_u^2 \neq 0.$$

ii. To detect autocorrelation the Breusch–Godfrey test should be proposed, which requires us to run the following auxiliary regression:

$$\widehat{u}_t = \gamma_0 + \gamma_1 X_t + \gamma_2 Y_{t-1} + \rho \widehat{u}_{t-1} + v_t.$$

The use of the Durbin–Watson test is incorrect, a Durbin h-test may be proposed as well. Candidates should clearly indicate H_0 and H_1, the test statistic and the rejection rule.

Question 3

For the population of men who grew up with disadvantaged backgrounds, let *poverty* be a dummy variable equal to one if a man is currently living below the poverty line, and zero otherwise. The variable *age* is age and *educ* is total years of schooling. Let *vocat* be an indicator equal to unity if a man's high school offered vocational training. Using a random sample of 850 men, you obtain

$$\mathbf{Pr}(\boldsymbol{poverty} = \mathbf{1} \widehat{\,|\, \boldsymbol{educ}}, \boldsymbol{age}, \boldsymbol{vocat}) = \boldsymbol{\Lambda}(\mathbf{0.453} - \mathbf{0.016}\boldsymbol{age} - \mathbf{0.087}\boldsymbol{educ} - \mathbf{0.049}\boldsymbol{vocat})$$

where $\boldsymbol{\Lambda(z) = \exp(z) = (1 + \exp(z))}$ is the logit function.

(a) It is argued that using the logit regression model is better than using the linear probability model when explaining the binary variable *poverty.* Discuss the benefits/drawback of using the logit regression model when trying to explain a binary variable.

(5 marks)

(b) For a 40-year old man, with 12 years of education, what is the estimated effect of having vocational training available in high school on the probability of currently living in poverty?

***Hint*: Clarity of computations required is enough, no need to give an exact number.**

(3 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 10.1 (the linear probability model), Chapter 10.2 (logit analysis), and Chapter 10.6 (introduction to maximum likelihood estimation.

Dougherty, C. Subject guide (2016): Chapter 10.

Approaching the question

(a) The logit model has two main advantages over the linear probability model (LPM): predicted probabilities are restricted to lie in $[0, 1]$ and MLE is (asymptotically) efficient whereas OLS (LPM) will be inefficient given the inherent presence of heteroskedasticity.

The main drawbacks of the logit model relative to the linear probability model are that the coefficients cannot be directly interpreted as the marginal effects of the regressor(s) of interest and it is also computationally more complicated.

(b) Candidates will here need to observe that we need to compare predicted probabilities using the logit specification of the probabilities:

$$\Lambda(z) = \frac{1}{1 + \exp(-z)}.$$

For a 40-year old man with 12 years of education with vocational training the estimated probability of living in poverty is given by:

$$\widehat{\Pr(y_i = 1 \,|\, age_i = 40, educ_i} = 12, vocat_i = 1) = \frac{\exp(z_1)}{1 + \exp(z_1)} \approx 0.218$$

where $z_1 = 0.453 - 0.016 \times 40 - 0.087 \times 12 - 0.049 \times 1 \approx -1.28$.

The estimated probability of living in poverty for the same man without the vocational training is given by:

$$\widehat{\Pr(y_i = 1 \,|\, age_i = 40; educ_i} = 12; vocat_i = 0) = \frac{\exp(z_2)}{1 + \exp(z_2)} \approx 0.226$$

where $z_2 = z_1 + 0.049 \times 1 \approx -1.23$.

Therefore, for a 40-year old man with 12 years of education having vocational training in high school decreases the probability of living in poverty by 0.8 percentage points.

Question 4

The following model jointly determines monthly child support payments and monthly visitation rights for divorced couples with children:

$$\boldsymbol{support = \alpha_1 + \alpha_2 visits + \alpha_3 finc + \alpha_4 fremarr + \alpha_5 dist + \varepsilon_1}$$
$$\boldsymbol{visits = \beta_1 + \beta_2 support + \beta_3 mremarr + \beta_4 dist + \varepsilon_2}$$

We assume that children live with their mothers, so that fathers pay child support. Thus, the first equation is the father's 'reaction function': it describes the amount of child support paid for any given level of visitation rights and the other exogenous variables *finc* (father's income), *fremarr* (binary indicator if father remarried), and *dist* (miles currently between the mother and father's residence). Similarly the second equation is the mother's reaction function: it describes visitation rights for a given amount of child support; *mremarr* is a binary indicator for whether the woman is remarried.

(a) **Examine the identification of each structural equation.**

(3 marks)

(b) **Your friend suggests you should implement the IV estimator to estimate the β parameters consistently. He tells you to use *finc* as instrument for *support.* Provide a critical discussion of this suggestion.**

(5 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 8.3 (Instrumental variables), and Chapter 9 (Simultaneous equations estimation).

Dougherty, C. Subject guide (2016): Chapter 9.

Approaching the question

(a) The first equation is *exactly identified* since there is one exogenous variable ($mremarr_i$) available as instrument to the endogenous regressor ($visits_i$).

The second equation is *overidentified* since there are two exogenous variables ($finc_i$ and $fremarr_i$) available as instruments to the endogenous regressor ($support_i$).

Let G denote the number of equations in the system of simultaneous equations. It is true that in both equations there are $G-1$ endogenous variables (but you may be given a setting where there are fewer endogenous variables), and in each case you have excluded enough exogenous variables (k) you can use as instruments for these bad variables. Answers tend to be very vague.

(b) The IV approach suggested by the friend would yield consistent estimates for β parameters since $finc_i$ is exogenous and correlated with support (assuming $\alpha_3 \neq 0$). However, given this equation is overidentified, we could obtain more efficient estimates by using a two-stage least squares approach in which both $fremarr_i$ and $finc_i$ are included in the vector of instruments.

Question 5

Consider the simple linear regression model

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + u_i}$$

under the classical linear regression model assumptions, where $\boldsymbol{X_i}$ is fixed under repeated sampling. The usual OLS estimators $\boldsymbol{\widehat{\beta}_0}$ and $\boldsymbol{\widehat{\beta}_1}$ are unbiased for their respective population parameters. Let $\boldsymbol{\tilde{\beta}_1}$ be the estimator of $\boldsymbol{\beta_1}$ obtained by assuming the intercept is zero.

(a) Show that the restricted least squares estimator of $\boldsymbol{\beta_1}$ is given by

$$\boldsymbol{\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}}.$$

(4 marks)

(b) Find E($\boldsymbol{\tilde{\beta}_1}$) in terms of the $\boldsymbol{X_i}$, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$. Verify that $\boldsymbol{\tilde{\beta}_1}$ is unbiased for $\boldsymbol{\beta_1}$ when the population intercept is zero. Are there other cases where $\boldsymbol{\tilde{\beta}_1}$ is unbiased?

(4 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 1.3 (Derivation of the regression coefficients) and Chapter 2.3 (The random components and unbiasedness of the OLS regression coefficients).

Dougherty, C. Subject guide (2016): Chapters 1 and 2.

Approaching the question

(a) Formally, restricted least squares estimates of β_0 and β_1 solve the following problem:

$$(\tilde{\beta}_0, \tilde{\beta}_1) = \min_{b_0,b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2, \quad \text{subject to } b_0 = 0.$$

This is the same as performing OLS on the model where $\beta_0 = 0$, that is performing OLS while leaving out the intercept.

Therefore, $\tilde{\beta}_0 = 0$ and $\tilde{\beta}_1 : \min_{b_1} \sum_{i=1}^{n} (Y_i - b_1 X_i)^2$. The first-order condition is given by:

$$-2\sum_{i=1}^{n} (Y_i - \tilde{\beta}_1 X_i) X_i = 0 \qquad \Leftrightarrow \qquad \tilde{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

(b) While candidates found the discussion of restricted least squares difficult, there was no reason not to answer the second part which was standard. Plug in the true model and use properties of sums to obtain:

$$\begin{aligned}\tilde{\beta}_1 &= \left(\sum_{i=1}^{n} X_i^2\right)^{-1} \left(\sum_{i=1}^{n} X_i(\beta_0 + \beta_1 X_i + u_i)\right)\\ &= \left(\sum_{i=1}^{n} X_i^2\right)^{-1} \left(\sum_{i=1}^{n} X_i\right)\beta_0 + \beta_1 + \left(\sum_{i=1}^{n} X_i^2\right)^{-1} \left(\sum_{i=1}^{n} X_i u_i\right).\end{aligned}$$

Taking expectations using the fact that X_is are fixed and $\mathrm{E}(u_i) = 0$, we have:

$$\begin{aligned}\mathrm{E}(\tilde{\beta}_1) &= \left(\sum_{i=1}^{n} X_i^2\right)^{-1} \left(\sum_{i=1}^{n} X_i\right)\beta_0 + \beta_1 + \left(\sum_{i=1}^{n} X_i^2\right)^{-1} \left(\sum_{i=1}^{n} X_i \mathrm{E}(u_i)\right)\\ &= \left(\sum_{i=1}^{n} X_i^2\right)^{-1} \left(\sum_{i=1}^{n} X_i\right)\beta_0 + \beta_1.\end{aligned}$$

Therefore, $\tilde{\beta}_1$ is unbiased if either: (i) $\beta_0 = 0$ or (ii) $\sum_{i=1}^{n} X_i = 0$.

Section B

Answer three questions from this section.

Question 6

Let us consider the estimation of a hedonic price function for houses. The hedonic price refers to the implicit price of a house given certain attributes (e.g., the number of bedrooms). The data contains the sale price of 546 houses sold in the summer of 1987 in Canada along with their important features. The following characteristics are available: the lot size of the property in square feet (*lotsize*), the numbers of bedrooms (*bedrooms*), the number of full bathrooms (*bathrooms*), and a dummy indicating the presence of airconditioning (*airco*).

Consider the following ordinary least squares results

$$\begin{aligned}\widehat{\log(price)}_i = \underset{\substack{(.232)\\ [.233]}}{7.094} + \underset{\substack{(.028)\\ [.028]}}{0.400} \log(lotsize)_i + \underset{\substack{(.015)\\ [.017]}}{0.078} bedrooms_i + \qquad (6.1)\\ \underset{\substack{(.023)\\ [.024]}}{0.216} bathrooms_i + \underset{\substack{(.024)\\ [.023]}}{0.212} airco_i \quad n = 546, \; RSS = 32.622\end{aligned}$$

The usual standard errors are in parentheses, the heteroskedasticity robust standard errors are in square brackets, and RSS measures the residual sum of squares.

(a) **Interpret the parameter estimates on log($lotsize$), $bedrooms$, and $airco$. Briefly discuss the statistical significance of the results.**

(5 marks)

(b) **Suppose that lot size was measured in square metres rather than square feet. How would this affect the parameter estimates of the slopes and intercept? How would this affect the fitted values? *Note*: the conversion (approximate) $1m^2 = 10ft^2$.**

(5 marks)

(c) **We are interested in testing the hypothesis $H_0 : \beta_{bedrooms} = \beta_{bathrooms}$ against the alternative $H_1 : \beta_{bedrooms} \neq \beta_{bathrooms}$. Discuss a test for this hypothesis that makes use of the following restricted regression result**

$$\log\widehat{(price)}_i = \underset{(.234)}{6.994} + \underset{(.282)}{0.408}\log(lotsize)_i + \underset{(.011)}{0.127}bbrooms_i + \underset{(.024)}{0.215}airco_i \qquad (6.2)$$

$$n = 546, \;\; RSS = 33.758$$

where $bbrooms = bedrooms + bathrooms$. Clearly indicate the assumptions you are making for this test to be valid.

(5 marks)

(d) **You are interested in testing for the presence of heteroskedasticity. Say you are told that the variance is increasing with log($lotsize$). Discuss how you would test for the presence of heteroskedasticity. What is the name of the test you are proposing?**

(5 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 1.4 (Interpretation of a regression equation – units of measurement), Chapter 4.2 (Logarithmic transformations), Chapter 2.6 (Testing hypotheses relating to the regression coefficients), Chapter 7.1 (Heteroskedasticity and its implications), and Chapter 7.2 (Detection of heteroskedasticity).

Dougherty, C. Subject guide (2016): Chapter 7.

Approaching the question

(a) Clear discussion of interpretation required (units not always clear): On average, holding the remaining variables in the regression constant, (i) a 1% increase in lot size is associated with a 0.4% increase in house price, (ii) each extra bedroom is associated with a 7.8% increase in house price, and (iii) houses with air conditioning are 21.2% more expensive than those without. All estimates are statistically significant at 5% significance levels. Candidates should clearly indicate H_0 and H_1, the test statistic and the rejection rule.

(b) Let $lotsize_i$ be the lot size in square feet and $\widetilde{lotsize}_i$ be the lot size in square metres. We have that $\widetilde{lotsize}_i = (10)^{-1} lotsize_i$ and $\log(\widetilde{lotsize}_i) = \log((10)^{-1}) + \log(lotsize_i)$. Since this is an additive transformation of one of the explanatory variables we have that (i) the regression slopes will not be affected, (ii) the intercept will change to $7.094 - \log((10)^{-1}) \times 0.4$, and (iii) the fitted values will also not be affected. Many candidates made an error here, ignoring the fact that the variable whose measurement was changed entered in log form.

(c) Candidates would have to recognise that (6.2) is a restricted version of (6.1) where $\beta_{bedrooms} = \beta_{bathrooms}$ is imposed. We therefore need to use the F test. Test statistic:

$$F = \frac{RRSS - URSS}{URSS} \times \frac{n-K}{J} = \frac{33.758 - 32.622}{32.622} \times \frac{541}{1} \approx 18.84.$$

Assuming the Gauss–Markov assumptions plus normality of the error term hold: under H_0, $F \sim F_{1,541}$. At the 5% signicance level we reject H_0 since $F > 3.86$. Conclusion: The effect of one extra bathroom is different from the effect of one extra bedroom. Some candidates did not recognise this and were proposing a test on the coefficient of *bbrooms* equalling zero which is wrong.

(d) Assuming Gauss–Markov assumptions plus the normality of the error hold, he can use the Goldfeld–Quandt test for heteroskedasticity. For that purpose, we should first order the 546 observations by the magnitude of $\log(lotsize_i)$. Fit one regression for the first n^* observations and another for the last n^* observations (usually n^* equals one-third of the sample). Let RSS_1 and RSS_2 denote the sum of squared residuals in each of these regressions, respectively.

- $\mathrm{H}_0 : \sigma_2^2 = \sigma_1^2$ vs. $\mathrm{H}_1 : \sigma_2^2 > \sigma_1^2$.
- Test statistic: $GQ = RSS_2/RSS_1$.
- Under H_0, $GQ \sim F_{n^*-k,\, n^*-k}$.
- Reject if GQ is greater than the 95th percentile of the F distribution above.

Candidates need to be careful not to state $\mathrm{H}_0 : RSS_2 = RSS_1$ vs. $\mathrm{H}_1 : RSS_2 > RSS_1$. Both RSS_1 and RSS_2 are random variables. Because the sample sizes are identical it is also correct to state H_0 : RSS_1 and RSS_2 are not statistically different. Note that simply by having one sample larger than the other, you could also have a larger residual sum of squares.

Question 7

The following question concerns the effects of background characteristics and admission assessment scores on the performance of students in the final university examinations in a UK university. The following equation was estimated by Ordinary Least Squares:

$$\begin{aligned}\widehat{finalavg} = \underset{(2.78)}{53.89} + \underset{(.04)}{0.03}tst\_reas + \underset{(.02)}{0.05}tst\_quan + \underset{(.02)}{0.06}interview \\ -\underset{(.78)}{0.04}indep + \underset{(.66)}{0.67}male + \underset{(.97)}{0.06}indep{*}male \\ n = 325,\ R^2 = .06, \qquad (7.1)\end{aligned}$$

where *finalavg* is the average finals score (the outcome), *tst_reas* and *tst_quan* are the pre-admission reasoning and quantitative test scores respectively, *interview* is the pre-admission interview score, *indep* indicates whether the student attended an independent school (1 = yes, 0 = no), and *male* indicates whether the student is male (1 = yes, 0 = no). The usual standard errors are in parentheses.

(a) We want to test whether gender has a significant impact on students' finals performance. Clearly indicating the null and the alternative hypothesis, provide the test statistic and the rejection rule. Discuss what information you would need to enable you to implement this test. You are expected to provide the assumptions which underlie your test.

(5 marks)

(b) If we do not include the interaction term *indep*male* in our regression model, what are we implicitly assuming about the effect of gender and school background on finals performance?

(5 marks)

(c) Suppose students who did not attend an independent school, attended a state school. Using the results in (7.1), provide the parameter estimates you would obtain if you had applied Ordinary Least Squares to the equation

$$\begin{aligned}finalavg = \beta_0 + \beta_1 tst\_reas + \beta_2 tst\_quan + \beta_3 interview \qquad (7.2) \\ \beta_4 state + \beta_5 male + \beta_6 state{*}male + \varepsilon,\end{aligned}$$

where *state* indicates whether the student attended a state school (1 = yes, 0 = no).

(5 marks)

(d) **Discuss any problem you may have in estimating the model if all males in your sample have attended an independent school prior to attending university. What name does this problem have and what can you do to mitigate this problem?**

(5 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 5.1–5.3 (Dummy variables).

Dougherty, C. Subject guide (2016): Chapter 5.

Approaching the question

(a) For given values of the remaining explanatory variables, the gender performance gap measured as the difference between male and female average final scores is given by β_5 if $indep_i = 0$ or $\beta_5 + \beta_6$ if $indep_i = 1$. Therefore, we want to perform the following test:

- $\text{H}_0 : \beta_5 = 0$ and $\beta_6 = 0$ vs. $\text{H}_1 : \beta_5 \neq 0$ or $\beta_6 \neq 0$.
- Need the R^2 of a regression of *finalavg* on a constant, *tst_reas*, *tst_quant*, *interview* and *indep* (the restricted model).
- Compute the test statistic:

$$F = \frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} \times \frac{n - K}{J} = \frac{0.06 - R^2_r}{1 - 0.06} \times \frac{325 - 7}{2}.$$

- Under Gauss–Markov assumptions plus normality, $F \sim F_{2,\,318}$.
- At the 5% signicance level, reject H_0 if $F > 3.07$.

(b) If we do not include the interaction term we are implicitly assuming that the effect of gender on performance is the same in independent and non-independent schools. We are also assuming that the effect of attending an independent school on final performance is the same for both male and female students.

(c) Let $\widehat{\theta}$s indicate the estimated parameters from the initial regression. From the original regression we have:

$$\begin{aligned}
\widehat{y}_i &= \widehat{\theta}_0 + \cdots + \widehat{\theta}_4 indep_i + \widehat{\theta}_5 male_i + \widehat{\theta}_6 indep_i \times male_i \\
&= \widehat{\theta}_0 + \cdots + \widehat{\theta}_4 (1 - state_i) + \widehat{\theta}_5 male_i + \widehat{\theta}_6 (1 - state_i) \times male_i \\
&= (\widehat{\theta}_0 + \widehat{\theta}_4) + \cdots - \widehat{\theta}_4 state_i + (\widehat{\theta}_5 + \widehat{\theta}_6) male_i - \widehat{\theta}_6 state_i \times male_i
\end{aligned}$$

where $\cdots = \widehat{\theta}_1 tst\_reas_i + \widehat{\theta}_2 tst\_quant_i + \widehat{\theta}_3 interview_i$. Mapping original estimates into estimated parameters in (7.2) yields:

- $\widehat{\beta}_0 = \widehat{\theta}_0 + \widehat{\theta}_4 = 53.89 - 0.04 = 53.85$
- $\widehat{\beta}_4 = -\widehat{\theta}_4 = -(-0.04) = 0.04$
- $\widehat{\beta}_6 = -\widehat{\theta}_6 = -0.06$
- $\widehat{\beta}_5 = \widehat{\theta}_5 + \widehat{\theta}_6 = 0.67 + 0.06 = 0.73$
- $\widehat{\beta}_2$, $\widehat{\beta}_3$ and $\widehat{\beta}_4$ remain the same from (7.1).

(d) If all males attended an independent school, then $male_i = male_i \times indep_i \; \forall \; i$, so the variables will be perfectly collinear and he would not be able to obtain the OLS estimates from (7.1). In this case, we could either drop the interaction term from the regression or obtain more data such that some men that have not attended independent school will also be in the sample.

Question 8

An OLS regression of y_t on x_t and x_{t-1} gives the following results (with the standard errors given in parentheses)

$$\hat{y}_t = \underset{(2.30)}{8.88} + \underset{(3.01)}{5.07}\,x_t - \underset{(3.01)}{3.18}\,x_{t-1}; \quad R^2 = .095,\ T = 209 \qquad (8.1)$$

(a) **What are the estimates of the short-run and long-run effect of x_t on y_t? Interpret these estimates.**

(4 marks)

(b) **Test the hypothesis that a one unit increase in x results in a ten unit increase in y in the same year. Under what assumptions is this test valid?**

(4 marks)

Let e_t be the OLS residuals from the above regression. An OLS regression of e_t on e_{t-1} yields

$$e_t = \underset{(2.12)}{0.55} + \underset{(0.18)}{0.44}\,e_{t-1} + \underset{(2.18)}{2.16}\,x_t - \underset{(.99)}{1.09}x_{t-1}; \quad R^2 = .175,\ T = 208 \qquad (8.2)$$

(c) **Using this result, test for evidence of autocorrelation, clearly indicating the null and alternative hypotheses, the test statistic, rejection rule and assumptions underlying the test. What name do we give this test?**

(5 marks)

(d) **You are interested in testing whether the long-run effect of x_t on y_t is statistically significant.**

i. **Discuss how to reparameterise (8.1) to ensure that your regression output will provide you with a standard error for the long-run effect.**

(4 marks)

ii. **Discuss the problem of implementing your test using the standard error obtained in (d)i. when you do find evidence of autocorrelation in (8.1). Briefly indicate how you proceed with your test.**

(3 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 6.5 (Testing a linear restriction), Chapter 11.3 (Models with lagged explanatory variables), and Chapter 12.1–12.3 (Definition, consequences and detection of autocorrelation; Fitting a model subject to AR(1) autocorrelation).

Dougherty, C. Subject guide (2016): Chapters 6 and 12.

Approaching the question

(a) Short-run effect: It shows the immediate effect a change in x with 1 unit has on y. Increases with 5.07 units.

Long-run effect: It shows the effect a permanent change in x with one unit has on y after 1 period (the last lagged response) has passed. Increases with $5.07 - 3.18 = 1.89$ units.

(b) Asked to perform the t test for $\text{H}_0 : \beta_1 = 10$ vs. $\text{H}_1 : \beta_1 \neq 10$ (standard bookwork). Validity of Gauss–Markov + normality.

(c) The test equation tells us that we have to perform the Breusch–Godfrey test for first-order autocorrelation. Let ρ denote the coefficient associated with e_{t-1}.

- $\text{H}_0 : \rho = 0$ (No autocorrelation) vs. $\text{H}_1 : \rho \neq 0$ (Autocorrelation).
- Test statistic: $LM = nR^2 = 28 \times 0.175 \approx 4.9$.

- Under H_0, $LM \overset{a}{\sim} \chi_1^2$.
- Reject H_0 in favour of autocorrelation since $LM > 3.84$.

(d) i. We need to reparameterise (8.1) to obtain the standard error of $\widehat{\beta}_1 + \widehat{\beta}_2$ directly:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + u_t$$

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t - \beta_2 x_t + \beta_2 x_{t-1} + u_t$$

$$y_t = \beta_0 + (\beta_1 + \beta_2) x_t - \beta_2 \Delta x_t + u_t.$$

Regress y_t on a constant, x_t and Δx_t and the standard error of the estimated coefficient on x_t is the standard error of the long-run effect.

ii. The usual standard errors will be invalidated so we need to use HAC standard errors instead. Given the robust standard errors, one could compute the test statistic robust to autocorrelation as $t^{robust} = \widehat{\theta}/se(\widehat{\theta})^{HAC}$ and conclude the long-run effect is statistically significant at the 5% significance level if $|t^{robust}| > 1.96$.

Question 9

Consider the model

$$y_t = \alpha + \beta t + \varepsilon_t, \quad t = 1, \ldots, T \qquad (9.1)$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t, \quad \text{and}$$

v_t is an i.i.d. (0, σ^2) innovation which is independent of the past. Let $|\rho| \leq 1$.

(a) What name do we give the ε_t process given above? Provide the condition(s) that ensures that ε_t is stationary. In your answer discuss what we mean by the concept of stationarity (more precisely 'covariance stationarity').

(4 marks)

(b) It will be important to distinguish between the above process for y_t being 'trend stationary' as opposed to 'difference stationary'.

i. Explain these concepts clearly. Why is it important to distinguish between these two types of non-stationarity?

(4 marks)

ii. Show that under the condition you provided in (a) that y_t is trend stationary.

(2 marks)

iii. Show that if ε_t is difference stationary then y_t is difference stationary.

(2 marks)

(c) Show that you can rewrite the above model in the following form

$$\Delta y_t = \gamma_1 + \gamma_2 t + \gamma_3 y_{t-1} + v_t. \qquad (9.2)$$

Clearly indicate the relation between $(\gamma_1, \gamma_2, \gamma_3)$ and (α, β, ρ).

(4 marks)

(d) What problem do you see here with using (9.2) to conducting the Dickey–Fuller Test to distinguish between trend and difference stationarity when v_t exhibits autocorrelation? What solution do you suggest we adopt?

(4 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 12.1 (Definition and consequences of autocorrelation), Chapter 13.1 (Stationarity and nonstationarity), and Chapter 13.4–13.5 (Tests of nonstationarity).

Dougherty, C. Subject guide (2016): Chapter 13.

Approaching the question

(a) Standard bookwork.

(b) i. A process $\{y_t\}_{t=-\infty}^{\infty}$ is said to be difference stationary if its first-difference is stationary ($\Delta y_t \sim I(0)$). It is said to be trend stationary if the process $y_t - \beta t$ is stationary. It is important to distinguish between the two because the source of non-stationarity has different implications on how we proceed to obtain a stationary time-series to use for regression analysis and for statistical inference.

ii. If $|\rho| < 1$ then ε_t is stationary, since $y_t - \beta t = \alpha + \varepsilon_t$ is also stationary. Therefore, y_t is trend stationary.

iii. Suppose that ε_t is difference stationary ($\Delta\varepsilon_t \sim I(0)$). Taking first differences of (9.1) we obtain $\Delta y_t = \beta + \Delta\varepsilon_t$, since β is just a constant we have that $\Delta y_t \sim I(0)$, that is, y_t is difference stationary.

(c) Lagging (9.1) by one period and multiplying both sides by ρ we obtain:

$$\rho y_{t-1} = \rho\alpha + \rho\beta(t-1) + \rho\varepsilon_{t-1}.$$

Subtracting the above from (9.1) yields:

$$y_t - \rho y_{t-1} = (1-\rho)\alpha + \beta t - \rho\beta t + \rho\beta + \varepsilon_t - \rho\varepsilon_{t-1}.$$

Let $v_t \equiv \varepsilon_t - \rho\varepsilon_{t-1}$ and rearrange:

$$y_t = ((1-\rho)\alpha + \rho\beta) + \beta(1-\rho)t + \rho y_{t-1} + v_t.$$

Finally, subtract y_{t-1} on both sides to obtain:

$$\Delta y_t = \underbrace{(1-\rho)\alpha + \rho\beta}_{\gamma_1} + \underbrace{\beta(1-\rho)t}_{\gamma_2} + \underbrace{\rho}_{\gamma_3} y_{t-1} + v_t.$$

(d) If v_t exhibits autocorrelation it must be eliminated before running the test regression otherwise the Dickey–Fuller test will not be valid. To eliminate it we should include the lags of Δy_t in the test equation (9.2). This test is then known as the augmented Dickey–Fuller test.

Question 10

Let $math10$ denote the percentage of students at a high school receiving a passing score on a standardised math test. We are interested in estimating the effect of per student spending on math performance. A simple model is

$$math10_i = \beta_0 + \beta_1 \log(expend_i) + \beta_2 \log(enroll_i) + \beta_3 poverty_i + u_i \qquad (10.1)$$

where, for each high school i; $poverty_i$ is the percentage of students living in poverty, $expend_i$ is the spending per student and $enroll_i$ the number of registered students. You may assume that this model satisfies all Gauss–Markov assumptions.

You are faced with the fact that data is unavailable on a key variable: $poverty$.

(a) Discuss the properties (unbiasedness and consistency) of the estimators when you drop the variable poverty. Explain your answers.

(5 marks)

You do have information available on a closely related variable: the percentage of students eligible for the federally funded school lunch program, $lnchprg_i$. Let us consider using $lnchprg_i$ as a proxy for $poverty_i$.

(b) Briefly discuss why $lnchprg_i$ is a sensible proxy variable for the unobserved variable $poverty_i$.

(2 marks)

(c) It is unlikely that $lnchprg_i$ is an ideal proxy, in the sense that there is an exact linear relationship between them, instead, we will assume that

$$poverty_i = \alpha_0 + \alpha_1 lnchprg_i + v_i, \quad \alpha_1 \neq 0 \qquad (10.2)$$

Discuss the assumptions you need to make to enable consistent parameter estimators of β_1 and β_2 using your estimable equation

$$math10_i = \gamma_0 + \gamma_1 \log(expend_i) + \gamma_2 \log(enroll_i) + \gamma_3 lnchprg_i + e_i,$$

***Hint*: Consider the relation between the γ and the β parameters and express e_i in terms of u_i and v_i.**

(5 marks)

(d) The OLS results with and without $lnchprg_i$ as an explanatory variable are given by (standard errors in parentheses):

$$\widehat{math}10_i = -\underset{(26.72)}{69.24} + \underset{(3.30)}{11.13} \log expend_i + \underset{(0.615)}{0.022} \log emroll_i$$
$$N = 428, \; R^2 = 0.0297$$

$$\widehat{math}10_i = -\underset{(24.99)}{23.14} + \underset{(3.04)}{7.75} \log expend_i - \underset{(0.58)}{1.26} \log emroll_i - \underset{(0.036)}{0.324} lnchprg_i$$
$$N = 428, \; R^2 = 0.1893$$

i. Interpret the coefficient on $lnchprg$. What does this parameter tell us regarding the parameter of interest β_3?

(4 marks)

ii. Give an intuitive discussion explaining why the effect of expenditures on $math10_i$ is lower in the regression where $lnchprg_i$ is included than where it is excluded.

(4 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 6.2 (The effect of omitting a variable that ought to be included), Chapter 6.4 (Proxy variables).

Dougherty, C. Subject guide (2016): Chapter 6.

Approaching the question

(a) Consider rewriting (10.1) as:

$$math10_i = \beta_0 + \beta_1 \log(expend_i) + \beta_2 \log(enroll_i) + \varepsilon_i$$

where $\varepsilon_i = \beta_3 poverty_i + u_i$. Assuming $\beta_3 \neq 0$, if poverty is correlated with either (log) expenditures and/or (log) enrollment the model will suffer from endogeneity due to omitted variables and OLS will be biased and inconsistent. That is likely to be the case since schools that have smaller expenditures tend to be located in poorer neighbourhoods and hence to have more students living in poverty conditions.

(b) It is a sensible proxy because it is likely to be correlated with poverty and to capture some of the effect of poverty since usually students eligible for the lunch program tend to be those with low levels of family income.

(c) Plug in (10.2) for poverty in (10.1) to obtain:

$$math10_i = \underbrace{(\beta_0 + \beta_3\alpha_0)}_{\gamma_0} + \underbrace{\beta_1}_{\gamma_1} \log(expend_i) + \underbrace{\beta_2}_{\gamma_2} \log(enroll_i) + \underbrace{\beta_3\alpha_1}_{\gamma_3} lnchprg_i + \underbrace{(\beta_3 v_i + u_i)}_{\varepsilon_i}.$$

Given that the model (10.1) satisfies all the Gauss–Markov assumptions, to obtain consistent estimates of β_1 and β_2 we need that: (i) v_i is uncorrelated with $\log(expend_i)$, $\log(enroll_i)$ and $lnchprg_i$, and (ii) $lnchprg_i$ is uncorrelated with u_i.

(d) i. On average, holding expenditure and enrollment constant, a 1 percentage point increase in the number of students eligible for the lunch program is associated with a 0.324 percentage point fall in the percentage of students receiving a passing score in the standardised math test. Since $\gamma_3 = \alpha_1\beta_3$ and assuming $\alpha_1 > 0$, the direction of the effect (sign) of poverty of $math10$ is the same as the effect of $lnchprg$ on $math10$, in this case, with a negative coefficient on $lnchprg$ we can infer that poverty has a negative effect on $math10$.

ii. Omitting relevant variables will result in the remaining parameters attempting to pick up its effect through the correlation these omitted variables have with the included regressors. We, therefore, expect the effect to be smaller, as part of the effect we attribute to expenditure in the short regression is actually coming from the fact that high schools that have larger expenditures tend to have fewer students eligible for the lunch program and those students tend to perform worse in the standardised math test.

Examiners' commentaries 2018

EC2020 Elements of econometrics

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2017–18. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the subject guide (2016). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

Comments on specific questions – Zone B

Candidates should answer **EIGHT** of the following **TEN** questions: **ALL** of the questions in Section A (8 marks each) and **THREE** questions from Section B (20 marks each). **Candidates are strongly advised to divide their time accordingly.**

Section A

Answer all questions from this section.

Question 1

Consider the consumption function

$$C_t = \alpha + \lambda Y_t + \varepsilon_t \tag{1.1}$$

where C_t is aggregate consumption at t, λ is marginal propensity to consume ($0 < \lambda < 1$) and Y_t is aggregate income at t defined as

$$Y_t = C_t + A_t,$$

where A_t is the sum of investment and government consumption at t. Assume that A_t is uncorrelated with ε_t and that the shock ε_t is mean zero i.i.d. across t. A random sample of size n containing Y_t, C_t and A_t is available.

(a) **Provide the reduced form equation for Y_t.**

(2 marks)

(b) **Show that the OLS estimator of λ in (1.1) is inconsistent. You are asked to indicate the direction of this inconsistency.**

***Note*: you are not expected to derive the OLS estimator.**

(6 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 9.1 (Structural and reduced form equations) and Chapter 9.2 (Simultaneous equation bias).

Dougherty, C. Subject guide (2016): Chapter 9.

Approaching the question

(a) Plugging in the consumption function into the aggregate income identity:

$$Y_t = (\alpha + \lambda Y_t + \varepsilon_t) + A_t.$$

Rearranging yields, $Y_t = (1-\lambda)^{-1}\alpha + (1-\lambda)^{-1}A_t + (1-\lambda)^{-1}\varepsilon_t$.

(b) The OLS estimator of λ is given by:

$$\widehat{\lambda} = \frac{\sum_{t=1}^{T}(C_t - \bar{C})(Y_t - \bar{Y})}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2} = \frac{T^{-1}\sum_{t=1}^{T}(C_t - \bar{C})(Y_t - \bar{Y})}{T^{-1}\sum_{t=1}^{T}(Y_t - \bar{Y})^2}.$$

Taking probability limits and using a suitable LLN:

$$\text{plim}(\widehat{\lambda}) = \frac{\text{Cov}(C_t, Y_t)}{\text{Var}(Y_t)} = \frac{\text{Cov}(\alpha + \lambda Y_t + \varepsilon_t, Y_t)}{\text{Var}(Y_t)} = \lambda + \frac{\text{Cov}(\varepsilon_t, Y_t)}{\text{Var}(Y_t)}.$$

Using the reduced form from part (a) we can show OLS is inconsistent since:

$$\text{Cov}(\varepsilon_t, Y_t) = (1-\lambda)^{-1}\text{Cov}(\varepsilon_t, \alpha + A_t + \varepsilon_t) = (1-\lambda)^{-1}\text{Var}(\varepsilon_t) \neq 0$$

where the second equality uses that A_t is uncorrelated with ε_t. Finally, since $\lambda \in (0,1)$ and both $\text{Var}(Y_t)$ and $\text{Var}(\varepsilon_t)$ are positive, there is a positive inconsistency in OLS estimator, in other words, propensity to consume is overestimated in (1.1).

Question 2

Consider the simple linear regression model

$$\boldsymbol{Y_t = \beta X_t + u_t, \qquad t = 1, \ldots, T}$$

where the errors u_t are distributed independently of the regressors X_t. You suspect that the, mean zero, errors exhibit autocorrelation.

(a) **Explain what we mean by the concept of autocorrelation.**

(2 marks)

(b) **Assume you are told that u_t follows an MA(1) process.**

i. **Discuss whether the OLS estimator $\widehat{\beta}$ is a consistent estimator for β. Justify your answers with suitable technical derivations.**

***Note*: you are not expected to derive the OLS estimator.**

(3 marks)

ii. Suppose you want to test $\mathbf{H_0} : \beta = 1$ against $\mathbf{H_1} : \beta < 1$. Discuss how you would conduct this test based on the OLS estimator, recognising the presence of autocorrelation in the error.

(3 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 6.5 (Testing a linear restriction), Chapter 12.1–12.3 (Definition, consequences and detection of autocorrelation; Fitting a model subject to AR(1) autocorrelation).

Dougherty, C. Subject guide (2016): Chapters 6 and 12.

Approaching the question

(a) Candidates should indicate clearly what autocorrelation is (standard bookwork).

(b) i. The OLS estimator is given by:

$$\widehat{\beta} = \frac{\sum_{t=1}^{T} X_t Y_t}{\sum_{t=1}^{T} X_t^2} = \frac{T^{-1} \sum_{t=1}^{T} X_t Y_t}{T^{-1} \sum_{t=1}^{T} X_t^2}.$$

To analyse consistency we take probability limits and apply a LLN to obtain:

$$\text{plim}(\widehat{\beta}) = \frac{\text{E}(X_t Y_t)}{\text{E}(X_t^2)} = \frac{\text{E}(X_t(\beta X_t + u_t))}{\text{E}(X_t^2)} = \beta + \frac{\text{E}(X_t u_t)}{\text{E}(X_t^2)}.$$

The OLS estimator is consistent as long as $\text{E}(X_t u_t) = 0$ (and $\text{E}(X_t^2) \neq 0$). This is satisfied since the zero mean errors u_t are assumed to be independent (and hence uncorrelated) of the regressors X_t.

ii. Recognising the presence of autocorrelation in the error when testing the single linear restriction we need to make use of HAC standard errors. This fact was ignored by many. The test statistic we use is $t = (\widehat{\beta} - 1)/se(\widehat{\beta})^{HAC}$ and we should reject H_0 when t is too small. Under H_0, $t \overset{a}{\sim} N(0,\, 1)$ (also accepted would be $t \sim t_{n-k}$). At the 5% significance level we reject H_0 if $t < -1.645$.

Question 3

Consider the simple linear regression model

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + u_i}$$

under the classical linear regression model assumptions, where X_i is fixed under repeated sampling. The usual OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased for their respective population parameters. Let $\tilde{\beta}_0$ be the estimator of β_0 when β_1 equals 1.

(a) Show that the restricted least squares estimator of β_0 is given by

$$\boldsymbol{\tilde{\beta}_0 = \bar{Y} - \bar{X}}$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(4 marks)

(b) Find $\text{E}(\tilde{\beta}_0)$ in terms of the X_i, β_0 and β_1. Verify that $\tilde{\beta}_0$ is unbiased for β_0 when $\beta_1 = 1$. Are there other cases where $\tilde{\beta}_0$ is unbiased?

(4 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 1.3 (Derivation of the regression coefficients) and Chapter 2.3 (The random components and unbiasedness of the OLS regression coefficients).

Dougherty, C. Subject guide (2016): Chapters 1 and 2.

Approaching the question

(a) Formally, restricted least squares estimates of β_0 and β_1 solve the following problem:

$$(\tilde{\beta}_0, \tilde{\beta}_1) = \min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2, \quad \text{subject to } b_1 = 1.$$

This is the same as performing OLS on the model where $\beta_1 = 1$.

Therefore, $\tilde{\beta}_1 = 1$ and $\tilde{\beta}_0 = \min_{b_0} \sum_{i=1}^{n} (Y_i - b_1 X_i)^2$. The first-order condition is given by:

$$-2 \sum_{i=1}^{n} (Y_i - \tilde{\beta}_0 - X_i) = 0 \qquad \Leftrightarrow \qquad \tilde{\beta}_0 = \bar{Y} - \bar{X}.$$

(b) While candidates found the discussion of restricted least squares difficult, there was no reason not to answer the second part which was standard. Using our estimator, notice that averaging the true model yields $\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u}$. Plugging in our estimator $\tilde{\beta}_0$, we have:

$$\tilde{\beta}_0 = (\beta_0 + \beta_1 \bar{X} + \bar{u}) - \bar{X} = \beta_0 + (\beta_1 - 1)\bar{X} + \bar{u}.$$

Taking expectations using the fact that X_is are fixed and $\mathrm{E}(u_i) = 0$, we have:

$$\begin{aligned} E(\tilde{\beta}_0) &= \beta_0 + (\beta_1 - 1)\bar{X} + n^{-1} \sum_{i=1}^{n} \mathrm{E}(u_i) \\ &= \beta_0 + (\beta_1 - 1)\bar{X}. \end{aligned}$$

Therefore, $\tilde{\beta}_0$ is unbiased if either: (i) $\beta_1 = 1$, or (ii) $\bar{X} = 0$.

Question 4

We are interested in investigating the factors governing the precision of regression coefficients. Consider the model

$$\boldsymbol{Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i}$$

with OLS parameter estimates $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$. Under the Gauss–Markov assumptions, we have

$$\mathbf{Var}\left(\widehat{\beta}_2\right) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} (X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r_{X_2 X_3}^2},$$

where σ_ε^2 is the variance of ε and $r_{X_2X_3}$ is the sample correlation between X_2 and X_3.

(a) **Provide four factors that help with obtaining more precise parameter estimates for, say, $\widehat{\beta}_2$.**

(4 marks)

(b) **Assume that the true value of $\beta_3 = 0$, so that the above model includes an irrelevant variable. Discuss the effect of including this irrelevant variable on the unbiasedness and precision of $\widehat{\beta}_2$.**

(4 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 3.3 (Properties of the multiple linear regression coefficients), Chapter 6.3 (The effect of including a variable that ought not to be included).

Dougherty, C. Subject guide (2016): Chapter 3.

Approaching the question

(a) Candidates should note that the expression of the variance can be rewritten as:

$$\text{Var}(\widehat{\beta}_2) = \frac{\sigma_\varepsilon^2}{n \times MSD(X_2) \times (1 - r^2_{X_2X_3})}$$

where $MSD(X_2) \equiv n^{-1} \sum (X_{2i} - \bar{X}_2)^2$. The four factors that affect the precision then are: n, $MSD(X_2)$, $r^2_{X_2X_3}$ and σ_ε^2. Therefore, to obtain more precise parameter estimates of β_2 it is desirable to have: (i) small error variance, (ii) large sample size, (iii) large sample variability of the regressor X_2, and (iv) small correlation among the regressors X_2 and X_3.

Some candidates gave a discussion of the Gauss–Markov assumptions which is not the answer.

(b) Concept and consequences of including irrelevant variables are standard bookwork – unbiased but less precise because the irrelevant variable typically is correlated with the included regressor, i.e. $r^2_{X_2X_3} \neq 0$.

More precisely: Let $\tilde{\beta}_2$ be the OLS estimator in a regression without including X_3 and $\widehat{\beta}_2$ be the OLS estimator in a regression with X_3. If $r_{X_2,X_3} \neq 0$ we have:

$$\text{Var}\left(\widehat{\beta}_2\right) = \frac{\sigma_\varepsilon^2}{\sum\limits_{i=1}^{n} (X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r^2_{X_2X_3}} > \frac{\sigma_\varepsilon^2}{\sum\limits_{i=1}^{n} (X_{2i} - \bar{X}_2)^2} \equiv \text{Var}(\tilde{\beta}_2).$$

Question 5

A probit model to explain whether a firm is taken over by another firm during a given year postulates

$$\mathbf{Pr}(\boldsymbol{takeover} = 1 \mid \boldsymbol{x}) = \boldsymbol{\Phi}(\boldsymbol{\beta_0} + \boldsymbol{\beta_1 avgprof} + \boldsymbol{\beta_2 mtkval} + \boldsymbol{\beta_3 debtearn} + \boldsymbol{\beta_4 ceoten} + \boldsymbol{\beta_5 ceosal} + \boldsymbol{\beta_6 ceoage})$$

where $\Phi(z)$ is the cumulative standardised normal distribution. *takeover* is a binary response variable, *avgprof* is the firm's average profit margin over several prior years, *mktval* is the market value of the firm, *debtearn* is the debt-to-earnings ratio, and *ceoten*, *ceosal*, and *ceoage* are the tenure, annual salary, and age of the chief executive officer, respectively.

(a) **It is argued that using the probit regression model is better than using the linear probability model when explaining the binary variable *takeover*. Discuss the benefits/drawback of using the probit regression model when trying to explain a binary variable.**

(5 marks)

(b) **Discuss how you would implement the LR test that variables related to the CEO have no effect on the probability of takeover, other factors being equal. Clearly indicate the null, alternative, test statistic and rejection rule.**

(3 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 10.1 (The linear probability model), Chapter 10.3 (Probit analysis), and Chapter 10.6 (Introduction to maximum likelihood estimation.

Dougherty, C. Subject guide (2016): Chapter 10.

Approaching the question

(a) The probit model has two main advantages over the linear probability model (LPM): predicted probabilities are restricted to lie in $[0, 1]$ and MLE is (asymptotically) efficient whereas OLS (LPM) will be inefficient given the inherent presence of heteroskedasticity.

The main drawbacks of the probit model relative to the linear probability model is that the coefficients cannot be directly interpreted as the marginal effects of the regressor(s) of interest and it is also computationally more complicated.

(b) Let $\ln L^U$ be the maximised value of the log-likelihood of the unrestricted model and $\ln L^R$ be the maximised value of the log-likelihood of a model that excludes all the CEO-related variables (*ceoten*, *ceosal* and *ceoage*). Perform a likelihood-ratio test.

- $\text{H}_0 : \beta_4 = \beta_5 = \beta_6 = 0$ vs. $\text{H}_1 : \beta_4 \neq 0$ or $\beta_5 \neq 0$ or $\beta_6 \neq 0$.
- Test statistic: $LR = 2 \times (\ln L^U - \ln L^R)$.
- Under H_0, $LR \overset{a}{\sim} \chi^2_3$.
- At the 5% significance level, reject H_0 if $LR > 7.815$.

Section B

Answer three questions from this section.

Question 6

The following question concerns the effects of background characteristics on student's performance in the SAT (Scholastic Assessment Test). The SAT test is used for college admissions in the US.

$$\widehat{sat} = \underset{(6.29)}{1,028.10} + \underset{(3.83)}{19.3}\,hsize + \underset{(.53)}{-2.19}hsize^2 - \underset{(4.29)}{45.09}female - \underset{(12.71)}{169.81}black$$
$$+\underset{(12.71)}{62.31}female*black$$
$$n = 4,127, \ R^2 = .0858$$

The variable *hsize* is the size of the student's high school graduating class, in hundreds, *female* is a gender dummy variable (1 = female, 0 = male), and *black* is a race dummy variable (1 = black, 0 = otherwise). The standard errors are in parentheses.

(a) **What is the economic rationale for including $hsize^2$ in the above regression? Using this equation, determine for a given gender and race, what the graduating class size would be at which the predicted SAT scores are maximised.**

(5 marks)

(b) Holding *hsize* fixed, what is the estimated difference in SAT scores between nonblack females and nonblack males? Is this difference statistically significant? Interpret this result.

(5 marks)

(c) What is the estimated difference in SAT score between black females and nonblack females? What would you need to do to test whether the difference is statistically significant?

(5 marks)

(d) Discuss any problem you may have in estimating the model if all females in your sample are black. What name does this problem have and what can you do to mitigate this problem?

(5 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 4.3 (Models with quadratic variables), Chapter 6.5 (Testing a linear restriction), Chapter 5.1–5.3 (Dummy variables).

Dougherty, C. Subject guide (2016): Chapter 5.

Approaching the question

(a) The economic rationale for including a quadratic term in class size is that it allows for diminishing effects of class size on student's performance. The optimal class size is given by:

$$\left.\frac{\partial\widehat{SAT}}{\partial hsize}\right]_{hsize^*} = 0 \quad \Leftrightarrow \quad 19.3 - 2 \times 2.19 \times hsize^* = 0 \quad \Leftrightarrow \quad hsize^* = \frac{19.3}{4.38} \approx 4.4.$$

The optimal graduating class size is about 440 students.

(b) The expected difference in SAT scores between nonblack females and nonblack males is given by:

$$\mathrm{E}(sat \,|\, hsize, black = 0, female = 1) - \mathrm{E}(sat \,|\, hsize, black = 0, female = 0) = \beta_3.$$

The estimated difference is -45.03 which means that on average, holding the size of graduating class constant, nonblack females have a SAT score that is 45 points below nonblack males. To test the significance of this difference we perform a hypothesis test:

- $\mathrm{H}_0 : \beta_3 = 0$ vs. $\mathrm{H}_1 : \beta_3 \neq 0$.
- Test statistic: $t = \widehat{\beta}_3/se(\widehat{\beta}_3) = -45.09/4.29 \approx -10.5$.
- Under H_0 and assuming Gauss–Markov plus normality, $t \sim t_{4121}$.
- At the 5% significance level, we reject H_0 since $|t| > 1.96$.

(c) The expected difference in SAT scores between black females and nonblack females is given by:

$$\mathrm{E}(sat \,|\, hsize, black = 1, female = 1) - \mathrm{E}(sat \,|\, hsize, black = 0, female = 1) = \beta_4 + \beta_5.$$

The estimated difference is given by $-169.81 + 62.31 = -107.5$ which means that, holding size of graduating class constant, on average black females have a sat score that is 107.5 points lower than their white counterparts. To perform the test of $\mathrm{H}_0 : \beta_4 + \beta_5 = 0$ vs. $\mathrm{H}_1 : \beta_4 + \beta_5 \neq 0$ one would need either: (i) $\widehat{\mathrm{Cov}}(\widehat{\beta}_4, \widehat{\beta}_5)$ to compute the t statistic, or (ii) the R^2 of the restricted model to compute the F statistic.

(d) If all females are black, then $female_i = female_i \times black_i \ \forall \ i$, so the variables will be perfectly collinear and he would not be able to obtain the OLS estimates. In this case, we could either drop the interaction term from the regression at the expense of not being able to identify any heterogeneity in the effects or obtain more data such that the sample contains some white females.

Question 7

Let us consider the estimation of a hedonic price function for houses. The hedonic price refers to the implicit price of a house given certain attributes (e.g., the number of bedrooms). The data contains the sale price of 546 houses sold in the summer of 1987 in Canada along with their important features. The following characteristics are available: the lot size of the property in square feet (*lotsize*), the numbers of bedrooms (*bedrooms*), the number of full bathrooms (*bathrooms*), and a dummy indicating the presence of airconditioning (*airco*).

Consider the following ordinary least squares results

$$\begin{aligned}\widehat{\log(price)}_i = \underset{\substack{(.232)\\ [.233]}}{7.094} + \underset{\substack{(.028)\\ [.028]}}{0.400}\log(lotsize)_i + \underset{\substack{(.015)\\ [.017]}}{0.078}bedrooms_i + \\ \underset{\substack{(.023)\\ [.024]}}{0.216}bathrooms_i + \underset{\substack{(.024)\\ [.023]}}{0.212}airco_i \quad n = 546,\ RSS = 32.622\end{aligned} \tag{7.1}$$

The usual standard errors are in parentheses, the heteroskedasticity robust standard errors are in square brackets, and *RSS* measures the residual sum of squares.

(a) Interpret the parameter estimates on log(*lotsize*), *bedrooms*, and *airco*. Briefly discuss the statistical significance of the results.

(5 marks)

(b) Suppose that lot size was measured in square metres rather than square feet. How would this affect the parameter estimates of the slopes and intercept? How would this affect the fitted values? *Note*: the conversion (approximate) $1m^2 = 10ft^2$.

(5 marks)

(c) We are interested in testing the hypothesis $\mathrm{H}_0 : \beta_{bedrooms} = \beta_{bathrooms}$ against the alternative $H_1 : \beta_{bedrooms} \neq \beta_{bathrooms}$. Discuss a test for this hypothesis that makes use of the following restricted regression result

$$\begin{aligned}\widehat{\log(price)}_i = \underset{(.234)}{6.994} + \underset{(.282)}{0.408}\log(lotsize)_i + \underset{(.011)}{0.127}bbrooms_i + \underset{(.024)}{0.215}airco_i \\ n = 546,\ RSS = 33.758\end{aligned} \tag{7.2}$$

where *bbrooms* = *bedrooms* + *bathrooms*. Clearly indicate the assumptions you are making for this test to be valid.

(5 marks)

(d) You are interested in testing for the presence of heteroskedasticity. Say you are told that the variance is increasing with log(*lotsize*). Discuss how you would test for the presence of heteroskedasticity. What is the name of the test you are proposing?

(5 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 1.4 (Interpretation of a regression equation – units of measurement), Chapter 4.2 (Logarithmic transformations), Chapter 2.6 (Testing hypotheses relating to the regression coefficients), Chapter 7.1 (Heteroskedasticity and its implications), and Chapter 7.2 (Detection of heteroskedasticity).

Dougherty, C. Subject guide (2016): Chapter 7.

Approaching the question

(a) Clear discussion of interpretation required (units not always clear): On average, holding the remaining variables in the regression constant, (i) a 1% increase in lot size is associated with a 0.4% increase in house price, (ii) each extra bedroom is associated with a 7.8% increase in house price, and (iii) houses with air conditioning are 21.2% more expensive than those without. All estimates are statistically significant at 5% significance levels. Candidates should clearly indicate H_0 and H_1, the test statistic and the rejection rule.

(b) Let $lotsize_i$ be the lot size in square feet and $\widetilde{lotsize}_i$ be the lot size in square metres. We have that $\widetilde{lotsize}_i = (10)^{-1} lotsize_i$ and $\log(\widetilde{lotsize}_i) = \log((10)^{-1}) + \log(lotsize_i)$. Since this is an additive transformation of one of the explanatory variables we have that (i) the regression slopes will not be affected, (ii) the intercept will change to $7.094 - \log((10)^{-1}) \times 0.4$, and (iii) the fitted values will also not be affected. Many candidates made an error here, ignoring the fact that the variable whose measurement was changed entered in log form.

(c) Candidates would have to recognise that (6.2) is a restricted version of (6.1) where $\beta_{bedrooms} = \beta_{bathrooms}$ is imposed. We therefore need to use the F test. Test statistic:

$$F = \frac{RRSS - URSS}{URSS} \times \frac{n-K}{J} = \frac{33.758 - 32.622}{32.622} \times \frac{541}{1} \approx 18.84.$$

Assuming the Gauss–Markov assumptions plus normality of the error term hold: under H_0, $F \sim F_{1,\,541}$. At the 5% signicance level we reject H_0 since $F > 3.86$. Conclusion: The effect of one extra bathroom is different from the effect of one extra bedroom. Some candidates did not recognise this and were proposing a test on the coefficient of *bbrooms* equalling zero which is wrong.

(d) Assuming Gauss–Markov assumptions plus the normality of the error hold, he can use the Goldfeld–Quandt test for heteroskedasticity. For that purpose, we should first order the 546 observations by the magnitude of $\log(lotsize_i)$. Fit one regression for the first n^* observations and another for the last n^* observations (usually n^* equals one-third of the sample). Let RSS_1 and RSS_2 denote the sum of squared residuals in each of these regressions, respectively.

- $\mathrm{H}_0 : \sigma_2^2 = \sigma_1^2$ vs. $\mathrm{H}_1 : \sigma_2^2 > \sigma_1^2$.
- Test statistic: $GQ = RSS_2/RSS_1$.
- Under H_0, $GQ \sim F_{n^*-k,\, n^*-k}$.
- Reject if GQ is greater than the 95th percentile of the F distribution above.

Candidates need to be careful not to state $\mathrm{H}_0 : RSS_2 = RSS_1$ vs. $\mathrm{H}_1 : RSS_2 > RSS_1$. Both RSS_1 and RSS_2 are random variables. Because the sample sizes are identical it is also correct to state $\mathrm{H}_0 : RSS_1$ and RSS_2 are not statistically different. Note that simply by having one sample larger than the other, you could also have a larger residual sum of squares.

Question 8

Let $math10$ denote the percentage of students at a high school receiving a passing score on a standardised math test. We are interested in estimating the effect of per student spending on math performance. A simple model is

$$math10_i = \beta_0 + \beta_1 \log(expend_i) + \beta_2 \log(enroll_i) + \beta_3 poverty_i + u_i \qquad (8.1)$$

where, for each high school i; $poverty_i$ is the percentage of students living in poverty, $expend_i$ is the spending per student and $enroll_i$ the number of registered students. You may assume that this model satisfies all Gauss–Markov assumptions.

You are faced with the fact that data is unavailable on a key variable: *poverty*.

(a) Discuss the properties (unbiasedness and consistency) of the estimators when you drop the variable poverty. Explain your answers.

(5 marks)

You do have information available on a closely related variable: the percentage of students eligible for the federally funded school lunch program, $lnchprg_i$. Let us consider using $lnchprg_i$ as a proxy for $poverty_i$.

(b) Briefly discuss why $lnchprg_i$ is a sensible proxy variable for the unobserved variable $poverty_i$.

(2 marks)

(c) It is unlikely that $lnchprg_i$ is an ideal proxy, in the sense that there is an exact linear relationship between them, instead, we will assume that

$$poverty_i = \alpha_0 + \alpha_1 lnchprg_i + v_i, \quad \alpha_1 \neq 0 \qquad (8.2)$$

Discuss the assumptions you need to make to enable consistent parameter estimators of β_1 and β_2 using your estimable equation

$$math10_i = \gamma_0 + \gamma_1 \log(expend_i) + \gamma_2 \log(enroll_i) + \gamma_3 lnchprg_i + e_i,$$

***Hint*: Consider the relation between the γ and the β parameters and express e_i in terms of u_i and v_i.**

(5 marks)

(d) The OLS results with and without $lnchprg_i$ as an explanatory variable are given by (standard errors in parentheses):

$$\widehat{math10}_i = \underset{(26.72)}{-69.24} + \underset{(3.30)}{11.13} \log expend_i + \underset{(0.615)}{0.022} \log emroll_i$$
$$N = 428, \ R^2 = 0.0297$$

$$\widehat{math10}_i = \underset{(24.99)}{-23.14} + \underset{(3.04)}{7.75} \log expend_i - \underset{(0.58)}{1.26} \log emroll_i - \underset{(0.036)}{0.324} lnchprg_i$$
$$N = 428, \ R^2 = 0.1893$$

i. Interpret the coefficient on $lnchprg$. What does this parameter tell us regarding the parameter of interest β_3?

(4 marks)

ii. Give an intuitive discussion explaining why the effect of expenditures on $math10_i$ is lower in the regression where $lnchprg_i$ is included than where it is excluded.

(4 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 6.2 (The effect of omitting a variable that ought to be included), Chapter 6.4 (Proxy variables).

Dougherty, C. Subject guide (2016): Chapter 6.

Approaching the question

(a) Consider rewriting (8.1) as:

$$math10_i = \beta_0 + \beta_1 \log(expend_i) + \beta_2 \log(enroll_i) + \varepsilon_i$$

where $\varepsilon_i = \beta_3 poverty_i + u_i$. Assuming $\beta_3 \neq 0$, if poverty is correlated with either (log) expenditures and/or (log) enrollment the model will suffer from endogeneity due to omitted variables and OLS will be biased and inconsistent. That is likely to be the case since schools that have smaller expenditures tend to be located in poorer neighbourhoods and hence to have more students living in poverty conditions.

(b) It is a sensible proxy because it is likely to be correlated with poverty and to capture some of the effect of poverty since usually students eligible for the lunch program tend to be those with low levels of family income.

(c) Plug in (8.2) for poverty in (8.1) to obtain:

$$math10_i = \underbrace{(\beta_0 + \beta_3\alpha_0)}_{\gamma_0} + \underbrace{\beta_1}_{\gamma_1} \log(expend_i) + \underbrace{\beta_2}_{\gamma_2} \log(enroll_i) + \underbrace{\beta_3\alpha_1}_{\gamma_3} lnchprg_i + \underbrace{(\beta_3 v_i + u_i)}_{\varepsilon_i}.$$

Given that the model (8.1) satisfies all the Gauss–Markov assumptions, to obtain consistent estimates of β_1 and β_2 we need that: (i) v_i is uncorrelated with $\log(expend_i)$, $\log(enroll_i)$ and $lnchprg_i$, and (ii) $lnchprg_i$ is uncorrelated with u_i.

(d) i. On average, holding expenditure and enrollment constant, a 1 percentage point increase in the number of students eligible for the lunch program is associated with a 0.324 percentage point fall in the percentage of students receiving a passing score in the standardised math test. Since $\gamma_3 = \alpha_1\beta_3$ and assuming $\alpha_1 > 0$, the direction of the effect (sign) of poverty of $math10$ is the same as the effect of $lnchprg$ on $math10$, in this case, with a negative coefficient on $lnchprg$ we can infer that poverty has a negative effect on $math10$.

ii. Omitting relevant variables will result in the remaining parameters attempting to pick up its effect through the correlation these omitted variables have with the included regressors. We, therefore, expect the effect to be smaller, as part of the effect we attribute to expenditure in the short regression is actually coming from the fact that high schools that have larger expenditures tend to have fewer students eligible for the lunch program and those students tend to perform worse in the standardised math test.

Question 9

Let us consider monthly data on the short-term interest rate (the three month Treasury Bill rate) and on the AAA corporate bond yield in the USA. The data run from January 1950 to December 1999. Let $DUS3MT$ denote the changes in three-month Treasury Bill rate, and $DAAA$ denote the changes in AAA bond rate. We consider the following results (with the standard errors given in parentheses)

$$\widehat{DAAA}_t = \underset{(.007)}{0.006} + \underset{(.015)}{0.275} DUS3MT_t \quad t = 1, \ldots, 600 \qquad (9.1)$$

$$RSS = 17.486; \ DW = 1.447$$

where RSS is the residual sum of squares and DW is the Durbin–Watson test.

A researcher interpreting the residuals suggests that the errors show a positive correlation over time.

(a) What are the consequences of this correlation for the above regression results?

(5 marks)

(b) Use the results above to test for the presence of first-order positive autocorrelation. Clearly specify the null and alternative hypothesis, test statistic, assumptions underlying the test, and the acceptance/rejection rule.

(5 marks)

(c) In an attempt to remove the autocorrelation you consider the following specification

$$\widehat{DAAA}_t = \underset{(.007)}{0.005} + \underset{(.015)}{0.252} DUS3MT_t - \underset{(.018)}{0.080} DUS3MT_{t-1} + \underset{(.040)}{0.290} DAAA_{t-1} \quad (9.2)$$

$$RSS = 16.087; \ DW = 1.897$$

Comment on the following statement 'The Durbin–Watson statistic is closer to 2, indicating that we have succesfully removed the autocorrelation'. If you disagree with this statement, suggest what you would need to do instead.

(5 marks)

(d) **Discuss the Common Factor Test as a model specification suitable for this model. What extra information do you need to conduct this test.**

(5 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 12.1–12.3 (Definition, consequences and detection of autocorrelation; Fitting a model subject to AR(1) autocorrelation).

Dougherty, C. Subject guide (2016): Chapters 12.

Approaching the question

(a) Standard bookwork discussion of autocorrelation in the setting where there are no lagged dependent variables.

(b) Discussion of Durbin–Watson test expected – bookwork. Candidates should clearly indicate the H_0 and H_1, test statistic and rejection rule. Assumptions underlying the test: tests only AR(1) autocorrelation, no lagged dependent variables (deterministic regressors only) in presence of intercept.

(c) We disagree with the statement because the Durbin–Watson test is not valid in the presence of lagged dependent variables since in this case regressors cannot be strictly exogenous. Instead, we should use the Breusch–Godfrey test which is asymptotically valid with predetermined regressors a weaker requirement than strict exogeneity. Alternatively, the Durbin h test can be proposed.

(d) Assuming AR(1) autocorrelation and letting $Y_t = DAAA_t$ and $X_t = DUS3MT_t$ we can remove the autocorrelation by rewriting the model as:

$$Y_t = (1-\rho)\beta_1 + \rho Y_{t-1} + \beta_2 X_t - \beta_2 \rho X_{t-1} + \varepsilon_t.$$

This is a restricted version of a more general ADL(1,1) model:

$$Y_t = \lambda_1 + \lambda_2 Y_{t-1} + \lambda_3 X_t + \lambda_4 X_{t-1} + \varepsilon_t$$

under the restriction $\lambda_4 = -\lambda_2\lambda_3$. We can test this restriction using the Common Factor Test. Let RSS_r and RSS_u be the restricted sums of squares of the restricted and unrestricted model, respectively.

- Test statistic: $CF = n \times \log(RSS_r/RSS_u)$.
- Under H_0, $CF \overset{a}{\sim} \chi^2_1$.
- Extra information needed: RSS_r (since RSS_u is given in (9.2)).

Question 10

Consider the model

$$\boldsymbol{y_t = \alpha + \beta x_t + \varepsilon_t, \quad t = 1, \ldots, T} \qquad (10.1)$$

where y_t and x_t are both integrated of order one.

(a) **Explain what it means to say that y_t is integrated of order one. Discuss how you would test for this. In your answer make sure that it is clear how to implement your test.**

(6 marks)

(b) **Give an example of an economic variable that is potentially integrated of order one and give an intuitive explanation why you expect this process to be integrated of order 1.**

(2 marks)

(c) It will be important to distinguish whether the above relationship is 'spurious' as opposed to 'cointegrating'.

i. Explain what it means to say that y_t and x_t have a cointegrating relationship and how does that contrast to a spurious relationship.

(4 marks)

ii. Discuss how you can test for evidence of a cointegrating relationship.

(4 marks)

(d) Suppose that

$$\varepsilon_t = \rho\varepsilon_{t-1} + v_t, \quad |\rho| < 1,$$

and v_t is an i.i.d. $(0, \sigma^2)$ innovation which is independent of ε_{t-1}. Show that you can rewrite equation (10.1) in terms of an error correction model:

$$\Delta y_t = \delta_1 \Delta x_t + \delta_2 (y_{t-1} - \alpha - \beta x_{t-1}) + v_t.$$

Clearly indicate the relation between (δ_1, δ_2) and (α, β, ρ). Give an economic intuition behind this result.

(4 marks)

Reading for this question

Dougherty, C. *Introduction to econometrics* (fifth edition): Chapter 13.1 (Stationarity and nonstationarity), Chapter 13.4–13.5 (Tests of nonstationarity), and Chapter 13.6 (Cointegration).

Dougherty, C. Subject guide (2016): Chapter 13.

Approaching the question

(a) Integrated of order one (or simply I(1)) means that the process can be made stationary by differencing once. One could test for this using a Dickey–Fuller or an Augmented Dickey–Fuller test. Standard bookwork. For instance, suppose $y_t = \rho y_{t-1} + \varepsilon_t$. We can conduct the Dickey–Fuller test by running the auxiliary regression of the form:

$$\Delta y_t = \gamma y_{t-1} + \varepsilon_t$$

where $\gamma \equiv \rho - 1$. Provide test statistic and discuss rejection rule, clearly indicating H_0 and H_1.

(b) One example could be CPI due to its clearly trending behaviour over time it can be suggestive of a time-series integrated of order 1.

(c) i. Two variables integrated of order one are said to be cointegrated if there exists a linear combination of them that is integrated of order zero. In those cases, a regression of (10.1) will capture the long-run relationship between y_t and x_t and we should prefer an Error Correction Model to capture both the short-term and the long-term relationships among the two variables. In contrast, if the two series are not cointegrated a regression like (10.1) is likely to indicate a very strong relationship between the two variables even when there is no relationship at all. Such a regression is called a spurious regression.

ii. If y_t and x_t are both found to be integrated of order one, one could test for cointegration as follows:

- Run a regression of y_t on a constant and x_t and collect residuals.
- Perform a DF or ADF test on the residual series.
- If reject the null of non-stationarity conclude the series are cointegrated.

(d) Lag (10.1) by one period and multiply both sides by ρ to obtain:

$$\rho y_{t-1} = \rho\alpha + \rho\beta x_{t-1} + \rho\varepsilon_{t-1}.$$

Subtract the above from (10.1), hence:

$$y_t - \rho y_{t-1} = (1-\rho)\alpha + \beta x_t - \rho\beta x_{t-1} + \underbrace{\varepsilon_t - \rho\varepsilon_{t-1}}_{v_t}.$$

Rearranging:

$$y_t = \alpha + \rho(y_{t-1} - \alpha - \beta x_{t-1}) + \beta x_t + v_t.$$

Subtract $y_{t-1} = \alpha + \beta x_{t-1} + \varepsilon_{t-1}$, hence:

$$\Delta y_t = \rho(y_{t-1} - \alpha - \beta x_{t-1}) + \beta(x_t - x_{t-1}) + v_t - \varepsilon_{t-1}.$$

Finally, using $\varepsilon_{t-1} = y_{t-1} - \alpha - \beta x_{t-1}$ we obtain:

$$\Delta y_t = \beta\Delta x_t + (\rho - 1)(y_{t-1} - \alpha - \beta x_{t-1}) + v_t.$$

Therefore, $\delta_1 = \beta$ and $\delta_2 = \rho - 1$. Economic intuition: this is a process that reverts to the long-run equilibrium after a shock (δ_1 captures the short-run dynamics and δ_2 captures the speed of the reversal to the long-run equilibrium).