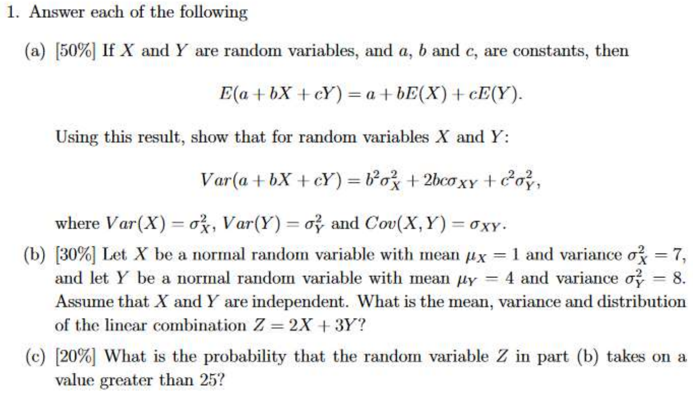
Probability and Statistics

2019, A



1. Answer each of the following

(a) [50%] If X and Y are random variables, and a, b and c, are constants, then

$$E(a+bX+cY)=a+bE(X)+cE(Y).$$

Using this result, show that for random variables X and Y:

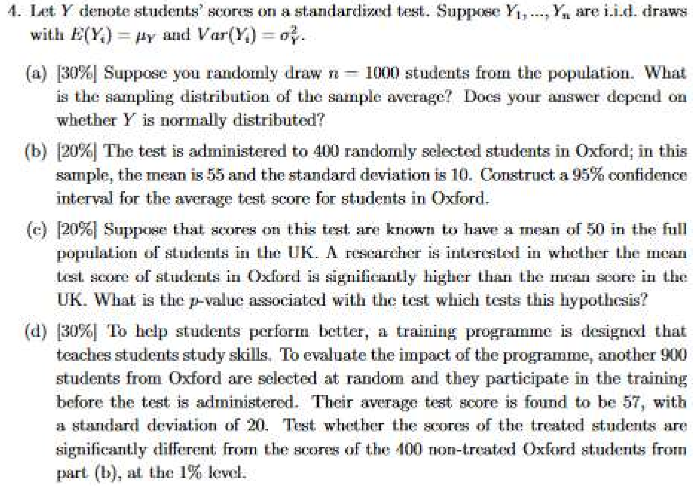
$$Var(a+bX+cY)=b^2\sigma_X^2+2bc\sigma_{XY}+c^2\sigma_Y^2,$$

where $Var(X)=\sigma_X^2$, $Var(Y)=\sigma_Y^2$ and $Cov(X,Y)=\sigma_{XY}$.

(b) [30%] Let X be a normal random variable with mean $\mu_X=1$ and variance $\sigma_X^2=7$, and let Y be a normal random variable with mean $\mu_Y=4$ and variance $\sigma_Y^2=8$. Assume that X and Y are independent. What is the mean, variance and distribution of the linear combination $Z=2X+3Y$?

(c) [20%] What is the probability that the random variable Z in part (b) takes on a value greater than 25?

2019, B



4. Let Y denote students' scores on a standardized test. Suppose $Y_1,...,Y_n$ are i.i.d. draws with $E(Y_i)=\mu_Y$ and $Var(Y_i)=\sigma_Y^2$.

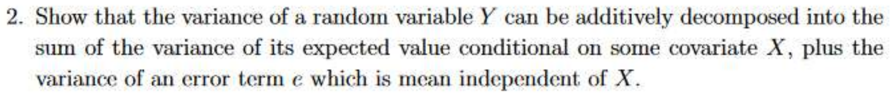
(a) [30%] Suppose you randomly draw $n=1000$ students from the population. What is the sampling distribution of the sample average? Does your answer depend on whether Y is normally distributed?

(b) [20%] The test is administered to 400 randomly selected students in Oxford; in this sample, the mean is 55 and the standard deviation is 10. Construct a 95% confidence interval for the average test score for students in Oxford.

(c) [20%] Suppose that scores on this test are known to have a mean of 50 in the full population of students in the UK. A researcher is interested in whether the mean test score of students in Oxford is significantly higher than the mean score in the UK. What is the p-value associated with the test which tests this hypothesis?

(d) [30%] To help students perform better, a training programme is designed that teaches students study skills. To evaluate the impact of the programme, another 900 students from Oxford are selected at random and they participate in the training before the test is administered. Their average test score is found to be 57, with a standard deviation of 20. Test whether the scores of the treated students are significantly different from the scores of the 400 non-treated Oxford students from part (b), at the 1% level.

2018, A



2. Show that the variance of a random variable Y can be additively decomposed into the sum of the variance of its expected value conditional on some covariate X, plus the variance of an error term e which is mean independent of X.

2018, B

4. Let X_i be a Bernoulli random variable with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$.

(a) What are the density and distribution functions of X_i? [10%]

(b) Find the expected value, variance and skewness of X_i. [20%]

For $X_1, ..., X_n$ independent, identically Bernoulli(p), let $\hat{p} = n^{-1} \sum_{i=1}^{n} X_i$.

(c) What is the standard error of $\hat{p}$? [20%]

(d) Explain why

$$t = \frac{\hat{p} - p}{se(\hat{p})} \xrightarrow{\mathsf{D}} \mathsf{N}(0, 1),$$

where $se(\hat{p})$ is the standard error of $\hat{p}$. [30%]

(e) Suppose that in a sample of size $n = 100$, we obtain $\hat{p} = 0.3$. Construct an approximate 95% confidence interval for p. State all the results being used. [20%]

4. Let Y denote students' scores on a standardized test. Suppose $Y_1, ..., Y_n$ are i.i.d. draws with $E(Y_i) = \mu_Y$ and $Var(Y_i) = \sigma_Y^2$.

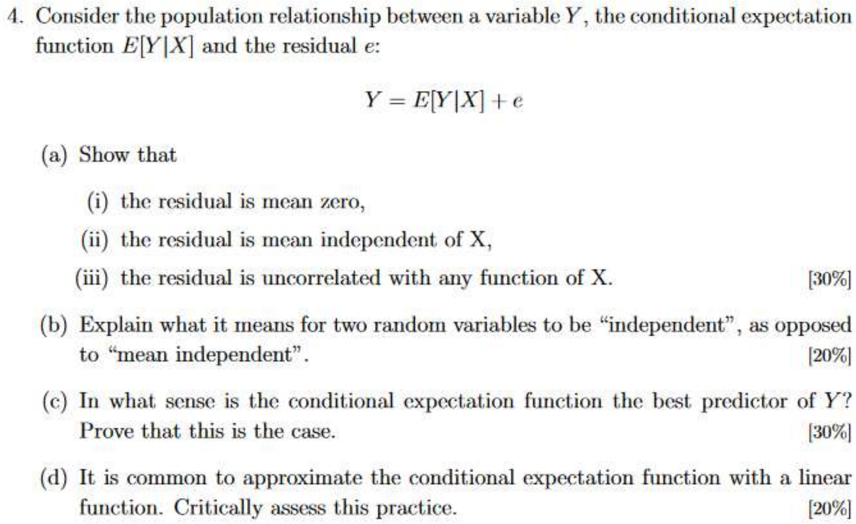
(a) [30%] Suppose you randomly draw $n = 1000$ students from the population. What is the sampling distribution of the sample average? Does your answer depend on whether Y is normally distributed?

(b) [20%] The test is administered to 400 randomly selected students in Oxford; in this sample, the mean is 55 and the standard deviation is 10. Construct a 95% confidence interval for the average test score for students in Oxford.

(c) [20%] Suppose that scores on this test are known to have a mean of 50 in the full population of students in the UK. A researcher is interested in whether the mean test score of students in Oxford is significantly higher than the mean score in the UK. What is the p-value associated with the test which tests this hypothesis?

(d) [30%] To help students perform better, a training programme is designed that teaches students study skills. To evaluate the impact of the programme, another 900 students from Oxford are selected at random and they participate in the training before the test is administered. Their average test score is found to be 57, with a standard deviation of 20. Test whether the scores of the treated students are significantly different from the scores of the 400 non-treated Oxford students from part (b), at the 1% level.

2017, B



4. Consider the population relationship between a variable Y, the conditional expectation function $E[Y|X]$ and the residual e:

$$Y = E[Y|X] + e$$

(a) Show that

(i) the residual is mean zero,

(ii) the residual is mean independent of X,

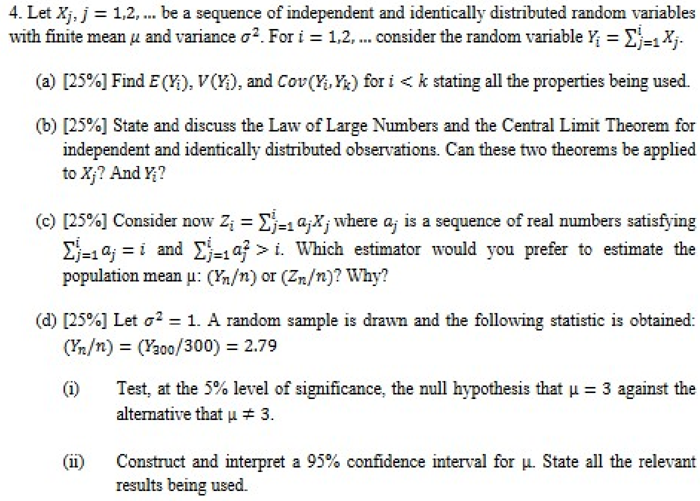
(iii) the residual is uncorrelated with any function of X. [30%]

(b) Explain what it means for two random variables to be "independent", as opposed to "mean independent". [20%]

(c) In what sense is the conditional expectation function the best predictor of Y? Prove that this is the case. [30%]

(d) It is common to approximate the conditional expectation function with a linear function. Critically assess this practice. [20%]

2016, B



4. Let X_j, $j = 1,2, \ldots$ be a sequence of independent and identically distributed random variables with finite mean μ and variance σ^2. For $i = 1,2, \ldots$ consider the random variable $Y_i = \sum_{j=1}^{i} X_j$.

(a) [25%] Find $E(Y_i)$, $V(Y_i)$, and $Cov(Y_i, Y_k)$ for $i < k$ stating all the properties being used.

(b) [25%] State and discuss the Law of Large Numbers and the Central Limit Theorem for independent and identically distributed observations. Can these two theorems be applied to X_j? And Y_i?

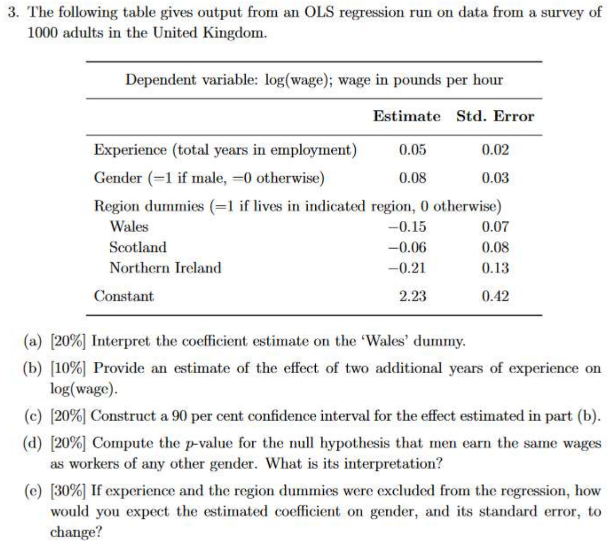
(c) [25%] Consider now $Z_i = \sum_{j=1}^{i} a_j X_j$ where a_j is a sequence of real numbers satisfying $\sum_{j=1}^{i} a_j = i$ and $\sum_{j=1}^{i} a_j^2 > i$. Which estimator would you prefer to estimate the population mean μ: (Y_n/n) or (Z_n/n)? Why?

(d) [25%] Let $\sigma^2 = 1$. A random sample is drawn and the following statistic is obtained: $(Y_n/n) = (Y_{300}/300) = 2.79$

(i) Test, at the 5% level of significance, the null hypothesis that $\mu = 3$ against the alternative that $\mu \neq 3$.

(ii) Construct and interpret a 95% confidence interval for μ. State all the relevant results being used.

Introduction to Regression and Advanced Topics in Regression



3. The following table gives output from an OLS regression run on data from a survey of 1000 adults in the United Kingdom.

Dependent variable: log(wage); wage in pounds per hour

	Estimate	Std. Error
Experience (total years in employment)	0.05	0.02
Gender (=1 if male, =0 otherwise)	0.08	0.03
Region dummies (=1 if lives in indicated region, 0 otherwise)		
Wales	-0.15	0.07
Scotland	-0.06	0.08
Northern Ireland	-0.21	0.13
Constant	2.23	0.42

(a) [20%] Interpret the coefficient estimate on the 'Wales' dummy.

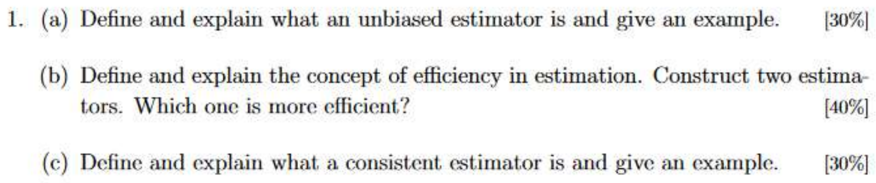
(b) [10%] Provide an estimate of the effect of two additional years of experience on log(wage).

(c) [20%] Construct a 90 per cent confidence interval for the effect estimated in part (b).

(d) [20%] Compute the p-value for the null hypothesis that men earn the same wages as workers of any other gender. What is its interpretation?

(e) [30%] If experience and the region dummies were excluded from the regression, how would you expect the estimated coefficient on gender, and its standard error, to change?

2018, A



1. (a) Define and explain what an unbiased estimator is and give an example. [30%]

(b) Define and explain the concept of efficiency in estimation. Construct two estimators. Which one is more efficient? [40%]

(c) Define and explain what a consistent estimator is and give an example. [30%]

2018, B

5. Let us define positive assortative mating in wealth as the situation when men and women with the same level of wealth marry more frequently than what would be expected under a marriage pattern that is random in terms of wealth. Assume that all individuals have positive wealth and that you have data on all first marriages occurring in 1960, 1970, 1980, 1990, 2000, and 2010, and wife's log wealth ($\log(x_w)$) and husband's log wealth ($\log(x_h)$) at the time of marriage.

Table 1: Descriptive statistics, population quantities

	1960	1970	1980	1990	2000	2010
$COV(\log(x_w), \log(x_h))$	10	11	12	13	14	15
$SD(\log(x_w))$	5	5.1	5.2	5.3	5.4	5.5
$SD(\log(x_h))$	4.5	4.75	5	5.25	5.5	5.75

Note: COV denotes covariance and SD denotes standard deviation.

(a) Using the population quantities in Table 1, compute and plot the degree of assortative mating by year in wealth as measured by

(i) the regression coefficient on wife's log wealth on a regression of husband's log wealth on wife's log wealth, and

(ii) the regression coefficient on husband's log wealth on a regression of wife's log wealth on husband's log wealth. [20%]

(b) Can your findings in (a) be explained by the presence of measurement error in husband's log wealth? Discuss. [20%]

(c) Suppose that you can link the information on all first marriages in 1960, 1970 and 1980, with information on children born of these marriages, so that you can run the following regression:

$$\log(y) = \alpha + \beta_m \log(x_m) + \beta_f \log(x_f) + \gamma X + e,$$

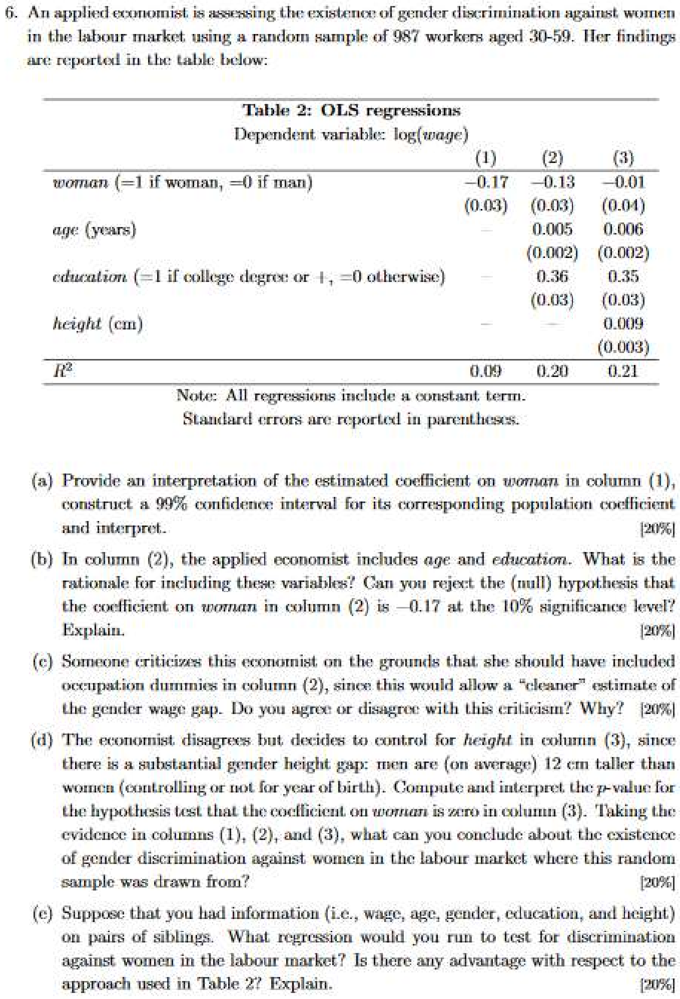
where $\log(y)$ is children's log wealth at age 35, $\log(x_m)$ is mother's log wealth at marriage, $\log(x_f)$ is father's log wealth at marriage and X is a vector of control variables. Write down the expressions for the OLS estimands of β_m and β_f. Do these estimands reflect causal effects? Explain. [20%]

(d) Describe how you would test whether, conditional on X, parental log wealth predicts children's log wealth. [20%]

(e) Suppose that father's log wealth is unobservable in your data, so that you are forced to run the following regression:

$$\log(y) = \alpha^0 + \beta_m^0 \log(x_m) + \gamma^0 X + u.$$

What is the relationship between β_m^0 and β_m? Explain. [20%]



6. An applied economist is assessing the existence of gender discrimination against women in the labour market using a random sample of 987 workers aged 30-59. Her findings are reported in the table below:

Table 2: OLS regressions

Dependent variable: log(*wage*)

	(1)	(2)	(3)
woman (=1 if woman, =0 if man)	−0.17	−0.13	−0.01
	(0.03)	(0.03)	(0.04)
age (years)	–	0.005	0.006
		(0.002)	(0.002)
education (=1 if college degree or +, =0 otherwise)	–	0.36	0.35
		(0.03)	(0.03)
height (cm)	–	–	0.009
			(0.003)
R^2	0.09	0.20	0.21

Note: All regressions include a constant term.
Standard errors are reported in parentheses.

(a) Provide an interpretation of the estimated coefficient on *woman* in column (1), construct a 99% confidence interval for its corresponding population coefficient and interpret. [20%]

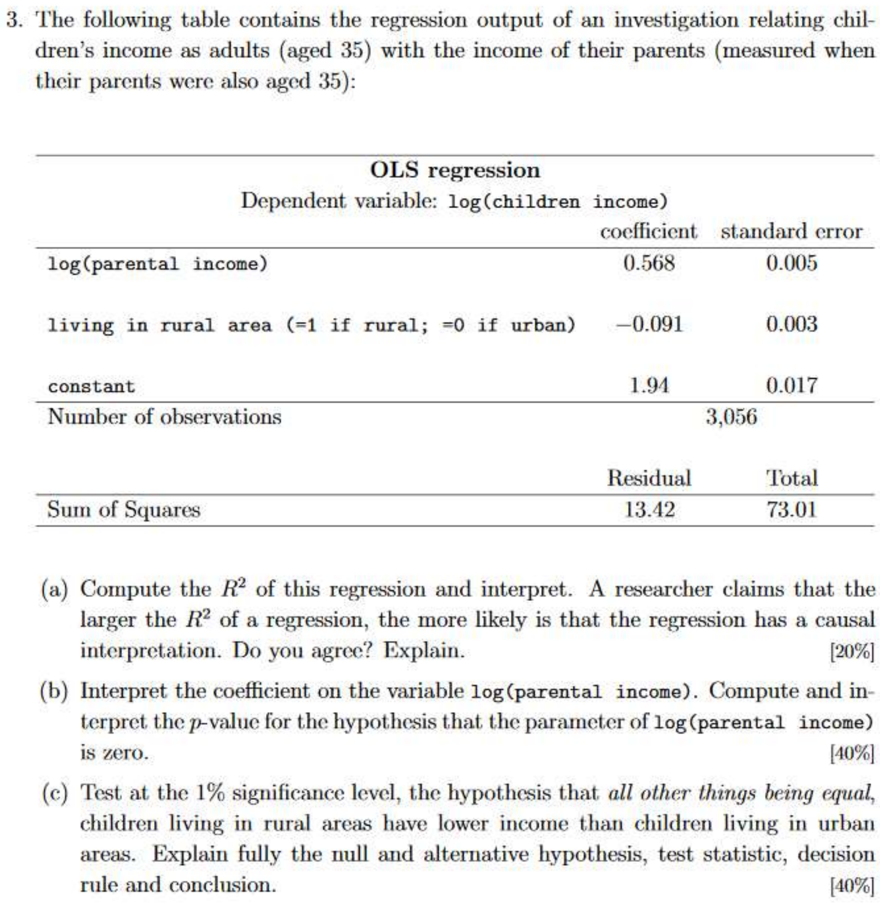
(b) In column (2), the applied economist includes *age* and *education*. What is the rationale for including these variables? Can you reject the (null) hypothesis that the coefficient on *woman* in column (2) is −0.17 at the 10% significance level? Explain. [20%]

(c) Someone criticizes this economist on the grounds that she should have included occupation dummies in column (2), since this would allow a "cleaner" estimate of the gender wage gap. Do you agree or disagree with this criticism? Why? [20%]

(d) The economist disagrees but decides to control for *height* in column (3), since there is a substantial gender height gap: men are (on average) 12 cm taller than women (controlling or not for year of birth). Compute and interpret the p-value for the hypothesis test that the coefficient on *woman* is zero in column (3). Taking the evidence in columns (1), (2), and (3), what can you conclude about the existence of gender discrimination against women in the labour market where this random sample was drawn from? [20%]

(e) Suppose that you had information (i.e., wage, age, gender, education, and height) on pairs of siblings. What regression would you run to test for discrimination against women in the labour market? Is there any advantage with respect to the approach used in Table 2? Explain. [20%]

2017, A



3. The following table contains the regression output of an investigation relating children's income as adults (aged 35) with the income of their parents (measured when their parents were also aged 35):

OLS regression

Dependent variable: `log(children income)`

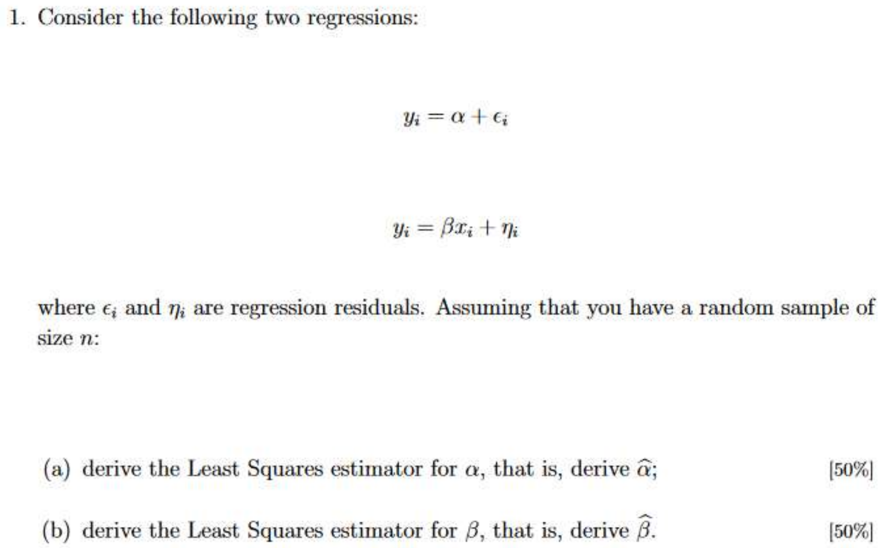
	coefficient	standard error
`log(parental income)`	0.568	0.005
`living in rural area (=1 if rural; =0 if urban)`	-0.091	0.003
`constant`	1.94	0.017
Number of observations	3,056	
	Residual	Total
Sum of Squares	13.42	73.01

(a) Compute the R^2 of this regression and interpret. A researcher claims that the larger the R^2 of a regression, the more likely is that the regression has a causal interpretation. Do you agree? Explain. [20%]

(b) Interpret the coefficient on the variable `log(parental income)`. Compute and interpret the p-value for the hypothesis that the parameter of `log(parental income)` is zero. [40%]

(c) Test at the 1% significance level, the hypothesis that *all other things being equal*, children living in rural areas have lower income than children living in urban areas. Explain fully the null and alternative hypothesis, test statistic, decision rule and conclusion. [40%]

2017, A



1. Consider the following two regressions:

$$y_i = \alpha + \epsilon_i$$

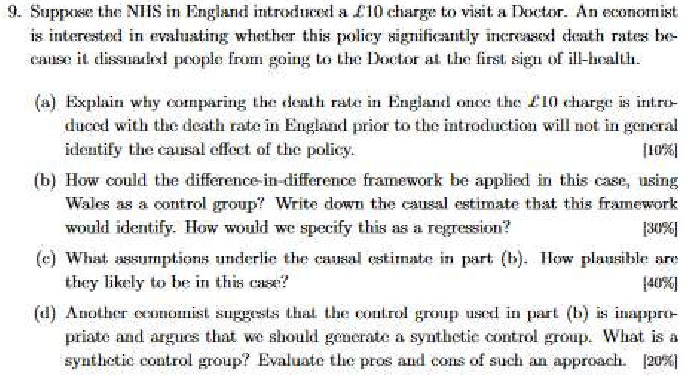
$$y_i = \beta x_i + \eta_i$$

where ϵ_i and η_i are regression residuals. Assuming that you have a random sample of size n:

(a) derive the Least Squares estimator for α, that is, derive $\widehat{\alpha}$; [50%]

(b) derive the Least Squares estimator for β, that is, derive $\widehat{\beta}$. [50%]

2017, B



9. Suppose the NHS in England introduced a £10 charge to visit a Doctor. An economist is interested in evaluating whether this policy significantly increased death rates because it dissuaded people from going to the Doctor at the first sign of ill-health.

(a) Explain why comparing the death rate in England once the £10 charge is introduced with the death rate in England prior to the introduction will not in general identify the causal effect of the policy. [10%]

(b) How could the difference-in-difference framework be applied in this case, using Wales as a control group? Write down the causal estimate that this framework would identify. How would we specify this as a regression? [30%]

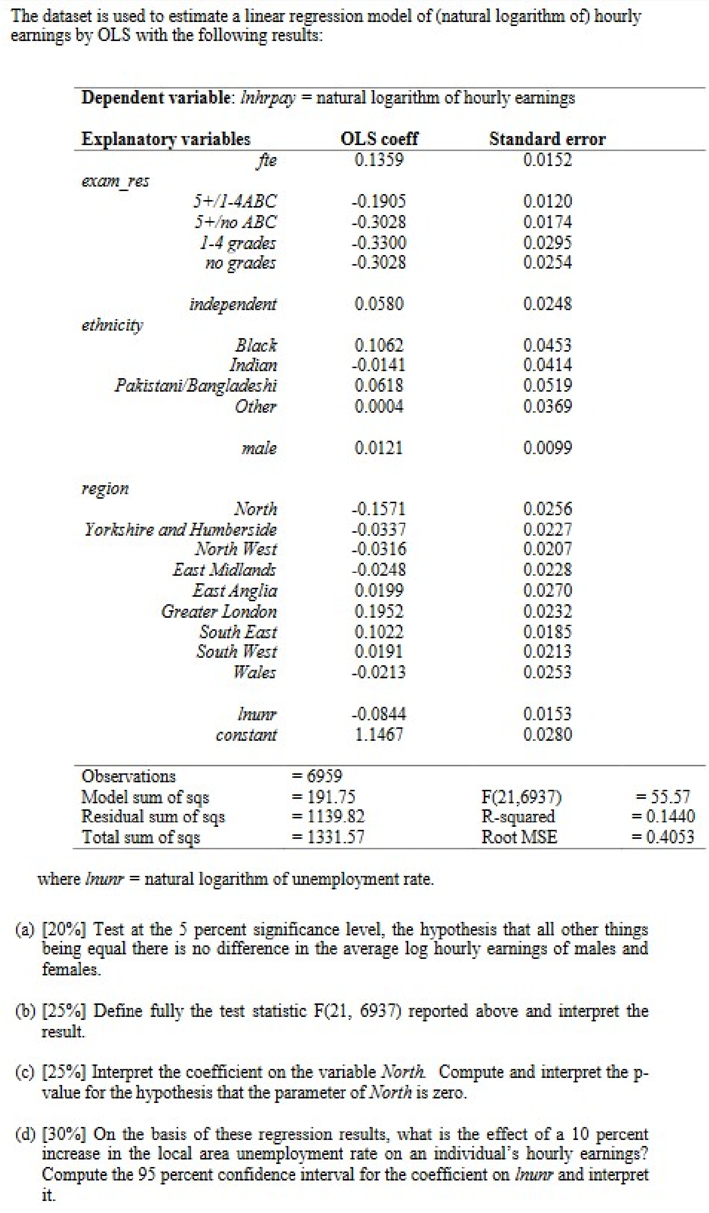
(c) What assumptions underlie the causal estimate in part (b). How plausible are they likely to be in this case? [40%]

(d) Another economist suggests that the control group used in part (b) is inappropriate and argues that we should generate a synthetic control group. What is a synthetic control group? Evaluate the pros and cons of such an approach. [20%]

2016, A

3. Our dataset contains information on the hourly earnings of 6959 young adults aged 16-17 years who were in employment in spring 1998, together with details of a range of individual characteristics as follows:

Variable	Definition	Sample mean
hrpay	Hourly earnings from employment	2.9102
fte	1 if the individual is in full-time employment; 0 otherwise.	0.1315
exam_res	Classification of results obtained in GCSE examinations taken at age 16. Variable takes value 1 if individual belongs to category; 0 otherwise.	
	5+ ABC - 5 or more GCSEs at grades A to C.	0.5841
	5+/ 1-4ABC - 5 or more GCSEs, of which 1 to 4 at grades A to C.	0.2473
	5+/ noABC - 5 or more GCSEs, with no grades A to C.	0.0973
	1-4 grades - up to 4 GCSEs, any grade.	0.0296
	no grades at GCSE.	0.0417
independent	1 if attended a private school; 0 otherwise.	0.5231
male	1 individual male; 0 otherwise.	0.4551
ethnicity	Individual's ethnic classification. Variable takes value 1 if individual belongs to category; 0 otherwise.	
	White	0.9463
	Black	0.0122
	Indian	0.0145
	Pakistani or Bangladeshi	0.0089
	Other - none of the categories identified above	0.0181
region	Individual's region of residence. Variable takes value 1 if individual belongs to category; 0 otherwise.	
	North	0.0619
	Yorkshire and Humberside	0.0822
	North West	0.1170
	East Midlands	0.0782
	West Midlands	0.1089
	East Anglia	0.0468
	Greater London	0.0815
	South East	0.2610
	South West	0.1040
	Wales	0.0585
unr	Unemployment rate in local area	4.7725



The dataset is used to estimate a linear regression model of (natural logarithm of) hourly earnings by OLS with the following results:

Dependent variable: *lnhrpay* = natural logarithm of hourly earnings

Explanatory variables	**OLS coeff**	**Standard error**
fte	0.1359	0.0152
exam_res		
5+/1-4ABC	-0.1905	0.0120
5+/no ABC	-0.3028	0.0174
1-4 grades	-0.3300	0.0295
no grades	-0.3028	0.0254
independent	0.0580	0.0248
ethnicity		
Black	0.1062	0.0453
Indian	-0.0141	0.0414
Pakistani/Bangladeshi	0.0618	0.0519
Other	0.0004	0.0369
male	0.0121	0.0099
region		
North	-0.1571	0.0256
Yorkshire and Humberside	-0.0337	0.0227
North West	-0.0316	0.0207
East Midlands	-0.0248	0.0228
East Anglia	0.0199	0.0270
Greater London	0.1952	0.0232
South East	0.1022	0.0185
South West	0.0191	0.0213
Wales	-0.0213	0.0253
lnunr	-0.0844	0.0153
constant	1.1467	0.0280

Observations	= 6959		
Model sum of sqs	= 191.75	F(21,6937)	= 55.57
Residual sum of sqs	= 1139.82	R-squared	= 0.1440
Total sum of sqs	= 1331.57	Root MSE	= 0.4053

where *lnunr* = natural logarithm of unemployment rate.

(a) [20%] Test at the 5 percent significance level, the hypothesis that all other things being equal there is no difference in the average log hourly earnings of males and females.

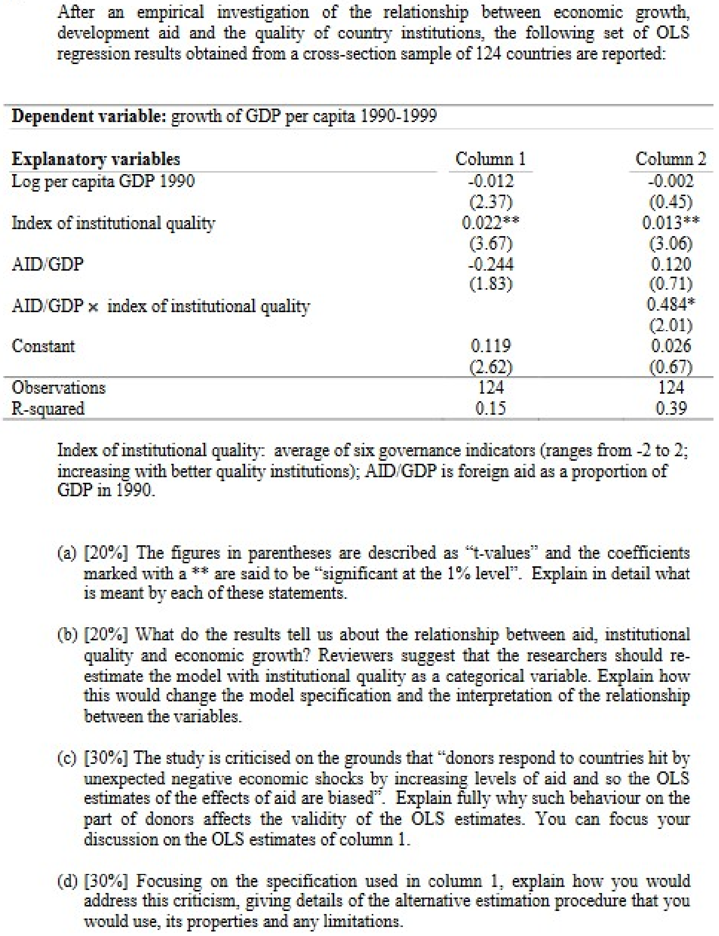
(b) [25%] Define fully the test statistic F(21, 6937) reported above and interpret the result.

(c) [25%] Interpret the coefficient on the variable *North*. Compute and interpret the p-value for the hypothesis that the parameter of *North* is zero.

(d) [30%] On the basis of these regression results, what is the effect of a 10 percent increase in the local area unemployment rate on an individual's hourly earnings? Compute the 95 percent confidence interval for the coefficient on *lnunr* and interpret it.

2016, B

7.



After an empirical investigation of the relationship between economic growth, development aid and the quality of country institutions, the following set of OLS regression results obtained from a cross-section sample of 124 countries are reported:

Dependent variable: growth of GDP per capita 1990-1999

Explanatory variables	Column 1	Column 2
Log per capita GDP 1990	-0.012	-0.002
	(2.37)	(0.45)
Index of institutional quality	0.022**	0.013**
	(3.67)	(3.06)
AID/GDP	-0.244	0.120
	(1.83)	(0.71)
AID/GDP × index of institutional quality		0.484*
		(2.01)
Constant	0.119	0.026
	(2.62)	(0.67)
Observations	124	124
R-squared	0.15	0.39

Index of institutional quality: average of six governance indicators (ranges from -2 to 2; increasing with better quality institutions); AID/GDP is foreign aid as a proportion of GDP in 1990.

(a) [20%] The figures in parentheses are described as "t-values" and the coefficients marked with a ** are said to be "significant at the 1% level". Explain in detail what is meant by each of these statements.

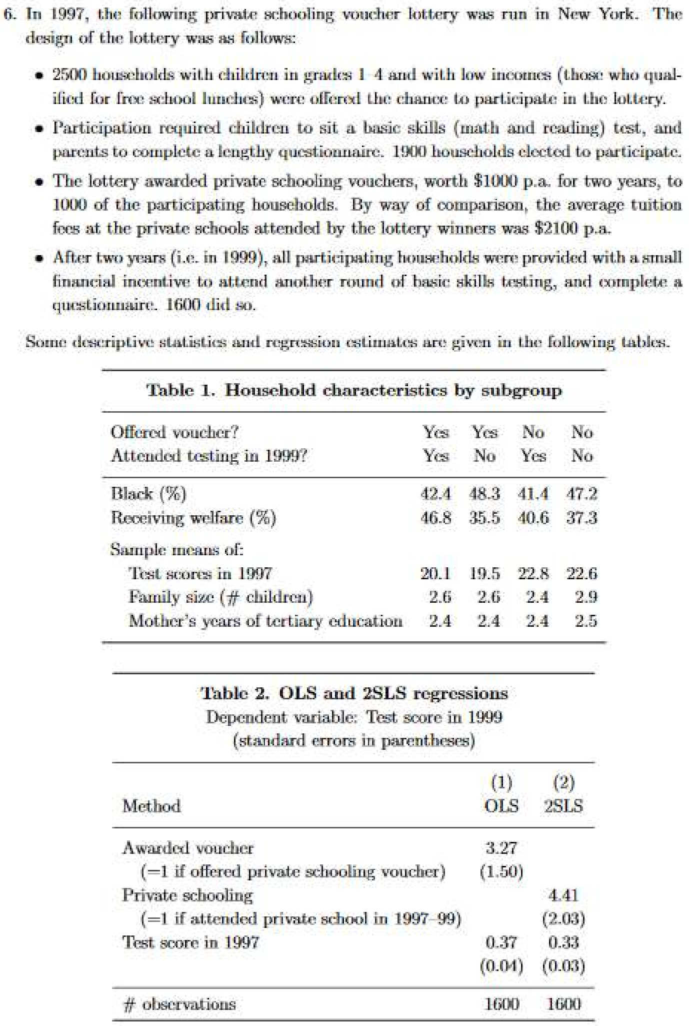
(b) [20%] What do the results tell us about the relationship between aid, institutional quality and economic growth? Reviewers suggest that the researchers should re-estimate the model with institutional quality as a categorical variable. Explain how this would change the model specification and the interpretation of the relationship between the variables.

(c) [30%] The study is criticised on the grounds that "donors respond to countries hit by unexpected negative economic shocks by increasing levels of aid and so the OLS estimates of the effects of aid are biased". Explain fully why such behaviour on the part of donors affects the validity of the OLS estimates. You can focus your discussion on the OLS estimates of column 1.

(d) [30%] Focusing on the specification used in column 1, explain how you would address this criticism, giving details of the alternative estimation procedure that you would use, its properties and any limitations.

Endogeneity

2019, B



6. In 1997, the following private schooling voucher lottery was run in New York. The design of the lottery was as follows:

   - 2500 households with children in grades 1–4 and with low incomes (those who qualified for free school lunches) were offered the chance to participate in the lottery.
   - Participation required children to sit a basic skills (math and reading) test, and parents to complete a lengthy questionnaire. 1900 households elected to participate.
   - The lottery awarded private schooling vouchers, worth $1000 p.a. for two years, to 1000 of the participating households. By way of comparison, the average tuition fees at the private schools attended by the lottery winners was $2100 p.a.
   - After two years (i.e. in 1999), all participating households were provided with a small financial incentive to attend another round of basic skills testing, and complete a questionnaire. 1600 did so.

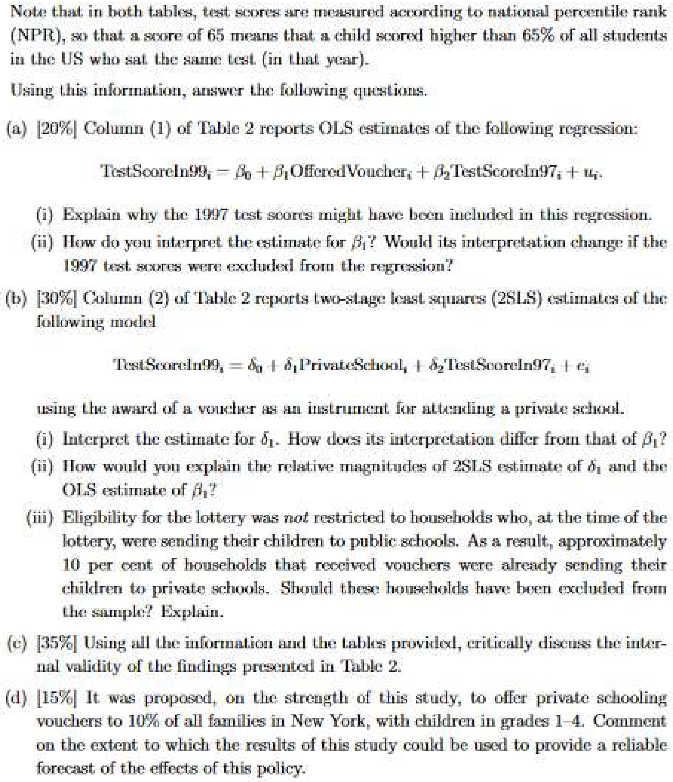
   Some descriptive statistics and regression estimates are given in the following tables.

Table 1. Household characteristics by subgroup

Offered voucher? Attended testing in 1999?	Yes Yes	Yes No	No Yes	No No
Black (%)	42.4	48.3	41.4	47.2
Receiving welfare (%)	46.8	35.5	40.6	37.3
Sample means of:				
Test scores in 1997	20.1	19.5	22.8	22.6
Family size (# children)	2.6	2.6	2.4	2.9
Mother's years of tertiary education	2.4	2.4	2.4	2.5

Table 2. OLS and 2SLS regressions
Dependent variable: Test score in 1999
(standard errors in parentheses)

Method	(1) OLS	(2) 2SLS
Awarded voucher	3.27	
(=1 if offered private schooling voucher)	(1.50)	
Private schooling		4.41
(=1 if attended private school in 1997–99)		(2.03)
Test score in 1997	0.37	0.33
	(0.04)	(0.03)
# observations	1600	1600



Note that in both tables, test scores are measured according to national percentile rank (NPR), so that a score of 65 means that a child scored higher than 65% of all students in the US who sat the same test (in that year).

Using this information, answer the following questions.

(a) [20%] Column (1) of Table 2 reports OLS estimates of the following regression:

$$\text{TestScoreIn99}_i = \beta_0 + \beta_1 \text{OfferedVoucher}_i + \beta_2 \text{TestScoreIn97}_i + u_i.$$

(i) Explain why the 1997 test scores might have been included in this regression.

(ii) How do you interpret the estimate for β_1? Would its interpretation change if the 1997 test scores were excluded from the regression?

(b) [30%] Column (2) of Table 2 reports two-stage least squares (2SLS) estimates of the following model

$$\text{TestScoreIn99}_i = \delta_0 + \delta_1 \text{PrivateSchool}_i + \delta_2 \text{TestScoreIn97}_i + e_i$$

using the award of a voucher as an instrument for attending a private school.

(i) Interpret the estimate for δ_1. How does its interpretation differ from that of β_1?

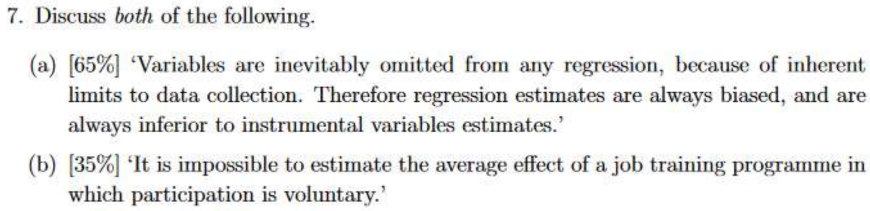
(ii) How would you explain the relative magnitudes of 2SLS estimate of δ_1 and the OLS estimate of β_1?

(iii) Eligibility for the lottery was *not* restricted to households who, at the time of the lottery, were sending their children to public schools. As a result, approximately 10 per cent of households that received vouchers were already sending their children to private schools. Should these households have been excluded from the sample? Explain.

(c) [35%] Using all the information and the tables provided, critically discuss the internal validity of the findings presented in Table 2.

(d) [15%] It was proposed, on the strength of this study, to offer private schooling vouchers to 10% of all families in New York, with children in grades 1–4. Comment on the extent to which the results of this study could be used to provide a reliable forecast of the effects of this policy.

2019, B

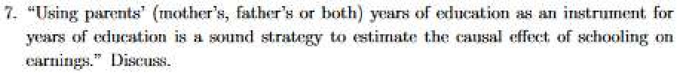


7. Discuss *both* of the following.

(a) [65%] 'Variables are inevitably omitted from any regression, because of inherent limits to data collection. Therefore regression estimates are always biased, and are always inferior to instrumental variables estimates.'

(b) [35%] 'It is impossible to estimate the average effect of a job training programme in which participation is voluntary.'

2018, B



7. "Using parents' (mother's, father's or both) years of education as an instrument for years of education is a sound strategy to estimate the causal effect of schooling on earnings." Discuss.

2017, B

6. A labour economist estimates the returns to schooling using a sample of *identical* (i.e., monozygotic) twins. Her findings are reported in the table below:

OLS regressions

Dependent variable: log(*wage*)

	(1)	(2)
education (years)	0.08	0.10
	(0.01)	(0.02)
age (years)	0.06	–
	(0.02)	
male (=1 if male, =0 if female)	0.20	–
	(0.05)	
constant	0.20	–
	(0.02)	
Number of observations	500	250
R^2	0.26	0.09

Note: Standard errors are reported in parentheses.

where in column (1) each observation is one individual and in column (2) each observation is a sibling difference. In column (1) the regression being estimated is:

$$\log(wage_{ij}) = \rho_0 + \rho_1 education_{ij} + \rho_2 age_{ij} + \rho_3 male_{ij} + u_{ij} \qquad (1)$$

where i denotes individual and j denotes sibling pair. The regression being estimated in column (2) is:

$$\Delta \log(wage_j) = \gamma_0 + \gamma_1 \Delta education_j + \Delta u_j \qquad (2)$$

where $\Delta \log(wage_j)$ is the difference in log wages within a pair of siblings and $\Delta education_j$ is the difference in years of education within a pair of siblings.

(a) Interpret the coefficient ρ_1 and compute a 99% confidence interval for its estimate and interpret it. [15%]

(b) Compare the estimates of ρ_1 and γ_1. Explain the main advantage of running regression (2) over regression (1). Is there any limitation in running regression (2)? Explain. [25%]

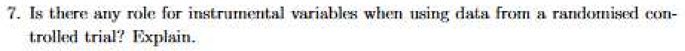
(c) Suppose that someone suggests you to run the following regression:

$$\Delta education_j = \delta_0 + \delta_1 \Delta BW_j + \Delta e_j \qquad (3)$$

where ΔBW_j is the difference in birth weight (in grams) within a pair of siblings. After carefully thinking, you run this regression and you obtain an estimate of 0.0025 for δ_1 with a corresponding standard error of 0.0010. Test whether differences in birth weight explain differences in education at the 5% significance level. Explain fully the null and alternative hypothesis, test statistic, decision rule and conclusion. If differences in birth weight are positively related to differences in log wages, what are the implications of this test in interpreting regression (2)? Explain. [40%]

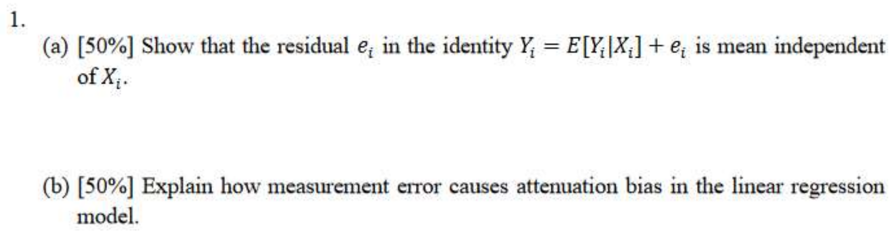
(d) Can you suggest an alternative way to estimate the causal effect of education on earnings? Explain. [20%]

2017, B



7. Is there any role for instrumental variables when using data from a randomised controlled trial? Explain.

2016, A



1.

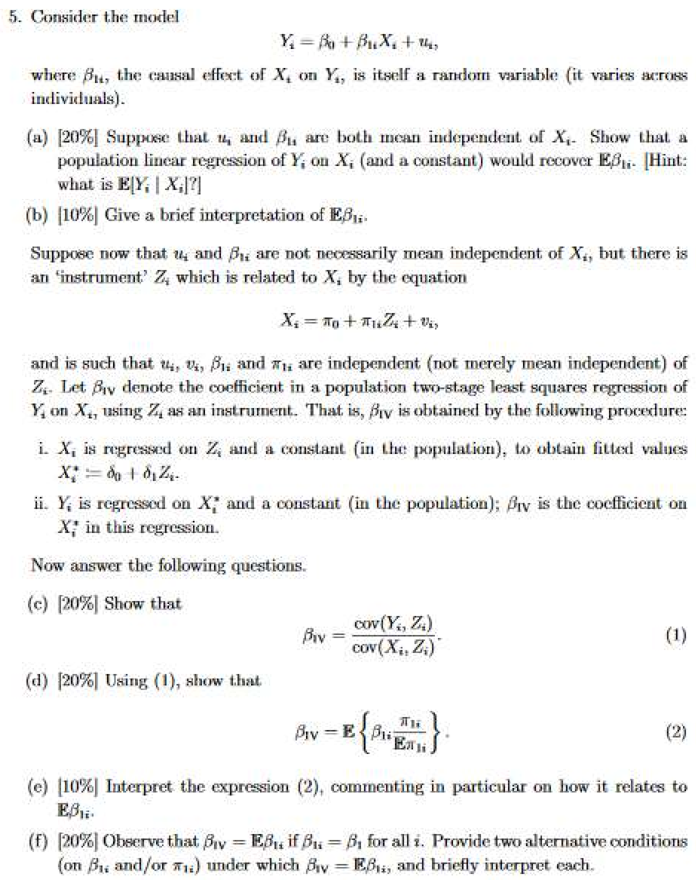
(a) [50%] Show that the residual e_i in the identity $Y_i = E[Y_i|X_i] + e_i$ is mean independent of X_i.

(b) [50%] Explain how measurement error causes attenuation bias in the linear regression model.

Kommentiert [CL1]: Done.

Heterogeneity

2019, B



5. Consider the model

$$Y_i = \beta_0 + \beta_{1i} X_i + u_i,$$

where β_{1i}, the causal effect of X_i on Y_i, is itself a random variable (it varies across individuals).

(a) [20%] Suppose that u_i and β_{1i} are both mean independent of X_i. Show that a population linear regression of Y_i on X_i (and a constant) would recover $\mathbb{E}\beta_{1i}$. [Hint: what is $\mathbb{E}[Y_i \mid X_i]$?]

(b) [10%] Give a brief interpretation of $\mathbb{E}\beta_{1i}$.

Suppose now that u_i and β_{1i} are not necessarily mean independent of X_i, but there is an 'instrument' Z_i which is related to X_i by the equation

$$X_i = \pi_0 + \pi_{1i} Z_i + v_i,$$

and is such that u_i, v_i, β_{1i} and π_{1i} are independent (not merely mean independent) of Z_i. Let β_{IV} denote the coefficient in a population two-stage least squares regression of Y_i on X_i, using Z_i as an instrument. That is, β_{IV} is obtained by the following procedure:

i. X_i is regressed on Z_i and a constant (in the population), to obtain fitted values $X_i^* := \delta_0 + \delta_1 Z_i$.

ii. Y_i is regressed on X_i^* and a constant (in the population); β_{IV} is the coefficient on X_i^* in this regression.

Now answer the following questions.

(c) [20%] Show that

$$\beta_{\mathrm{IV}} = \frac{\operatorname{cov}(Y_i, Z_i)}{\operatorname{cov}(X_i, Z_i)}. \tag{1}$$

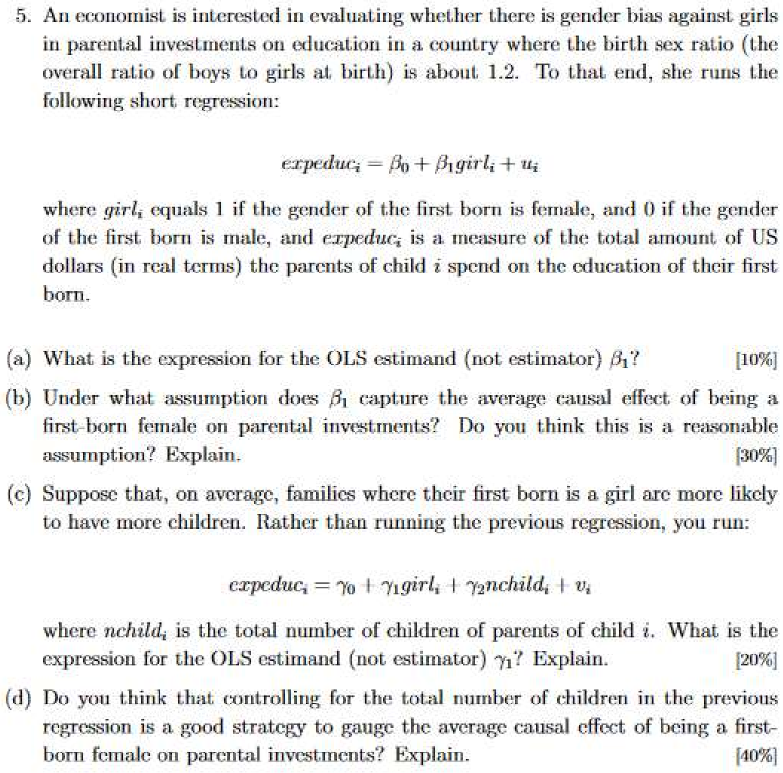
(d) [20%] Using (1), show that

$$\beta_{\mathrm{IV}} = \mathbb{E}\left\{\beta_{1i}\frac{\pi_{1i}}{\mathbb{E}\pi_{1i}}\right\}. \tag{2}$$

(e) [10%] Interpret the expression (2), commenting in particular on how it relates to $\mathbb{E}\beta_{1i}$.

(f) [20%] Observe that $\beta_{\mathrm{IV}} = \mathbb{E}\beta_{1i}$ if $\beta_{1i} = \beta_1$ for all i. Provide two alternative conditions (on β_{1i} and/or π_{1i}) under which $\beta_{\mathrm{IV}} = \mathbb{E}\beta_{1i}$, and briefly interpret each.

2017, B



5. An economist is interested in evaluating whether there is gender bias against girls in parental investments on education in a country where the birth sex ratio (the overall ratio of boys to girls at birth) is about 1.2. To that end, she runs the following short regression:

$$expeduc_i = \beta_0 + \beta_1 girl_i + u_i$$

where $girl_i$ equals 1 if the gender of the first born is female, and 0 if the gender of the first born is male, and $expeduc_i$ is a measure of the total amount of US dollars (in real terms) the parents of child i spend on the education of their first born.

(a) What is the expression for the OLS estimand (not estimator) β_1? [10%]

(b) Under what assumption does β_1 capture the average causal effect of being a first-born female on parental investments? Do you think this is a reasonable assumption? Explain. [30%]

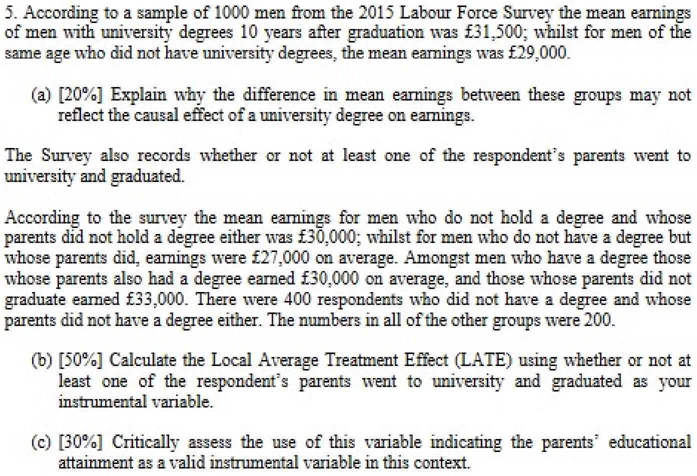
(c) Suppose that, on average, families where their first born is a girl are more likely to have more children. Rather than running the previous regression, you run:

$$expeduc_i = \gamma_0 + \gamma_1 girl_i + \gamma_2 nchild_i + v_i$$

where $nchild_i$ is the total number of children of parents of child i. What is the expression for the OLS estimand (not estimator) γ_1? Explain. [20%]

(d) Do you think that controlling for the total number of children in the previous regression is a good strategy to gauge the average causal effect of being a first-born female on parental investments? Explain. [40%]

2016, B



5. According to a sample of 1000 men from the 2015 Labour Force Survey the mean earnings of men with university degrees 10 years after graduation was £31,500; whilst for men of the same age who did not have university degrees, the mean earnings was £29,000.

(a) [20%] Explain why the difference in mean earnings between these groups may not reflect the causal effect of a university degree on earnings.

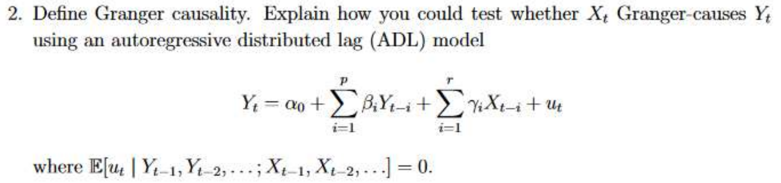
The Survey also records whether or not at least one of the respondent's parents went to university and graduated.

According to the survey the mean earnings for men who do not hold a degree and whose parents did not hold a degree either was £30,000; whilst for men who do not have a degree but whose parents did, earnings were £27,000 on average. Amongst men who have a degree those whose parents also had a degree earned £30,000 on average, and those whose parents did not graduate earned £33,000. There were 400 respondents who did not have a degree and whose parents did not have a degree either. The numbers in all of the other groups were 200.

(b) [50%] Calculate the Local Average Treatment Effect (LATE) using whether or not at least one of the respondent's parents went to university and graduated as your instrumental variable.

(c) [30%] Critically assess the use of this variable indicating the parents' educational attainment as a valid instrumental variable in this context.

Time Series
2019, A

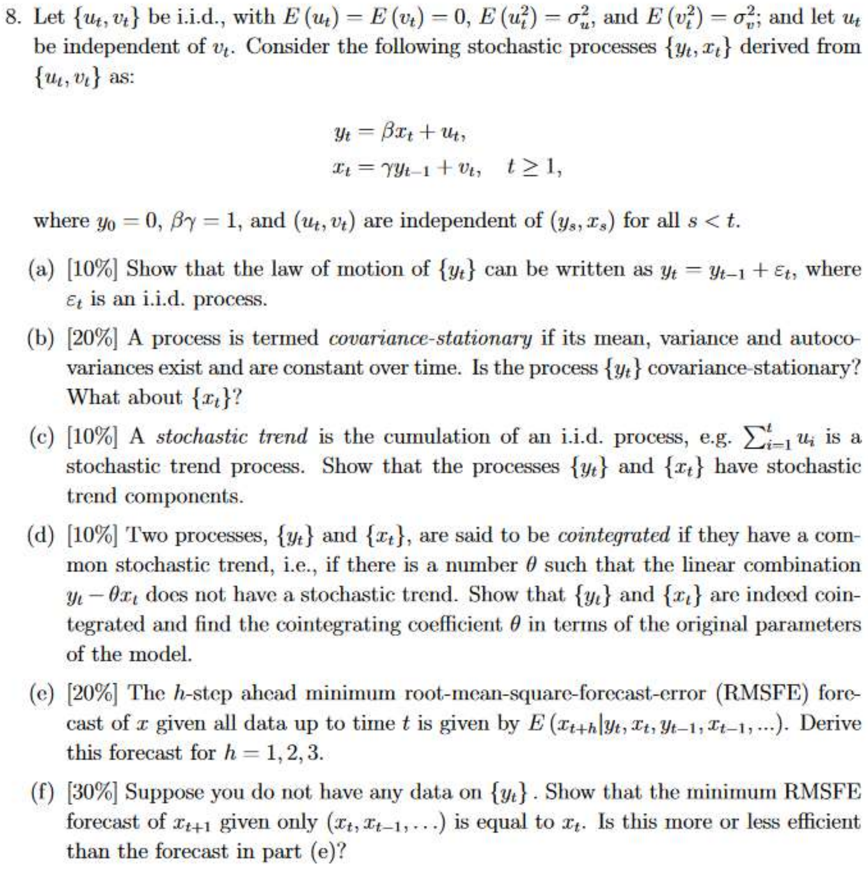


2. Define Granger causality. Explain how you could test whether X_t Granger-causes Y_t using an autoregressive distributed lag (ADL) model

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=1}^{r} \gamma_i X_{t-i} + u_t$$

where $\mathbb{E}[u_t \mid Y_{t-1}, Y_{t-2}, \ldots; X_{t-1}, X_{t-2}, \ldots] = 0$.

2019, B



8. Let $\{u_t, v_t\}$ be i.i.d., with $E(u_t) = E(v_t) = 0$, $E(u_t^2) = \sigma_u^2$, and $E(v_t^2) = \sigma_v^2$; and let u_t be independent of v_t. Consider the following stochastic processes $\{y_t, x_t\}$ derived from $\{u_t, v_t\}$ as:

$$\begin{aligned} y_t &= \beta x_t + u_t, \\ x_t &= \gamma y_{t-1} + v_t, \quad t \geq 1, \end{aligned}$$

where $y_0 = 0$, $\beta\gamma = 1$, and (u_t, v_t) are independent of (y_s, x_s) for all $s < t$.

(a) [10%] Show that the law of motion of $\{y_t\}$ can be written as $y_t = y_{t-1} + \varepsilon_t$, where ε_t is an i.i.d. process.

(b) [20%] A process is termed *covariance-stationary* if its mean, variance and autocovariances exist and are constant over time. Is the process $\{y_t\}$ covariance-stationary? What about $\{x_t\}$?

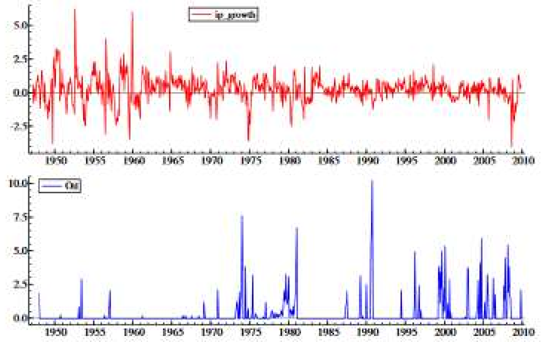
(c) [10%] A *stochastic trend* is the cumulation of an i.i.d. process, e.g. $\sum_{i=1}^{t} u_i$ is a stochastic trend process. Show that the processes $\{y_t\}$ and $\{x_t\}$ have stochastic trend components.

(d) [10%] Two processes, $\{y_t\}$ and $\{x_t\}$, are said to be *cointegrated* if they have a common stochastic trend, i.e., if there is a number θ such that the linear combination $y_t - \theta x_t$ does not have a stochastic trend. Show that $\{y_t\}$ and $\{x_t\}$ are indeed cointegrated and find the cointegrating coefficient θ in terms of the original parameters of the model.

(e) [20%] The h-step ahead minimum root-mean-square-forecast-error (RMSFE) forecast of x given all data up to time t is given by $E(x_{t+h} | y_t, x_t, y_{t-1}, x_{t-1}, ...)$. Derive this forecast for $h = 1, 2, 3$.

(f) [30%] Suppose you do not have any data on $\{y_t\}$. Show that the minimum RMSFE forecast of x_{t+1} given only $(x_t, x_{t-1}, \ldots)$ is equal to x_t. Is this more or less efficient than the forecast in part (e)?

9. Suppose you are interested in measuring the causal effect of changes in oil prices on economic activity. You are given an index of Industrial Production (IP) in the US, denoted IP_t, and a measure of oil price shocks defined as follows: O_t is the greater of zero and the percentage point difference between the oil price at date t and its maximum value during the previous 12 months. The figure below plots the time series $ip_t = 100 \times \Delta \ln IP_t$ (ip_growth) and O_t (Oil).



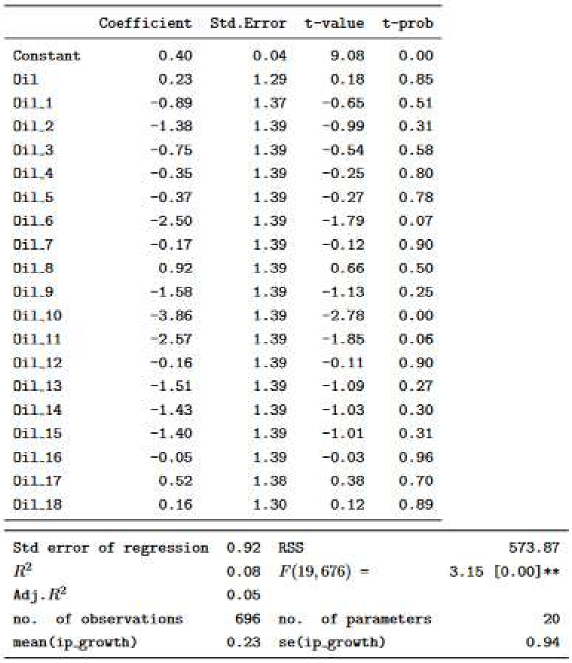
(a) [10%] Comment on the properties of the two time series plotted in the fgure. Why are so many observation of O_t at zero? Why aren't there any negative values?

(b) [20%] The table on the following page gives OLS estimates of a distributed lag model (DL) of ip_t on O_t. [Note: in the table `Oil` $= O_t$, `Oil_i` $= O_{t-i}$, and `ip_growth` $= ip_t$]. What do we need to assume in order for the coefficients of this DL model to admit a causal interpretation? Critically evaluate this assumption. Would this assumption be more plausible if you used data for the UK? [Note: oil prices are global.]

(c) [20%] Use the F statistic reported in the table (p-value in square brackets) to test whether the coefficients on O_t and its lags are all equal to zero. Give a causal interpretation of the outcome of this test.

(d) [30%] Suppose oil prices jump 10% above their previous peak and stay at this higher level (i.e. $O_t = 10$, and $O_{t+j} = 0$ for $j > 0$). What is your predicted impact on the growth rate of IP each month over the next 18 months (i.e. ip_{t+j} for $j = 1 \ldots, 18$)? What is your predicted effect on the log of IP after three months (i.e. $\ln IP_{t+3}$)?

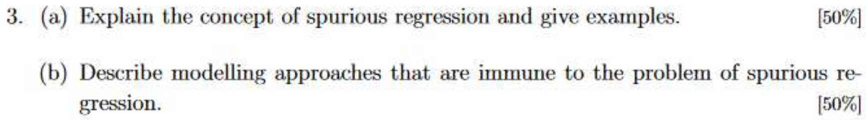
(e) [20%] What does this model predict will be the impact of an oil price shock on the growth rate of IP after two years? Suggest an alternative model that parsimoniously (i.e., with a small number of parameters) captures dynamic causal effects over long horizons.



	Coefficient	Std.Error	t-value	t-prob
Constant	0.40	0.04	9.08	0.00
Oil	0.23	1.29	0.18	0.85
Oil_1	-0.89	1.37	-0.65	0.51
Oil_2	-1.38	1.39	-0.99	0.31
Oil_3	-0.75	1.39	-0.54	0.58
Oil_4	-0.35	1.39	-0.25	0.80
Oil_5	-0.37	1.39	-0.27	0.78
Oil_6	-2.50	1.39	-1.79	0.07
Oil_7	-0.17	1.39	-0.12	0.90
Oil_8	0.92	1.39	0.66	0.50
Oil_9	-1.58	1.39	-1.13	0.25
Oil_10	-3.86	1.39	-2.78	0.00
Oil_11	-2.57	1.39	-1.85	0.06
Oil_12	-0.16	1.39	-0.11	0.90
Oil_13	-1.51	1.39	-1.09	0.27
Oil_14	-1.43	1.39	-1.03	0.30
Oil_15	-1.40	1.39	-1.01	0.31
Oil_16	-0.05	1.39	-0.03	0.96
Oil_17	0.52	1.38	0.38	0.70
Oil_18	0.16	1.30	0.12	0.89

Std error of regression	0.92	RSS	573.87
R^2	0.08	$F(19,676)$ =	3.15 [0.00]**
Adj.R^2	0.05		
no. of observations	696	no. of parameters	20
mean(ip_growth)	0.23	se(ip_growth)	0.94

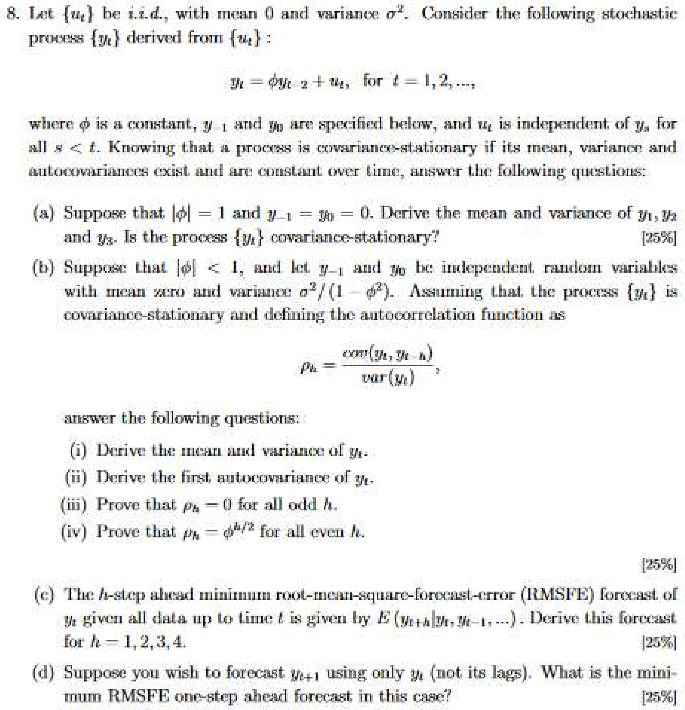
2018, A



3. (a) Explain the concept of spurious regression and give examples. [50%]

(b) Describe modelling approaches that are immune to the problem of spurious regression. [50%]

2018, B



8. Let $\{u_t\}$ be *i.i.d.*, with mean 0 and variance σ^2. Consider the following stochastic process $\{y_t\}$ derived from $\{u_t\}$:

$$y_t = \phi y_{t-2} + u_t, \quad \text{for } t = 1, 2, ...,$$

where ϕ is a constant, y_{-1} and y_0 are specified below, and u_t is independent of y_s for all $s < t$. Knowing that a process is covariance-stationary if its mean, variance and autocovariances exist and are constant over time, answer the following questions:

(a) Suppose that $|\phi| = 1$ and $y_{-1} = y_0 = 0$. Derive the mean and variance of y_1, y_2 and y_3. Is the process $\{y_t\}$ covariance-stationary? [25%]

(b) Suppose that $|\phi| < 1$, and let y_{-1} and y_0 be independent random variables with mean zero and variance $\sigma^2/(1-\phi^2)$. Assuming that the process $\{y_t\}$ is covariance-stationary and defining the autocorrelation function as

$$\rho_h = \frac{cov(y_t, y_{t-h})}{var(y_t)},$$

answer the following questions:

(i) Derive the mean and variance of y_t.

(ii) Derive the first autocovariance of y_t.

(iii) Prove that $\rho_h = 0$ for all odd h.

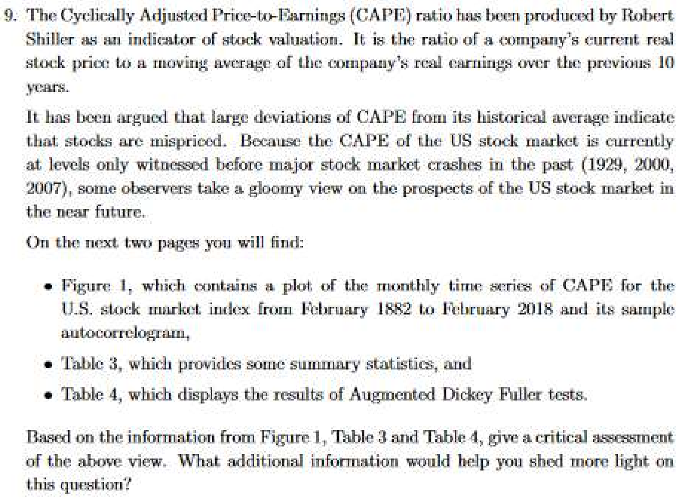
(iv) Prove that $\rho_h = \phi^{h/2}$ for all even h.

[25%]

(c) The h-step ahead minimum root-mean-square-forecast-error (RMSFE) forecast of y_t given all data up to time t is given by $E(y_{t+h}|y_t, y_{t-1}, ...)$. Derive this forecast for $h = 1, 2, 3, 4$. [25%]

(d) Suppose you wish to forecast y_{t+1} using only y_t (not its lags). What is the minimum RMSFE one-step ahead forecast in this case? [25%]

2018, B



9. The Cyclically Adjusted Price-to-Earnings (CAPE) ratio has been produced by Robert Shiller as an indicator of stock valuation. It is the ratio of a company's current real stock price to a moving average of the company's real earnings over the previous 10 years.

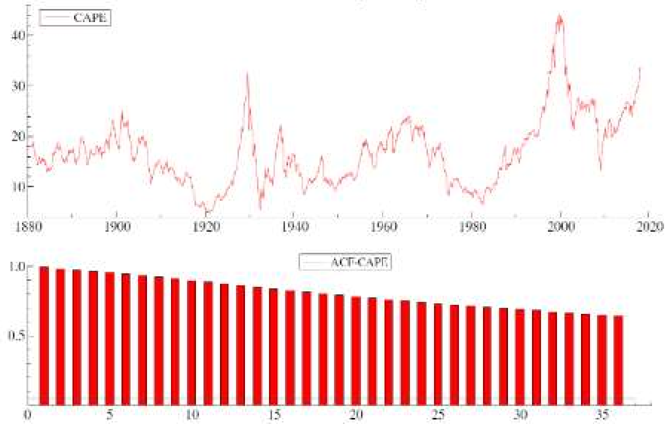
It has been argued that large deviations of CAPE from its historical average indicate that stocks are mispriced. Because the CAPE of the US stock market is currently at levels only witnessed before major stock market crashes in the past (1929, 2000, 2007), some observers take a gloomy view on the prospects of the US stock market in the near future.

On the next two pages you will find:

- Figure 1, which contains a plot of the monthly time series of CAPE for the U.S. stock market index from February 1882 to February 2018 and its sample autocorrelogram,
- Table 3, which provides some summary statistics, and
- Table 4, which displays the results of Augmented Dickey Fuller tests.

Based on the information from Figure 1, Table 3 and Table 4, give a critical assessment of the above view. What additional information would help you shed more light on this question?

Figure 1: Cyclically Adjusted Price-to-Earnings ratio for the U.S. Stock Market. Monthly averages.



Source: Robert Shiller

Table 3: Summary statistics

Sample	Mean	St. Dev.	First AC
1882(2)-2018(2)	16.830	6.7485	0.9928
1882(2)-1982(1)	14.734	4.6462	0.9890
1982(2)-2018(2)	22.639	8.1304	0.9905

2017, A

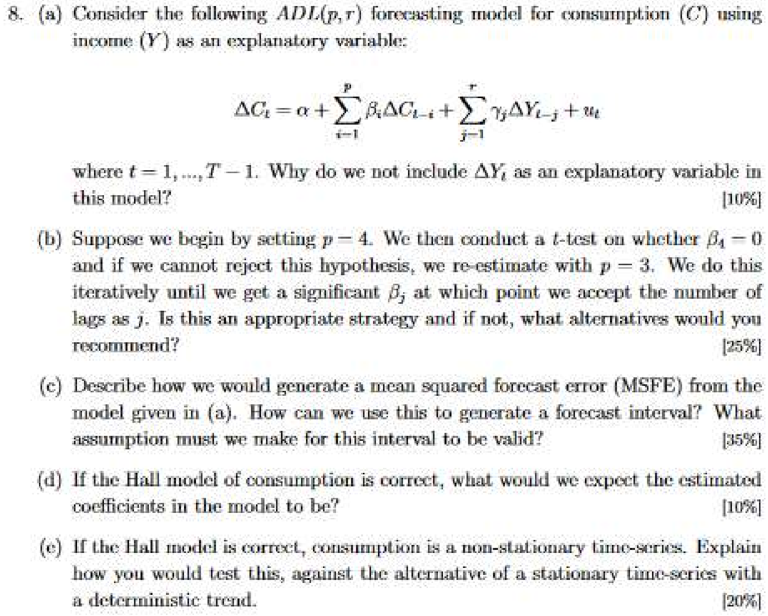
2. (a) If x_t and y_t are non-stationary time series that have a common stochastic trend, explain why there must exist a co-integrating relationship between x_t and y_t. [50%]

(b) An economist points out that since both x_t and y_t are non-stationary and co-integrated, we could estimate an equation of the form

$$\Delta x_t = \alpha + \beta \Delta y_t + \epsilon_t$$

using OLS and in that case standard statistical inference would be valid. Is this correct and are there any advantages to estimating the co-integrating relationship instead? [50%]

2017, B



8. (a) Consider the following $ADL(p,r)$ forecasting model for consumption (C) using income (Y) as an explanatory variable:

$$\Delta C_t = \alpha + \sum_{i=1}^{p} \beta_i \Delta C_{t-i} + \sum_{j=1}^{r} \gamma_j \Delta Y_{t-j} + u_t$$

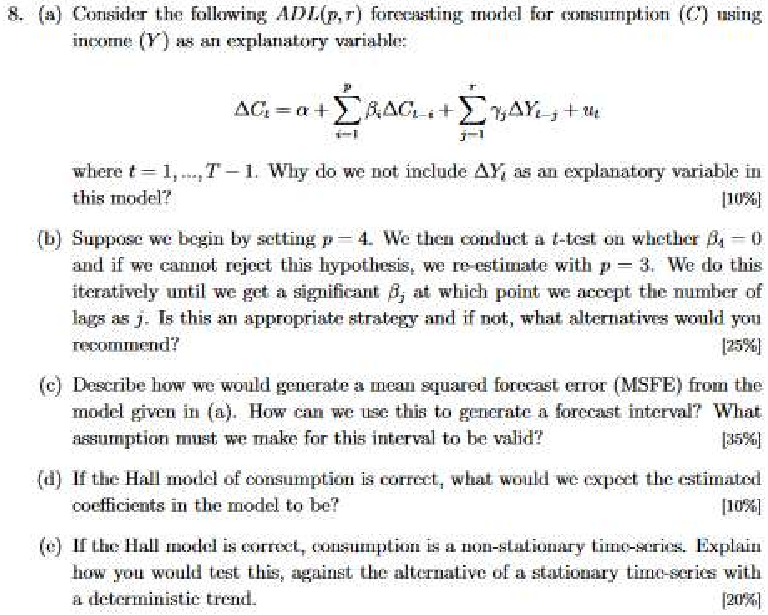
where $t = 1, ..., T-1$. Why do we not include ΔY_t as an explanatory variable in this model? [10%]

(b) Suppose we begin by setting $p = 4$. We then conduct a t-test on whether $\beta_4 = 0$ and if we cannot reject this hypothesis, we re-estimate with $p = 3$. We do this iteratively until we get a significant β_j at which point we accept the number of lags as j. Is this an appropriate strategy and if not, what alternatives would you recommend? [25%]

(c) Describe how we would generate a mean squared forecast error (MSFE) from the model given in (a). How can we use this to generate a forecast interval? What assumption must we make for this interval to be valid? [35%]

(d) If the Hall model of consumption is correct, what would we expect the estimated coefficients in the model to be? [10%]

(e) If the Hall model is correct, consumption is a non-stationary time-series. Explain how you would test this, against the alternative of a stationary time-series with a deterministic trend. [20%]



8. (a) Consider the following $ADL(p,r)$ forecasting model for consumption (C) using income (Y) as an explanatory variable:

$$\Delta C_t = \alpha + \sum_{i=1}^{p} \beta_i \Delta C_{t-i} + \sum_{j=1}^{r} \gamma_j \Delta Y_{t-j} + u_t$$

where $t = 1, ..., T-1$. Why do we not include ΔY_t as an explanatory variable in this model? [10%]

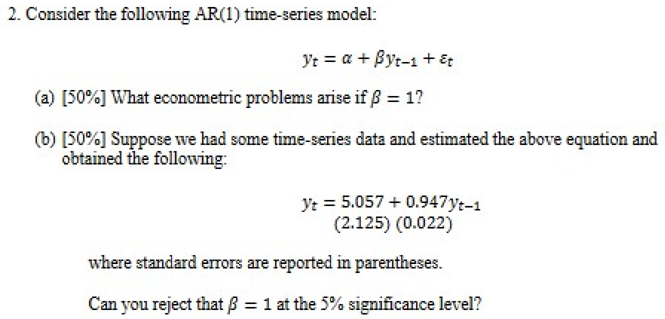
(b) Suppose we begin by setting $p = 4$. We then conduct a t-test on whether $\beta_4 = 0$ and if we cannot reject this hypothesis, we re-estimate with $p = 3$. We do this iteratively until we get a significant β_j at which point we accept the number of lags as j. Is this an appropriate strategy and if not, what alternatives would you recommend? [25%]

(c) Describe how we would generate a mean squared forecast error (MSFE) from the model given in (a). How can we use this to generate a forecast interval? What assumption must we make for this interval to be valid? [35%]

(d) If the Hall model of consumption is correct, what would we expect the estimated coefficients in the model to be? [10%]

(e) If the Hall model is correct, consumption is a non-stationary time-series. Explain how you would test this, against the alternative of a stationary time-series with a deterministic trend. [20%]

2016, A



2. Consider the following AR(1) time-series model:

$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t$$

(a) [50%] What econometric problems arise if $\beta = 1$?

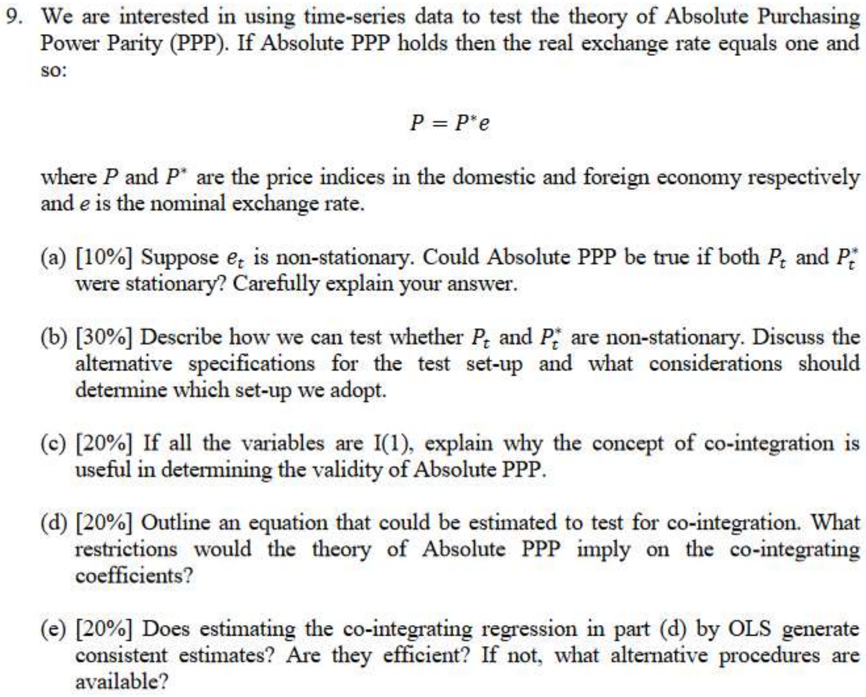
(b) [50%] Suppose we had some time-series data and estimated the above equation and obtained the following:

$$y_t = 5.057 + 0.947 y_{t-1}$$
$$(2.125)\ (0.022)$$

where standard errors are reported in parentheses.

Can you reject that $\beta = 1$ at the 5% significance level?

2016, B



9. We are interested in using time-series data to test the theory of Absolute Purchasing Power Parity (PPP). If Absolute PPP holds then the real exchange rate equals one and so:

$$P = P^{*}e$$

where P and P^{*} are the price indices in the domestic and foreign economy respectively and e is the nominal exchange rate.

(a) [10%] Suppose e_t is non-stationary. Could Absolute PPP be true if both P_t and P_t^{*} were stationary? Carefully explain your answer.

(b) [30%] Describe how we can test whether P_t and P_t^{*} are non-stationary. Discuss the alternative specifications for the test set-up and what considerations should determine which set-up we adopt.

(c) [20%] If all the variables are I(1), explain why the concept of co-integration is useful in determining the validity of Absolute PPP.

(d) [20%] Outline an equation that could be estimated to test for co-integration. What restrictions would the theory of Absolute PPP imply on the co-integrating coefficients?

(e) [20%] Does estimating the co-integrating regression in part (d) by OLS generate consistent estimates? Are they efficient? If not, what alternative procedures are available?