BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
SECOND SEMESTER – 2021-22
Course No.: ECON F342 **Course Title: APPLIED ECONOMETRICS**
Date: 14 MAY 2022 **Max. Marks: 80**
COMPREHENSIVE EXAMINATION

PART - A (CLOSED BOOK) (40 Marks)

NOTE: This is a closed-book and closed-notes exam. You may use a calculator if you wish. Answer to the point. No partial credits. A1 – A10 each question carries equal marks (2.0 each) and A11 is 20 Marks

A1. State whether the following is TRUE or FALSE, and briefly explain your answer - “If we apply a Hausman test of the hypothesis that the errors in a regression model are asymptotically uncorrelated with the regressors, we would use Instrumental variable (IV) estimation if the p-value for the test is large enough (say, greater than 10% or 20%).”

A2. “Non-invertible moving processes have no covariance generating function.”

A3. Explain what we mean by "nested models" (or "nested hypotheses"). Explain why the sum of squared residuals for a restricted model must higher than the sum of squared residuals for an unrestricted model, when the models are nested.

A4. Is it Possible to write an AR (1) process as an MA (∞) process? (∞ = infinity symbol)

A5. Explain what a "spurious regression" is? Why it occurs, and how to avoid it. Use all of the following keywords in your explanation. (Instruction: Five or six sentences in total should be enough.) Use the key words - nonstationary, OLS estimator, t-statistic, R^2, residual.

A6. Briefly explain why ordinary least squares (OLS) will not work when the dependent variable is a "count" variable.

A7. Specifically write the important assumptions of the Full Information Maximum Likelihood Method.

A8. What happens to Three Stage Least Squares (3SLS) if the random variables are contemporaneously independent? Also mention the properties of 3SLS?

A9. What are the basic similarities between the Limited Information Maximum Likelihood Method (LIML) and Two Stage Least Squares (2SLS)? (p460)

A10. Suppose that you have estimated this typical supply response model of cereal production:

$$\widehat{\log}(Q_t) = \hat{\beta}_0 + \underset{(0.02)}{0.08}\log(p_t) + \underset{(0.04)}{0.15}\log(p_{t-1}) + \underset{(0.05)}{0.10}\log(p_{t-2})$$

where is the quantity produced in million tons and the average cereal price in year t.

What is the short-term effect of a permanent cereal price increase by 20%? What will be the long-term effect of a permanent price increase by 20%? Why it is important to include the lagged variables in this model

A11. For the following multiple choice questions choose the correct best answer and put a tick (√) against that letter (A/B/C/D) and write the letter in the space provided in below table. **Corrections and overwriting answers are strictly invalid**. **(20 Marks)**

1) The Fixed Effects regression model
A. has n different intercepts.
B. the slope coefficients are allowed to differ across entities, but the intercept is “fixed” (remains unchanged).
C. has “fixed” (repaired) the effect of Heteroskedasticity.
D. in a log-log model may include logs of the binary variables, which control for the fixed effects.

2) In the Fixed Time Effects regression model, you should exclude one of the binary variables for the time periods when an intercept is present in the equation
A. because the first time period must always have excluded from your data set.
B. because there are already too many coefficients to estimate.
C. to avoid perfect multicollinearity.
D. to allow for some changes between time periods to take place.

3) Assume that for the T = 2 time periods case, you have estimated a simple regression in changes model and found a statistically significant positive intercept. This implies
A. a negative mean change in the LHS variable in the absence of a change in the RHS variable since you subtract the earlier period from the later period
B. that the panel estimation approach is flawed since differencing the data eliminates the constant (intercept) in a regression
C. a positive mean change in the LHS variable in the absence of a change in the RHS variable
D. that the RHS variable changed between the two sub periods

4) In panel data, the regression error
A. is likely to be correlated over time within an entity
B. should be calculated taking into account Heteroskedasticity but not autocorrelation
C. only exists for the case of $T > 2$
D. fits all of the three descriptions above

5) The linear probability model is
A. the application of the multiple regression model with a continuous left-hand side variable and a binary variable as at least one of the regressors.
B. an example of probit estimation.
C. another word for logit estimation.
D. the application of the linear multiple regression model to a binary dependent variable.

6) Nonlinear least squares
A. solves the minimization of the sum of squared predictive mistakes through sophisticated mathematical routines, essentially by trial and error methods.
B. should always be used when you have nonlinear equations.
C. gives you the same results as maximum likelihood estimation.
D. is another name for sophisticated least squares.

7) The rule-of-thumb for checking for weak instruments is as follows: for the case of a single endogenous regressor,
A. a first stage F must be statistically significant to indicate a strong instrument.
B. a first stage $F > 1.96$ indicates that the instruments are weak.
C. the t-statistic on each of the instruments must exceed at least 1.64.
D. a first stage $F < 10$ indicates that the instruments are weak.

8) Weak instruments are a problem because
A. the TSLS estimator may not be normally distributed, even in large samples.
B. they result in the instruments not being exogenous.
C. the TSLS estimator cannot be computed.
D. you cannot predict the endogenous variables any longer in the first stage.

9) Consider a model with one endogenous regressor and two instruments. Then the J-statistic will be large
A. if the number of observations are very large.
B. b. if the coefficients are very different when estimating the coefficients using one instrument at a time.
C. if the TSLS estimates are very different from the OLS estimates.
D. when you use Homoscedasticity-only standard errors.

10) The logic of control variables in IV regressions
A. parallels the logic of control variables in OLS
B. only applies in the case of homoscedastic errors in the first stage of two stage least squares estimation
C. is different in a substantial way from the logic of control variables in OLS since there are two stages in estimation
D. implies that the TSLS is efficient

11) The Granger Causality Test
A. uses the F-statistic to test the hypothesis that certain regressors have no predictive content for the dependent variable beyond that contained in the other regressors.
B. establishes the direction of causality (as used in common parlance) between X and Y in addition to correlation.
C. is a rather complicated test for statistical independence.
D. is a special case of the Augmented Dickey-Fuller test.

12) The AIC is a statistic
A. that is used as an alternative to the BIC when the sample size is small ($T < 50$)
B. often used to test for heteroscedasticity
C. used to help a researcher chose the number of lags in a time series with multiple predictors
D. all of the above

13) The formulae for the AIC and the BIC are different. The
A. AIC is preferred because it is easier to calculate
B. BIC is preferred because it is a consistent estimator of the lag length
C. difference is irrelevant in practice since both information criteria lead to the same conclusion
D. AIC will typically underestimate p with non-zero probability

14) The impact effect is the
A. zero period dynamic multiplier.
B. h period dynamic multiplier, $h>0$.
C. cumulative dynamic multiplier.
D. long-run cumulative dynamic multiplier

15) The interpretation of the coefficients in a distributed lag regression as causal dynamic effects hinges on
A. the assumption that X is exogenous
B. not having more than four lags when using quarterly data
C. using GLS rather than OLS
D. the use of monthly rather than annual data

16) A VAR with five variables, 4 lags and constant terms for each equation will have a total of

A. 21 coefficients.
B. 100 coefficients.
C. 105 coefficients.
D. 84 coefficients.

17) A VAR with k time series variables consists of

A. k equations, one for each of the variables, where the regressors in all equations are lagged values of all the variables
B. a single equation, where the regressors are lagged values of all the variables
C. k equations, one for each of the variables, where the regressors in all equations are never more than one lag of all the variables
D. k equations, one for each of the variables, where the regressors in all equations are current values of all the variables

18) In order to determine whether to use a fixed effects or random effects model, a researcher conducts a Hausman test. Which of the following statements is false?

A. For random effects models, the use of OLS would result in consistent but inefficient parameter estimation
B. If the Hausman test is not satisfied, the random effects model is more appropriate.
C. Random effects estimation involves the construction of "quasi-demeaned" data
D. Random effects estimation will not be appropriate if the composite error term is correlated with one or more of the explanatory variables in the model

19) The second stage in twostage least squares estimation of a simultaneous system would be to

A Estimate the reduced form equations
B Replace the endogenous variables that are onthe RHS of the structural equations with their reduced form fitted values
C Replace all endogenous variables in the structural equations with their reduced form fitted values
D Use the fitted values of the endogenous variables from the reduced forms as additional variables in the structural equations.

20) Consider the following AR(1) model with the disturbances having zero mean and unit variance
$y_t = 0.2 + 0.4\, y_{t-1} + u_t$ The (unconditional) mean of y will be given by

A. 0.2
B. 0.4
C. 0.5
D. 0.33

1	2	3	4	5	6	7	8	9	10

11	12	13	14	15	16	17	18	19	20

*******END of PART – A (CLOSED BOOK) *******

----Attempt PART-B (OPEN BOOK) ----

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
SECOND SEMESTER – 2021-22
Course No.: ECON F342 **Course Title: APPLIED ECONOMETRICS**
Date: 14 MAY 2022 **Max. Marks: 40**
COMPREHENSIVE EXAMINATION (PART B – OPEN BOOK)

Instructions: Open Book Examination. Do all the following problems. Please write legibly and organize your notation and arguments clearly. Write assumptions if any clearly. Start answering each question on a fresh page. Attempt all parts of the question at one place.

(B1 – 10 Marks) (B2-8 Marks) (B3 -10 Marks) (B4-12 Marks)

B1 This question has two parts I and II.

I. An econometrician conducted a research study and examined the effect of anti-piracy laws on music purchases. The specific objective was to determine whether COUNTRY X's passage of an anti-piracy law in 2019 affected the amount of legal music purchases. The country law allowed the government to track internet pirates who were illegally downloading music and seek criminal charges against pirates. A difference-in-difference methodology was used to estimate the effect of the law's passage on legal music purchases.

The below table reflects hypothetical averages that reflect the data in their study.

	Legal annual music sales per capita in Country X	Legal annual music sales per capita in countries other than Country X
2019	\$4.00	\$5.00
2020	\$5.50	\$6.00

a) What is the "diff-in"diff" estimate of the effect of Country X's anti-piracy law on legal music sales per capita? Suppose that, compared to other countries, Country X's economy has income growing at a slower rate. How would the slower income growth in Country X bias the diff-in-diff estimate of the law? Be sure to explain the direction of the bias (i.e. positive or negative) and how you decided on the direction.

b) Suppose that you have the following variables reflecting sales in Country X and the other countries. m_{it} is the amount of music legally purchases by person I in period t (t=2019, 2021); y2021t is a dummy variable that equals one in 2021 and is zero in 2019; Fi is a dummy variable indicating whether a person lives in Country X. Using the above variables, write out a regression equation that allows you to estimate the diff-in-diff effect of the anti-piracy law on legal music sales. If you need to create any other variables for your regression, be clear about their definition. If some of the above variables are unnecessary, don't include them in your regression. Indicate which coefficient(s) represent the "diff-in-diff" estimate of the effect of anti-piracy law on legal music sales.

II. Suppose you have panel data on people and their health expenditures, age, sex, and health insurance. (The data on total annual medical expenditures (mi) for a cross-section of people along with variables indicating each person's age (ai), a dummy variable indicating whether the person is female (fi) and a dummy variable indicating whether the person has health insurance (hi).)

Using the panel data, you estimate the following "fixed effects" model.

$$m_{it} = \beta_0 + \beta_1 h_{it} + \beta_2 a_{it} + \beta_3 f_{it} + \theta_i + u_{it}$$

where θi represent person specific fixed effects.

a) One approach to estimating this fixed effects model is to estimate a θ for each person by including a dummy variable for each person. This can be computationally burdensome since it would require 10,000 dummy variables for a sample with 10,000 people. Explain what you do with data so that it would allow estimation of the fixed effects model without including a dummy variable for each person. You must give a precise definition of how the variables are transformed to receive full credit.

b) When you estimate the fixed effects model, can the regression estimate a coefficient for the female dummy variable? Why or why not?

c) If you estimate the model with individual fixed effects added to the regression (instead of just OLS without fixed effects), do you expect the estimated coefficient on the health insurance dummy will rise or fall? Justify your prediction.

B2 This question has Two parts I and II

I. In this problem, consider and analyze U.S. employment data based on AR(p) models. Let {xt} be monthly total nonfarm payrolls, which are a key indicator on the U.S. macro economy.

The sample period covers October 1992 through November 2000 (n = 98 months). The target variable is $\Delta \ln x_t = \ln x_t - \ln x_{t-1}$, which is the log difference of nonfarm payrolls from the previous month.

See Figure 1 for time series plots of {ln xt} and {Δ ln xt}.

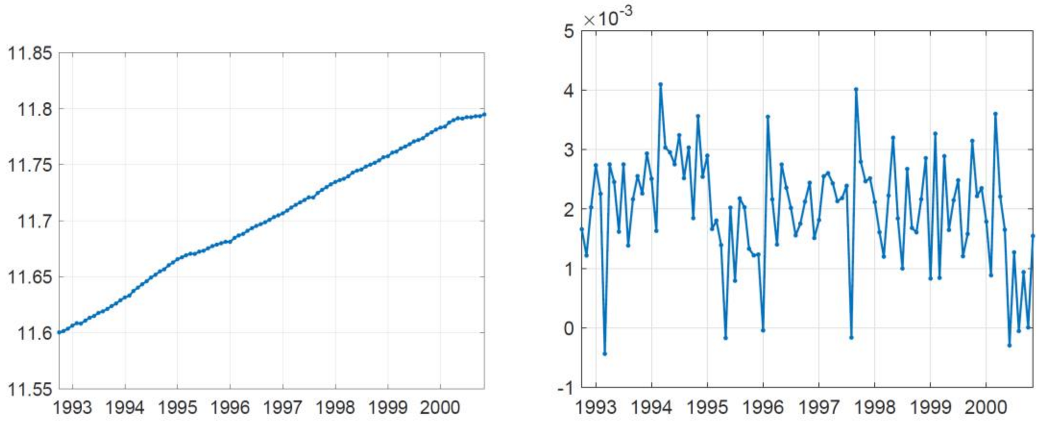
To simplify notation, redefine *yt* = Δln *xt*.

We fit AR(*p*) models for $\{y_t\}$:

$$y_t = c + \sum_{j=1}^{p} \phi_j y_{t-j} + u_t, \qquad p \in \{1, 2, 3, 4, 5, 6\},$$

where $\boldsymbol{\theta} = (c, \phi_1, \ldots, \phi_p)^\top$ are parameters to estimate. The ordinary least squares (OLS) is run to get $\hat{\boldsymbol{\theta}}_n = (\hat{c}_n, \hat{\phi}_{n1}, \ldots, \hat{\phi}_{np})^\top$.

Figure 1: Time series plots of monthly nonfarm payrolls in the U.S.



(a) $\ln x_t$ (b) $\Delta \ln x_t$

a) Akaike Information Criterion (AIC) for each lag length p is shown in Table given below. What is the optimal lag length p^* according to AIC?

b) Bayesian Information Criterion (BIC) for each lag length is also shown in Table. What is the optimal lag length p^* according to BIC?

Table 2: AIC and BIC

	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$
AIC	-13.964	-14.002	-14.013	-14.000	-13.981	-13.966
BIC	-13.911	-13.923	-13.908	-13.868	-13.822	-13.782

II. Consider the following panel data (small) set of observations for Y it.

		t		
		1	2	3
i	1	15	18	15
	2	11	17	11
	3	13	19	10

a) What is the entity demeaned value of Y for entity 3 in time 3?
b) What is the time demeaned value of Y for entity 2 in time 3?
c) What is the entity and time demeaned value of Y for entity 1 in time 3?

B3 The following model is a system of simultaneous equations to study whether the openness of the economy (open) leads to lower inflation rates (inf),

$$\begin{aligned} inf &= \delta_{10} + \gamma_{12} open + \delta_{11} \log(pcinc) + u_1 \\ open &= \delta_{20} + \gamma_{21} inf + \delta_{21} \log(pcinc) + \delta_{22} \log(land) + u_2. \end{aligned}$$

We assume that (the logarithms of) pcinc (per capita income) and land (land for farming) are exogenous in the whole exercise. The following estimations have been obtained by OLS and 2SLS .

a) Discuss the possible identification of each equation of the system, the weakness of the available instruments and perform the correspondent hypothesis tests whenever is possible.
b) Explain how you would perform a test of the exogeneity of the instruments used in the two-stage estimation for an equation and whether it is possible to apply it for the equations of the given system.
c) Test whether the effect of open over inf is lower than -0:2: If open were not a determinant of inf, (but inf is a determinant of open), explain the properties of the estimates of Output 1.

Output 1: OLS estimation using the 114 observations 1–114
Dependent variable: inf

Variable	Coefficient	Standard Dev.	*t* statistic	p-value
const	25,1040	15,2052	1,6510	0,1016
open	-0,215070	0,0946289	-2,2728	0,0250
lpcinc	0,0175673	1,97527	0,0089	0,9929

Mean of dependent variable	17,2640
Std. dev. of dependent variable	23,9973
Residual sum of squares	62127,5
Residual standard deviation ($\hat{\sigma}$)	23,6581
R^2	0,0452708
$\bar{R}^2$ corrected	0,0280685
$F(2, 111)$	2,63167
p-value for $F()$	0,0764453

Output 2: OLS estimation using the 114 observations 1–114
Dependent variable: open

Variable	Coefficient	Standard Dev.	*t* statistic	p-value
const	116,226	15,8808	7,3187	0,0000
inf	-0,0680353	0,0715556	-0,9508	0,3438
lpcinc	0,559501	1,49395	0,3745	0,7087
lland	-7,3933	0,834814	-8,8563	0,0000

Mean of dependent variable	37,0789
Std. dev. of dependent variable	23,7535
Residual sum of squares	34865,3
Residual standard deviation ($\hat{\sigma}$)	17,8033
R^2	0,453162
$\bar{R}^2$ corrected	0,438249
$F(3, 110)$	30,3855
p-value for $F()$	$<$ 0,00001

Output 3: OLS estimation using the 114 observations 1–114
Dependent variable: inf

Variable	Coefficient	Standard Dev.	*t* statistic	p-value
const	-12,615	21,0313	-0,5998	0,5498
lpcinc	0,191394	1,98158	0,0966	0,9232
lland	2,55380	1,08049	2,3635	0,0198

Mean of dependent variable	17,2640
Std. dev. of dependent variable	23,9973
Residual sum of squares	61903,2
Residual standard deviation ($\hat{\sigma}$)	23,6154
R^2	0,0487174
$\bar{R}^2$ corrected	0,0315772
$F(2, 111)$	2,84229
p-value for $F()$	0,0625432

Output 4: OLS estimation using the 114 observations 1–114
Dependent variable: open

Variable	Coefficient	Standard dev.	t statistic	p-value
const	117,085	15,8483	7,3878	0,0000
lpcinc	0,546479	1,49324	0,3660	0,7151
lland	-7,5671	0,814216	-9,2937	0,0000

Mean of dependent variable	37,0789
Std. dev. of dependent variable	23,7535
Residual sum of squares	35151,8
Residual standard deviation ($\hat{\sigma}$)	17,7956
R^2	0,448668
$\bar{R}^2$ corrected	0,438734
$F(2, 111)$	45,1654
p-value for $F()$	<0,00001

Output 5: 2SLS estimation using the 114 observations 1–114
Dependent variable: inf
Instruments: lland

Variable	Coefficient	Standard dev.	t statistic	p-value
const	26,8993	15,4012	1,7466	0,0807
open	-0,337487	0,144121	-2,3417	0,0192
lpcinc	0,375823	2,01508	0,1865	0,8520

Mean of dependent variable	17,2640
Std. dev. of dependent variable	23,9973
Residual sum of squares	63064,2
Residual standard deviation ($\hat{\sigma}$)	23,8358
$F(2, 111)$	2,62498
p-value for $F()$	0,0769352

Hausman Test –
Null hypothesis: OLS estimates are consistent
Asymptotic test statistic: $\chi^2_1 = 1{,}35333$
with p-value = 0,244697

B4 Taylor Swift wrote an op-ed article in The Wall Street Journal speaking out against pirating (i.e., illegally downloading) music and made news again recently because of her decision to remove her music from Spotify, an online streaming music website, The New York Times reported. Responding to critics, Daniel Ek, Spotify's CEO, states Spotify is ". . . a platform that protects them [music artists] from piracy and pays them for their amazing work." Musical piracy is an increasingly popular topic of debate with policymakers having to decide whether or not to change laws in order to adapt to rapidly changing technology. Further, the music industry puts more pressure on policymakers in order to protect their 15 billion dollar a year industry. As a consequence, economists have examined the impact of music piracy on music sales. Essentially, the following studies want to estimate an equation along the lines of

$$\mathbf{sales = \beta_0 + \beta_1 piracy + X\beta + u}$$

where sales is quantity of album sales, piracy is a variable representing the amount of illegally downloaded music, X is a matrix of explanatory (control) variables, and u is the idiosyncratic error term.

a) Anderson and Frenz (2010) use Canadian survey data and an instrumental variable approach to estimate a causal relationship of Equation (1). The paper finds illegal file sharing has a negligible effect (i.e., $\beta_1 \approx 0$) on music sales arguing such activities create a range of new business opportunities. Findings from the paper have been used as expert evidence in two British landmark court-cases dealing with peer-to-peer (P2P) file-sharing

i. Assuming the data used in the paper are sound. Why are the authors unable to use ordinary least squares (OLS) to obtain causal estimates? Explain

ii. The paper instruments for piracy by using internet skills (skills) whereby a respondent reports their respective level of internet expertise (sophistication). The first-stage regression is

$$\underset{\text{(robust } \hat{se})}{\widehat{piracy}} = \underset{(0.339)}{0.037} + \underset{(0.142)}{0.790}\,skills - \underset{(0.010)}{0.002}\,price - \underset{(0.146)}{0.030}\,student + \underset{(0.080)}{0.036}\,female - \underset{(0.077)}{0.249}\,region$$

What are the two important criteria for a valid ("good") instrumental variable (IV)? Can the two criteria be tested? Does the authors' IV meet these two criteria? Explain

b) Another researchers Adermon and Liang (2014) use a difference-in-differences (DID) approach, with Norway and Finland as control groups, exploiting the implementation of copyright protection reform in Sweden (i.e., treatment group) in April 2009 that suddenly increased the risk of being caught and punished for illegal file sharing. The paper's DID model is estimated with a regression to allow for controls:

$$\underset{\text{(robust } \hat{se})}{\ln(\widehat{sales})} = \underset{(0.0886)}{0.0942} + \underset{(0.0861)}{0.455}\,Sweden + \underset{(0.0960)}{0.254}\,Post2009 + \underset{(0.0722)}{0.364}\,reform + \boldsymbol{X}\hat{\beta} \qquad (2)$$

where the log of the continuous variable sales is the dependent variable, Sweden is the treated group dummy, Post2009 is a dummy for the period after the April 2009 policy change, reform is the dummy variable indicating treatment (i.e., Sweden_Post2009), and X is a matrix of explanatory (control) variables (Note: $\hat{\beta}$ are the respective estimates (not shown.). What is the key assumption underlying the validity of a DID estimate to be considered causal? Can it be tested? Explain. What is the estimated effect of Sweden's copyright protection reform? Use the coefficient and comment on statistical significance at the 95% confidence level.