Student Id._____

O.P. JINDAL GLOBAL
Institution of Eminence Deemed to be
UNIVERSITY
A Private University Promoting Public Service

Jindal School of Government and Public Policies
End-term Examination

Course Name : Statistical Methods and Data Analysis
Course Code : SMDA-6004
Programme : MAPP ONLINE
Duration : 48 HOURS
Maximum Marks : 50 (Answer **ANY FIVE** six-point questions, **ANY ONE** eight-point question, **AND** the twelve-point question)

This question paper has **(FIVE)** pages (including this page).

Instructions to students:

1. The datasets required for QUESTIONS 3, 4&7, 5&6 are attached.
2. Students undertaking the examination are requested to adhere to the University norms related to online examinations.

This is an Open Book examination.

Warning: Plagiarism in any form is prohibited. Anyone found using unfair means will be penalized severely.

Question 1 (8 marks)

The following table lists the number of agricultural holdings by land size class in India as per the 2001 census. Answer the following questions.

(i) Calculate the mean, standard deviation, median, and mode of the distribution of land holdings.
(ii) Based on your answers to (i), comment on the state of the distribution of land holdings in India in 2001.

Land size class	Number of agricultural holdings (in thousands)
0-1 hectare	75390
1-2 hectare	22687
2-5 hectare	16639
5-10 hectare	3948
10-20 hectare	1004
20-50 hectare	226

Question 2 (6 marks)

This question requires the use of the wine dataset available through the R package "ordinal" (download and call the package and type "data(wine)"). The dataset is adopted from an experiment on factors determining the bitterness of wine. Two treatment factors (temperature and contact) each have two levels. Temperature and contact between juice and skins can be controlled when crunching grapes during wine production. Nine judges each assessed wine from two bottles from each of the four treatment conditions. Hence there are 72 observations in all. The variables in the dataset are:

response scorings of wine bitterness on a 0—100 continuous scale.
rating ordered factor with 5 levels; a grouped version of response.
temp factor with two levels ("warm" and "cold")
contact factor with two levels ("no" and "yes").
bottle factor with eight levels.
judge factor with nine levels.

(i) Generate the appropriate bivariate plot of response by temp. Copy and paste your R code and output.
(ii) Based on the graphical illustration in (i), would you say that temperature matters in scoring of wine bitterness? Explain.

Question 3 (6 marks)

Consider the attached dataset reported in Column 1 in Table 1 in Doyle's study of remittances in Latin America.

(i) Use the appropriate test of significance to assess whether the average percentage of those who believed income distribution is fair differs systematically between recipients and non-recipients of remittances. Copy and paste your R code and R output for this step and provide comprehensive interpretation based on the output.
(ii) Based on your answer to (i), how would you describe the relationship between receiving remittances and attitude towards welfarism in remittance-receiving countries? Explain.

Question 4 (6 marks)

This exercise requires the use of the attached Titanic dataset from Varian's (2014) study of big data econometrics.
VARIABLE DESCRIPTIONS:

pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survived	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat

body	Body Identification Number
home.dest	Home/Destination

(i) Assess the relationship between the variable sex and the variable survived using the appropriate measure of association and test of significance. Copy and paste your R code and R output for this step and provide comprehensive interpretation based on the output.
(ii) Based on your response to (i), how would you describe the relationship between gender and the propensity of death during the Titanic accident?

Question 5 (6 marks)
You will need the attached excel file "LupPon.xlsx" for answering this question. This is the Replication data for Table 2, "Determinants of Redistribution", in Lupu, Noam and Jonas Pontusson. 2011. "The Structure of Inequality and the Politics of Redistribution." *American Political Science Review* 105(2): 316-336. Data structure is panels of OECD countries from 1969 to 2005. Data contains measurements of redistribution, various summaries of the earnings distribution, and controls. For details of the variables see Appendix of the article, which is enclosed with this question paper.

(i) Construct the variable skew, which is a measure of how skewed the income distribution is, by dividing the variable ratio9050 by the variable ratio5010. Generate the appropriate bivariate plot of redist, a measure of redistribution from rich to poor, by skew. Copy and paste your R code and R output.
(ii) Assess the relationship between the variable redist and the variable skew using the appropriate measure of association and test of significance. Copy and paste your R code and R output for this step and provide comprehensive interpretation based on the output.

Question 6 (12 marks)
You will need the attached excel file "LupPon.xlsx" for answering this question. This is the Replication data for Table 2, "Determinants of Redistribution", in Lupu, Noam and Jonas Pontusson. 2011. "The Structure of Inequality and the Politics of Redistribution." *American Political Science Review* 105(2): 316-336. Data structure is panels of OECD countries from 1969 to 2005. Data contains measurements of redistribution, various summaries of the earnings distribution, and controls. For details of the variables see Appendix of the article, which is enclosed with this question paper.

(i) Construct the variable skew, which is a measure of how skewed the income distribution is, by dividing the variable ratio9050 by the variable ratio5010. Run a bivariate regression of the variable redist, a measure of redistribution from rich to poor, on the variable skew. Copy and paste your R code and R output and provide comprehensive interpretation based on the output.
(ii) Run a regression of redist on skew, but now include the variables turnout (voter turnout in the most recent national election), union (annual net union density), and unempl (annual rate of unemployment) as control variables. Is the regression coefficient of skew in the multivariate regression different from the regression coefficient of skew in the bivariate regression? Why? Copy and paste your R code and R output and provide comprehensive interpretation based on the output.
(iii) Using the appropriate test assess whether the multivariate regression model you estimated in part (ii) violates the assumption of homoscedastic errors. Copy and paste your R code and R output and provide comprehensive interpretation based on the output.
(iv) If you detect heteroscedasticity in part (iii), use the appropriate correction to get heteroscedasticity-consistent standard errors. Is there is a significant change in your results from part (ii) when you correct for heteroscedasticity? Copy and paste your R code and R output for this step and provide comprehensive interpretation based on the output.

Question 7 (6 marks)

This exercise requires the use of the attached Titanic dataset from Varian's (2014) study of big data econometrics.

Run a regression of survived on age using the appropriate model. Copy and paste your R code and R output for this step and provide comprehensive interpretation based on the output.

Question 8 (6 marks)

This question requires the use of the CASchools dataset available through the R package "AER" (download and call the package and type "data(CASchools)").

The data frame contains 420 observations on 14 variables.

district District code
school School name
county County name
grades grade span of district
students Total enrollment
teachers Number of teachers
calworks Percent qualifying for CalWorks (income assistance)
lunch Percent qualifying for reduced-price lunch
computer Number of computers per classroom
expenditure Expenditure per student.
income District average income (in USD 1,000).
english Percent of English learners.
read Average reading score.
math Average math score.

(i) Using the appropriate model run a regression of average reading score on expenditure per student. Copy and paste your R code and output and provide comprehensive interpretation based on the output.
(ii) How will your answer to (i) change if you run a regression modeling the effect of expenditure on reading scores using a quadratic function? Explain. Copy and paste your R code and output.

Question 9 (6 marks)

This question requires the use of the Home Mortgage Disclosure Act dataset available through the R package "AER" (download and call the package and type "data(HMDA)").
HMDA is a a data frame containing 2,380 observations on 14 variables.

deny Was the mortgage denied?
pirat Payments to income ratio.
hirat Housing expense to income ratio.
lvrat Loan to value ratio.
chist Credit history: consumer payments.
mhist Credit history: mortgage payments.
phist Public bad credit record?
unemp 1989 Massachusetts unemployment rate in applicant's industry.
selfemp Is the individual self-employed?
insurance Was the individual denied mortgage insurance?
condomin Is the unit a condominium?

afam Is the individual African-American?
single Is the individual single?
hschool Does the individual have a high-school diploma?

(i) Construct a classification tree predicting deny using pirat and chist as predictors. Copy and paste your R code and output.
(ii) What is the misclassification rate of the tree generated in (i)?

Question 10 (8 marks)

This question requires the use of the Home Mortgage Disclosure Act dataset available through the R package "AER" (download and call the package and type "data(HMDA)").
HMDA is a a data frame containing 2,380 observations on 14 variables.
deny Was the mortgage denied?
pirat Payments to income ratio.
hirat Housing expense to income ratio.
lvrat Loan to value ratio.
chist Credit history: consumer payments.
mhist Credit history: mortgage payments.
phist Public bad credit record?
unemp 1989 Massachusetts unemployment rate in applicant's industry.
selfemp Is the individual self-employed?
insurance Was the individual denied mortgage insurance?
condomin Is the unit a condominium?
afam Is the individual African-American?
single Is the individual single?
hschool Does the individual have a high-school diploma?

(i) Split your dataset 60:40 into a training and testing sample. Use the training sample to grow a random forest predicting deny using pirat and chist as predictors. Copy and paste your R code for this step.
(ii) Apply the predictions from the random forest generated in (i) on the test sample. What is the misclassification rate?

APPENDIX: DEFINITIONS AND VARIABLES AND SOURCES

Variable	Definition	Source
50–10 ratio	Earnings of the worker with a median income as a share of the earnings of a worker in the 10th percentile of the earnings distribution	OECD (2007)
90–10 ratio	Earnings of a worker in the 90th percentile of the earnings distribution as a share of the earnings of a worker in the 10th percentile of the earnings distribution	OECD (2007)
90–50 ratio	Earnings of a worker in the 90th percentile of the earnings distribution as a share of the earnings of the worker with a median income	OECD (2007)
Elderly	Proportion of population older than 64	Armingeon et al. (n.d.)
Female labor	Proportion of working-age women in the labor force	OECD Labour Force Statistics
Globalization	Index of globalization constructed with principal component analysis of trade, FDI, portfolio investment, income payments to foreign nationals, hidden import barriers, mean tariff rate, taxes on international trade, and capital account restrictions	Dreher (2006)
GDP growth	Annual percentage change in GDP	World Development Indicators
Immigration	For Australia, Canada, and the U.S., proportion of the population that is foreign born; for other countries, noncitizens as proportion of the population	Dancygier data set and World Development Indicators
Partisanship	Index of the partisan left-right "center of gravity" of the cabinet based on the average of three expert classifications of government parties' placement on a left-right scale and weighted by their decimal share of cabinet portfolios (the index goes from left to right and is standardized here to vary between 0 and 1)	Cusack and Engelhardt (2002)
Redistribution	Percentage change in Gini coefficients as we move from gross market income (i.e., household income before taxes and transfers) to disposable income (i.e., income after taxes and transfers)	Kenworthy data set and Mahler and Jesuit (2006)
Skew	Ratio of the 90–50 ratio to the 50–10 ratio	OECD (2007)
Social spending	Total nonelderly government transfers (in percent GDP)	OECD Social Expenditure Database
Support for redistribution	Proportion of middle-income respondents to ISSP surveys (identified as those falling in the middle third of the distribution of respondent household incomes) who said they "strongly agree" or "agree" with the statement, "It is the responsibility of the government to reduce the differences in income between people with high incomes and those with low incomes"	ISSP survey modules (Environment 1993, 2000; Role of Government 1985, 1990, 1996; Social Inequality 1987, 1992, 1996); ESS 2002 and 2004
Unemployment	Annual rate of unemployment	Armingeon et al. (n.d.)
Unionization	Annual net union density	Visser (2009)
Vocational training	Enrollments in vocational training programs in percent of secondary school enrollments	Iversen (2005) and UNESCO database
Voter turnout	Turnout (as a percentage of eligible voters) in the most recent national election for each year	Armingeon et al. (n.d.)

ESS, European Social Survey; FDI, foreign direct investment; GDP, gross domestic product; ISSP, International Social Survey Programme; OECD, Organisation for Economic Co-operation and Development; UNESCO, United Nations Educational, Scientific and Cultural Organization.