Question 1

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

**Task:** Generate the appropriate univariate plot of rating. Copy and paste your R code and output.

**Question**: What are measures of central tendency and dispersion of rating? Explain how you arrived at your answer.

Question 2

Consider Column 2 from Table 1 from Doyle’s study of remittances in Latin America (see below).

**Task**: Use the appropriate test of significance to assess whether the average percentage of those who believed avoiding taxes is acceptable differs systematically between recipients and non-recipients of remittances. Provide a comprehensive interpretation based on the output.

**Question**: Based on your answer to the problem statement in the Task section, how would you describe the relationship between receiving remittances and attitude towards welfarism in remittance-receiving countries? Explain.

**Submissions:** R Notebook in HTML format. Word document with your interpretation.

|  |
| --- |
| **Individual Preferences of by Remittance Recipients and Non-Recipients – Latinobarometro 2009** |

|  |  |
| --- | --- |
|  | **Percentage Who BelieveAvoiding Tax is Acceptable** |

|  | **Receive remittances** | **Do not receive remittances** |
| --- | --- | --- |
| Argentina | 32 | 23 |
| Bolivia | 33.3 | 33.4 |
| Brazil | 70.9 | 26.3 |
| Colombia | 56.6 | 36.2 |
| Costa Rica | 58.1 | 39.9 |
| Chile | 36.3 | 28.5 |
| Ecuador | 60.5 | 27.8 |
| El Salvador | 47.7 | 30.3 |
| Guatemala | 40.3 | 41.3 |
| Honduras | 35.4 | 35.9 |
| Mexico | 52.5 | 32.9 |
| Nicaragua | 35.5 | 34.4 |
| Panama | 67.4 | 41.1 |
| Paraguay | 24.7 | 28.6 |
| Peru | 29.6 | 32.5 |
| Uruguay | 18.6 | 22.5 |
| Venezuela | 31.7 | 22.7 |
| Dominican Republic | 57.3 | 45.9 |

Question 3

This exercise requires the use of the Titanic dataset from Varian’s (2014) study of big data econometrics. The dataset is provided in the sandbox.

**Task:**

Assess the relationship between the variable sex and the variable survived using the appropriate measure of association. Provide comprehensive interpretation based on the output.

**Question:**

Based on your response to the above task, how would you describe the relationship between gender and the propensity of death during the Titanic accident? Explain.

Question 4

You will need the enclosed stata file "LupPon\_APSR.dta" 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 also enclosed. **Task:** Generate the appropriate bivariate plot of redist by skew (ratio9050 divided by ratio5010). Superimpose a regression line on the plot.

**Question:** What is value of the slope of the regression line? Provide comprehensive interpretation based on the output.

**Submissions:** R Notebook in HTML format. Word document with your interpretation.

Question 5

You will need the enclosed stata file "LupPon\_APSR.dta" 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 the Appendix of the article, which is also enclosed. **Task 1:** Using the appropriate model, estimate a regression of redistribution on skewness of the income distribution (ratio90-50 divided by ratio50-10) with turnout in the most recent national election, annual net union density, and the annual rate of unemployment as control variables. How does the regression coefficient from the multivariate regression differ from the regression coefficient of the bivariate regression identified in Question 4 (ii)? Which coefficient is higher? Why? Provide a comprehensive interpretation based on the output.

**Task 2:** Using the multivariate regression output in Task 1, show if there is a violation of OLS assumptions. Provide comprehensive interpretation based on the output.

**Task 3**: Explain how the results from Task 1 change when you correct for the potential violation in OLS assumptions identified in Task 2. Provide a comprehensive interpretation based on the output.

**Submissions:** R Notebook in HTML format. Word document with your interpretation.

Question 6

You will need the enclosed excel file "AfricaKF" for answering this question. The file has capital flight data for a representative sample of 30 African countries from 1970-2015 collected by Leonce Ndikumana and James Boyce, researchers at the University of Massachusetts, Amherst.

The description of the variables in the dataset are as follows:

| **Variable Name** | **Label** |
| --- | --- |
| countryname | country name |
| countrycode | country alpha code (World Bank) |
| ifscode | coutry IFS code (IMF) |
| year | year |
| kf | capital flight, nominal |
| realkf | capital flight, constant 2015 $ |
| stockkf | stock of capital flight |
| misinv | net trade misinvoicing, nominal |
| realmisinv | net trade misinvoicing,constant 2015$ |
| usgdpd\_2015 | US GDP deflator, base 2015=100 |
| gdpnominal | Nominal GDP (million $) |
| gdpconstant | Real GDP (million, constant 2015 $) |

**Notes:**

1. Latest download of input variables from the various sources in February 2018
2. All values are in million dollars
3. 'real' = constant value at 2015 dollars

Capital flight series computed by: Leonce Ndikumana and James K Boyce Political Economy Research Institute University of Massachusetts, Amherst Accessible online at: https://www.peri.umass.edu/capital-flight-from-africa Release: May 2018

**Task 1:** Using the appropriate model, estimate a regression of real capital flight on stock of capital flight, real net trade misinvoicing, and real GDP. Provide comprehensive interpretation based on the output.

**Task 2:** Using the appropriate test, identify potential violations of OLS assumptions in Task 1. Provide comprehensive interpretation based on the output.

**Task 3:** Explain how the results from Task 1 change when you account for the potential violations of OLS assumptions identified in Task 2. Provide comprehensive interpretation based on the output.

Question 7

his 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 five 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.

**Task**: Run a regression of rating on temp and contact using the appropriate model. Provide comprehensive interpretation based on the output.

**Submissions:** R Notebook in HTML format. Word document with your interpretation.

Question 8

This question requires the use of the CASchools dataset available through the R package "AER". This is already downloaded in the lab sandbox. Select the package on the packages tab and type "data(CASchools)".

The data frame contains 420 observations on 14 variables.

| **header** | **header** |
| --- | --- |
| 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 |

**Task 1:** 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.

**Task 2:** How will your answer to Task 1 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.