|  |  |
| --- | --- |
| **Learning objectives:**   1. **R with lists and lapply()** 2. **Correlation and elementary modelling using linear regression** | |
| Recall how apply() works. In this case, it is taking the mean of the rows and columns of a matrix. |  |
| lapply() can be applied to lists. In this example, we return the means of each vector in the list |  |
| We can embed a custom function in the lapply() function. In this case, we return the first elements of each vector in the list |  |
| In this example, we divide the first element in each vector in the list by the mean of that vector. |  |
| Download these data to your computer and save as a csv file:  <https://drive.google.com/file/d/1nGaUQ80FH86PPvXyjdpOkmnELKw4iy8N/view?usp=sharing> | |
| Import the data using read.csv() | |
| *When modelling data, you should have knowledge about the subject matter, and a clear motivation for creating the model in the first place.*  *To put things into context, we are trying to understand the main factors that could explain the differences in per capita CO2 emissions between countries. We will refer to CO2 emissions as our* ***dependent variable****. We will refer to other model variables as the* ***independent variables****. Independent variables include access to electricity (% of population), fossil fuel energy consumption (% of total), energy use (kg of oil equivalent per capita),population density (people per sq. km of land area), adjusted net national income per capita (current US$)and oil rents (% of GDP). Our working null hypothesis is that the independent variables have no independent association with the level of CO2.* This means that we should expect all model coefficients to equal zero if this null hypothesis is true. | |
| Use the following code to calculate some summary statistics: | |
| Use the following code to calculate histograms for the same three variables: | |
| **Q1. Briefly reflect on whether or not any of these three variables have a ‘Normal’ distribution. Use both the summary statistics and the histograms to make your case.** | |
| Use this code to transform these variables into their natural log. This may help ‘normalize’ these variables to some degree, which is sometimes helpful for analysis. | |
| Create a new set of histograms of the log-transformed variables |  |
| *You’ll note that some of the values for fossil fuels are ‘0’. These are missing data. Not only does this produce invalid values when using the log() function, but it also makes little sense to include these data in the analysis.*  *There are many ways to eliminate missing data. One is to delete all records with one or more missing elements. Another is to fill in missing values (‘imputation’) with sensible values. This method can work well for time series data, but in this case would be hard to justify. Another option is to avoid variables with missing data. Here, let’s try the third approach and focus on* ***energy and CO2 emissions.*** | |
| Another problem with this data set is that one record corresponds to a geographical region comprised of countries (‘Small states’).  **Q2. Write code to delete the record corresponding to ‘Small states’. Note: you must write this code in base R without using any external libraries or functions.** | |
| Now we’ll calculate the correlation between the variables:    **Q3. Do a little independent searching on the web, and explain (in no more than three sentences, and in simple language) the difference between the Pearson and Spearman correlation coefficients.** | |
| **Q4. Make a scatter plot (using plot()) of the log energy and log CO2 emissions variables. Be sure to label the plot.** | |
| There appears to be a positive association between the log of energy consumption and the log of CO2 emissions |  |
| Linear regression modelling is a method for predicting a continuous numerical outcome variable (also known as the **dependent** variable) as a function of one or more predictor variables (also known as **independent** variables or features). The goal of regression analysis is to find an equation that describes the relationship between the independent variable(s) and the dependent variable. This equation can then be used to make predictions about the dependent variable based on new values of the independent variables.  **Prediction** is the primary purpose and value of regression modelling. Sometimes regression models are created to ‘explain’ associations between variables, but can be less methodologically rigorous. We’ll discuss this in class. | |
| In simple regression (with one dependent and one independent variable) we can turn the resulting equation into a line. Let’s use regression to find a line that represents this association in mathematical form. This will give us a formula for prediction. |  |
| To the right are the parameters estimated using regression. The intercept (0.7136) and the coefficient for the log of energy (1.0342). From this, the equation for the line is:  Log of CO2 emissions = 0.7136 + 1.0342\*log of energy use  One way to interpret this is that for very 1 unit increase in the log of energy use we expect to see a 1.0342 unit increase in the log of CO2 emissions  Because these are both logged variables, we can also interpret as follows. For every 1% increase in energy use, we get a 1.0342% increase in CO2 emissions. |  |
| Note: The null hypothesis of the regression coefficients in regression is that the true unknown coefficients in the regression equation (the population parameters) equal 0. If the p-value is small, we might conclude that the null hypothesis is unlikely to be true. | |
| It is useful to get some more information from the modelling process to understand how useful this model is |  |
|  | |
| The problem with using logged variables is that they are sometimes hard to interpret in useful ways. | |
| **Q5. Use the lm() function to predict CO2 (dependent variable) with energy (independent variable) in their *natural* (non-logged) form** | |
|  | |
| Now our formula is CO2 emissions = 0.3686 + 2.20205\*energy use. This means that for every 1 unit increase in energy use, we see a 2.2 unit increase in CO2 emissions. To be more specific, for every 1 unit increase in energy use per capita we see a 2.20205 metric tonnes per capita increase in CO2 emissions. | |
| Let’s take a look at these values for Canada specifically |  |
| For Canada, we should see CO2 at 17.42608 and energy use at 8.365201. These are per capita values, so they can be interpreted as the average per Canadian. |  |
| **Q6. Write code to predict the CO2 emissions for all of Canada if there was a two unit increase in energy use per capita. Assume that Canada’s population is 40 million people.** Hint: you can do this by extracting the model coefficients (the intercept and the energy use) manually (writing them done from what you see on the screen) or extracting them programmatically. Once you have these coefficients, you can generate a prediction by simply writing out the regression formula in R. For example:  Say I have a coefficient of 0.1 for an independent variable A (Bx1) and a coefficient of 0.7 for an intercept term (BO1). Assume the data for variable A is a value of 1000 (x1). To get a prediction:  B0 <- 0.7  Bx1 <- 0.1  x1 <- 1000  Prediction <- B0 + Bx1\*x1 | |
| **Q7. Find some data (a dependent and independent variable) and use lm() to analyse their relationship. Briefly describe the data, put the regression results in a table, and offer a brief interpretation.**  **Save all the code and written answers (as appropriate) for the questions above into a document. Ensure that the document is nice, clean and readable. Save the document in a pdf file.**  **This assignment is due *before* next class. Please submit on avenue before next week’s class.** | |