SOC6707 Intermediate Data Analysis, Winter 2022

Assignment 1

Due date: 31 January, 11:59pm

Details

There are **100 points** in total.

You will need to submit both your answers to the questions and accompanying R code. You should submit:

- your R Markdown file; and
- the knitted PDF resulting from your R Markdown file.

Please submit both files via Quercus.

Remember to:

- Label the answers to each question
- Label any graphs clearly with suitable axis labels and titles
- Comment your code so that it is easy to understand

Question 1 (30 points)

This question relates to the GSS dataset. We will be looking at how age at the time of first birth varies by education and current age. Note there are a few different education variables in the GSS dataset but for this question, we will be focusing on the binary `has_bachelor_or_higher` variable.

a)

Report the following descriptive statistics:

i) What proportion of respondents have a non-missing observation for their age at the time of the birth of their first child?
ii) What proportion of respondents have a non-missing observation for their highest level of education (`has_bachelor_or_higher`)?
iii) For those respondents who have a non-missing education value:

- What is the number of respondents by education group (at least a Bachelor's degree, less than a Bachelor's degree) that have a non-missing observation for age at first birth?
- What is the proportion of respondents by education group (at least a Bachelor's degree, less than a Bachelor's degree) that have a non-missing observation for age at first birth?

Comment briefly on your calculations.

b)

For parts b) and c), we will be looking at the subset of respondents who have an education level reported, so you can filter out those respondents who have missing values of education.

Plot histograms of age at first birth by education level (at least a Bachelor's degree, less than a Bachelor's degree), with both histograms shown on the same chart but colored in different colors. Use `geom_histogram(position = "dodge")` so that the histograms are plotted next to each other. Interpret your chart.

c)

i) Calculate the correlation between age and age at first birth. Interpret your finding.
ii) Calculate the mean age of first birth by age group (`age_group`) and education level (`has_bachelor_or_higher`). Note that this question uses the categorical variable `age_group` (not the continuous variable `age`)
iii) Create a line chart of the results from part iii), plotting mean age of first birth (y axis) versus age group (x axis), with a separate line (and different color) for education level. Comment on your chart. Does the pattern over age agree with your findings from part i)? Why or why not?

Question 2 (30 points)

This question relates to the country indicators dataset. You will be looking at two country's fertility rate (TFR) and child mortality rate. Specifically, the two variables are

- `tfr` = total fertility rate, which is the average number births per woman in that particular country and year
- `child_mort` = under-five child mortality rate, which is the number of deaths to children aged 5 or less per 1,000 live births.

a)

Choose two countries. For each country, find

i) the mean TFR and child mortality rate
ii) the year which has the highest TFR
iii) the correlation between TFR and child mortality

b)

Focusing on child mortality, report the percent increase or decease in the child mortality rate observed from 2009 to 2017.

c)

Make one graph (only one!) that illustrates the trend in child mortality over time in the two countries of interest.

In all parts, comment briefly on what you observe.

Question 3 (40 points)

This question relates to the birthweight dataset, which lists the birth weight (in grams) of sample of 500 babies.

a)

Plot a histogram of the birthweights and comment briefly.

b)

Calculate the mean, median, standard deviation, and range of the birthweights.

c)

Assume that the sample of 500 babies are drawn from a Normal distribution with mean and standard deviation of what you calculated in part b).

i) What is the probability of observing a baby that has low birthweight (formally defined as less than 2500g)?
ii) What is the probability of observing a baby that has a birthweight of more than 3700g?
iii) What is the probability of observing a birthweight between 3400g and 3500g?

d)

i) How many standard deviations above the mean is a baby that weighs 4000g at birth?
ii) We are told a baby has a Z score of 0.13. What is the baby's weight?

e)

Suppose we observe 15 more birthweights (listed below). Discuss briefly with the aid of calculations and/or graphs whether you think these 15 babies are likely to be sampled from the same population as above.

birthweight
3984
3913
3864
3620
4117
4215
3857
3935
3886
4116
4110
4080
3897
4021
4243

Research project

Please describe

- your research question(s) of interest, and why they are of interest
- any hypotheses you may have
- the dataset you plan to use
- your main dependent variable of interest
- your main independent variables of interest

You should demonstrate that you can obtain and read in your dataset into R.

Notes:

- Please submit research proposal as a separate document to Assignment 1 (there will be a separate part of Quercus)
- References to past work are strongly encouraged.
- You can write the proposal in either RMarkdown or Word, but remember the final project is required to be in RMarkdown.