Assignment 1

Directions

Submit two files on Moodle before the deadline:

[1] your **solution set** as a *.pdf [2] the **R script** as a *.R type file

For both, the first line should include your student ID and the assignment number

If you worked with one or two other students write "group:" and then their student ID numbers

[1] your **solution** for each question should include:

- the question number/part
- your answer in words
- the command in R that was used (if relevant)
- the output from Rstudio (if relevant)

[2] the **R script** should include:

- all the commands you used to generate the solutions
- individually written comments that explain your code to the grader

You should set up a new folder for the Assignment 2 with sub-folders labeled "Data", "Code", "Assignment", and "Solutions" and populate it as before.

Part A Interpreting R output

The following part is based on the file `newdata.Rdata` that was produced from the first homework assignment. The variable monthly_hrs_wrk measures the number of hours worked in a month.

1. Using the command `cor`

a. What can we learn from the R output below about the relationship between hours worked and education?

```
> cor(newdata$monthly_hrs_wrk, newdata$schooly, use = "complete.obs")
[1] -0.04570389
```

b. Why do we set the argument "`use = complete.obs`" in the function `cor`?

2. You use the command `lm` in R to estimate the linear regression of education on monthly hours worked. The estimated model is below:

$$monthly_hrs_wrk = 184.4 - 0.859 \cdot schooly + \hat{u}$$

a. Interpret the intercept from the estimated model above.

b. Based on the estimates above, what is the predicted number of monthly hours worked for an individual with 12 years of education? What about an individual with 16?

c. What can we say about the relation between years of study and hours worked? Is the relation positive or negative? Is it strong or weak?

d. Can we make a causal claim for these results? Explain.

e. What other variables are likely to influence the monthly hours worked? Indicate at least 3.

(continued on next page)

Part B Birthweight and cigarette smoking

On the course website you will find two files that you will need:

- *BWGHT2-Description.txt* : a text file containing variable names and descriptions. Read this to familiarize yourself with the dataset before you begin.
- *BWGHT2_v2026.xslx this* : an excel file containing the data. This data was taken from a project that looked at the effect of smoking during pregnancy on newborn weight.

1. **Data exploration** [Heiss 1.6.4]
   - **a.** Import the dataset into the R-studio environment and name the data.frame object `data.BW`
   - **b.** Create a new variable called `non.smoker` indicating (= TRUE) if a mother is a non-smoker
   - **c.** What is the proportion of smokers in the sample? (hint: don't include the observations that are NA observations in your calculations)
2. **Ordinal factor variables** [Heiss 1.6.1]
   - **a.** Create a frequency table of the variable `monpre`, the month prenatal care begins. (*hint*: use the command `prop.table`)
   - **b.** What is the median month that prenatal care begins in the sample?
3. **Statistical Dependence**
   - **a.** What is the average daily number of cigarettes smoked during pregnancy in the entire sample?
   - **b.** What is the average daily number of cigarettes smoked during pregnancy among women who are smokers?
   - **c.** Compare the results from 3(a) and 3(b). Which average, if any, is a good measure for typical smoking behavior during pregnancy? Explain.
4. **OLS regression** [Heiss 2.1, 2.2]
   - **a.** Run the regression between birth weight (the dependent variable) and the number of cigarettes smoked by the mother during pregnancy (the independent variable). Report the estimation results in the form of an equation
   - **b.** Is it possible to apply a causal interpretation to these results?
   - **c.** What is the predicted birthweight for mothers who don't smoke? What is the predicted birthweight for mothers who smoke a pack of cigarettes a day? (assume there are 30 cigarettes per pack)
   - **d.** What value does `cigs` have to be in order to predict a birth weight of 3.4kg?
   - **e.** A woman smokes 20 cigarettes a day and has a newborn weighing 3.2kg. What is the residual for this woman?
5. **Scatter plots** [Heiss 1.4.1 – 1.4.5, 2.1]

   For this question create a scatterplot of birthweight (on the y-axis) versus number of cigarettes smoked (on the x-axis). It should have the following features:
   - Descriptive X and Y axis labels (what is being plotted and what are its units?)
   - A title for the graph
   - Appropriate scale for the X and Y axes
   - A dashed line corresponding to the OLS regression from `lm.bwght.cigs` (*hint*: you can use any command you prefer but in base plot you can use `abline`)