Homework 0 (review)

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 names

[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
- comments that explain your code to the grader

All references to the Heiss textbook are in brackets [Heiss X.X.X]

Part 0 – Set-up AKA boring computer stuff AKA fun with organization

Before beginning, create a folder for your first assignment. Create the subfolders: "Data", "Code", "Assignment", "Solutions". Download and save this file (which you are currently reading) in the subfolder "Assignment". Create a new file in "Solutions" to write up your solutions.

Download the following two files from the course website and save them in your "Data" folder. You will need both of these to complete the assignment:

- exp2015_v2025.xlsx – this is the excel file containing the results of a survey on household costs and incomes in Israel from 2015
- Variable_description-English.xlsx – this is the excel file with an explanation of the different variables and the codes used to represent them

Immediately upon opening R-studio, open a new script file to type your commands. This is the file that you will attach to your homework submission ([2] above). You may wish to refer [Heiss 1.1.1] and [Heiss 1.1.2]. Save your R-script in the "Code" folder.[1] You should name this file something like "HW1-Rcode-v1" and save it as a file type *.R . Refer to the HW00-Rcode-[template] file for how to organize this file.

Set your working directory for R as the assignment folder [Heiss 1.1.4]. You can do this using the setwd command (change the file path depending on your own computer directories):

```
setwd("C:/Users/YourUserName/Documents/University/Year2/IntroMetrics/HW1")
```

Include this at the beginning of your R-script.

[1] Version control: each time you make a major revision to your script (e.g. after meeting with your group) save a new version e.g. update v1 to v2 in the file name.

Make sure you have installed the package that allows you to read in *.xlsx file-types. If it is not installed, you can run the following command. You will need to be connected to the internet for this to work.

```
> install.packages("rio")
```

In your R-script be sure to attach the package `rio`, using the command `library(rio)`.

Part B – basic data manipulation in R

1) Read-in the data using `import` and `<-` commands. [Heiss 1.3.6] and [Heiss 1.3.1]

- **a.** Read the file exp2015_v2025.xlsx to R-studio and save it as an object in local R memory called `mydata`.
- **b.** Create a new dataframe object called `newdata` with the following variables selected: `sex age national marital ussrborn schoolty schooly diploma weekwhrs weekwrk workstat incomems chld017`

2) Data types using `summary`, `as.factor`, and `as.numeric`. [Heiss 1.2.3]

- **a.** Find the type of data each variable represents using the `summary` command. Which ones are characters?
- **b.** Transform each character variables (including age) into factor variables (use `as.factor`)
- **c.** Create a new variable, `age_numeric`, which is a transformation of the factor variable age into a numeric type variable (hint: use `as.numeric`).

3) Clean data using the `subset` command. [Heiss 1.3.2]

- **a.** Subset the observations in `newdata` to those with age between 25 and 62. Refer to the variable description file to interpret different values for the variable `age_numeric`. Why do we need to limit our observations to this age range?
- **b.** Limit the observations in `newdata` to individuals with less than 25 years of education (individuals with more than 24 years are considered data-entry errors)

4) Explore the data using the commands `summary`. [Heiss 1.6.4] (You should consult the variable description file.)

- **a.** Create a table with the following summary statistics for each numeric (non-factor) variable: average, median, minimum, maximum.
- **b.** What can be learned about the variable `incomems` from this table?
- **c.** What are the percentage of the observations are women? (hint: use `sex`)
- **d.** What percentage of the observations are for people born in the USSR? (hint: use `ussrborn`)

5) Add new variables/drop variables in `newdata`.

- **a.** Create a variable called `emp` which takes the value 0 if the person is unemployed and 1 otherwise. (hint: use `workstat`)
- **b.** Create a new variable for hourly wage called `wage_per_hour`. (hint: use `weekwrk`, `weekwhrs`, `incomems`)
- **c.** Drop (aka delete) the variable `age_numeric` you created in 2)c. (hint: use `subset`)
- **d.** Remove observations that have outlier hourly wage (higher than 1,000 NIS), while keeping those that have hourly wage of NA. Use the command `is.na`
- **e.** Create a new variable `sex_by_nation` that classifies individuals into one of the following groups (and NA otherwise): Jewish men, Arab men, Jewish women, Arab women. You can start with the following commands and add three more lines of code for the remaining groups, then save as a factor.

```
newdata$sex_by_nation = NA
newdata$sex_by_nation[newdata$sex=="male" & newdata$national=="arab"] <-
"arab_male"
```

6) Save your cleaned data. [Heiss 1.3.3]
- **a.** The object newdata will be used in other problems sets for this class. Save the cleaned version in your "Data" folder as "HW1-cleanData.RData" using the command `export`

7) Explore across different groups
- **a.** Calculate the average hourly wage for each of the groups in 5d
- **b.** Describe in a couple sentences the wage gaps between these different groups. What are some things that could cause these gaps?
- **c.** For the same groupings in 5d calculate the employment rates (hint: use `emp` from 5a)

8) Through independent analysis answer the following question: **does being married provide an advantage or a disadvantage in the labor market?**
- **a.** Address differences in both wage and employment between married and unmarried people.
- **b.** Does your answer changes between men and women?

Part A – Statistics and Mathematics review

9) Given: X is a random variable with mean μ_X and variance σ_X^2, and Y is a random variable with mean μ_Y and variance σ_Y^2. Let the covariance between the two be $\text{Cov}(X,Y) = \sigma_{XY}$. Evaluate the following:
- **a.** $\mathbb{E}[4 + 6X]$
- **b.** $(\mu_X)^2 + \sigma_X^2$
- **c.** $\text{Var}(11 - 3X)$
- **d.** $\text{Cov}(X - 8, Y - 2)$
- **e.** $\text{Cov}(6X, 2Y)$
- **f.** $\text{Var}(8X - 6Y)$

10) Prove the following equalities hold
- **a.**
$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$
- **b.**
$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \bar{X})\,X_i$$
- **c.**
$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) - \sum_{i=1}^{n}(X_iY_i) = -n\bar{X}\bar{Y}$$

11) Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables satisfying $\text{Var}(X_i) = \sigma^2$ and $\mathbb{E}[X_i] = \mu$ for all $i = 1, \ldots, n$. Furthermore let Y be a random variable defined as
$$Y = \frac{\sum_{i=1}^{n} b_i X_i}{n},$$
where $b_1, b_2, \ldots b_n$ are real numbers (non-random).
- **a.** Find $\mathbb{E}[Y]$ in terms of μ, the mean of X_i and the b_i's
- **b.** Find $\text{Var}(Y)$ in terms of σ^2 the variance of X_i and the b_i's