


This exercise contains 17 questions for the total of 80 points.

Please answer all (!) questions in a  script. Text should be written as a comment using the '#' to *comment out* the text.<sup>1</sup> Make sure that the script runs without errors.

1. (0 points) Write down your name, your matriculation number, and the date.
2. (5 points) Set your working directory.
3. (5 points) Clear your global environment.
4. (5 points) Load the following package(s): `tidyverse`

The following table stems from a survey carried out at the Campus of the German Sport University of Cologne at Opening Day (first day of the new semester) between 8:00am and 8:20am. The questions were asked in an open face-to-face communication. The survey consists of 6 individuals with the following information:

id	1	2	3	4	5	6
sex	f	f	f	m	m	m
age	21	19	23	18	20	61
weight	48	55	50	71	77	85
calories	1700	1800	2300	2000	2800	2500
sport	60	120	180	60	240	30

Data description:

**id:** Variable with an anonymized identifier for each participant.

**sex:** Gender, i.e., the participants replied to be either male (m) or female (f).

**age** The age in years of the participants at the time of the survey.

**weight** Number of kg the participants pretended to weight.

**calories** Estimate of the participants on their average daily consumption of calories.

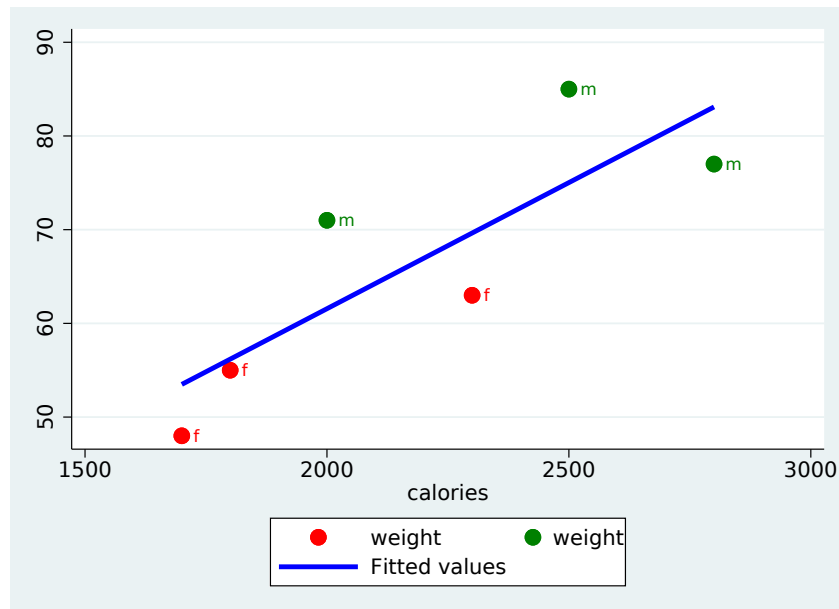
**sport** Estimate of the participants on their average daily time that they spend on doing sports (measured in minutes).

5. (5 points) Which type of data do we have here? (Panel data, repeated cross-sectional data, cross-sectional data, time Series data)
6. (10 points) Store each of the five variables in a vector and put all five variables into a dataframe with the title `df`. If you fail here, read in the data using this line of code:  

```
df <- read_csv("https://raw.githubusercontent.com/hubchev/courses/main/dta/df-calories.csv")
```
7. (5 points) Show for all numerical variables the summary statistics including the mean, median, minimum, and the maximum.

<sup>1</sup>The easiest way to create a multi-line comment in RStudio is to highlight the text and press Ctrl+Shift+C. For macOS, use Command+Shift+C.

8. (5 points) Show for all numerical variables the summary statistics including the mean and the standard deviation, **separated by male and female**. Use therefore the pipe operator.
9. (5 points) Suppose you want to analyze the general impact of average calories consumption per day on the weight. Discuss if the sample design is appropriate to draw conclusions on the population. What may cause some bias in the data? Discuss possibilities to improve the sampling and the survey, respectively.
10. (5 points) The following plot visualizes the two variables weight and calories. Discuss what can be improved in the graphical visualization.



11. (5 points) Make a scatterplot matrix containing all numerical variables.
12. (5 points) Calculate the Pearson Correlation Coefficient of the two variables
  - a) calories and sport, and
  - b) weight and calories.
13. (5 points) Make a scatterplot with `weight` in the y-axis and `calories` on the x-axis. Additionally, the plot should contain a linear fit and the points should be labeled with the `sex` just like in the figure shown above.
14. Estimate the following regression specification using the OLS method:

$$weight_i = \beta_0 + \beta_1 calories_i + \epsilon_i.$$

Show a summary of the estimates that look like this:

```
Call:
lm(formula = weight ~ calories, data = df)

Residuals:
1      2      3      4      5      6 
-5.490 -1.182 -6.640  9.435 -6.099  9.976
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.730275 20.197867 0.383 0.7214
calories    0.026917 0.009107 2.956 0.0417 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.68 on 4 degrees of freedom
Multiple R-squared: 0.6859, Adjusted R-squared: 0.6074
F-statistic: 8.735 on 1 and 4 DF, p-value: 0.04174
```

15. (5 points) Interpret the results. In particular, interpret how many kg the estimated weight increases—on average and *ceteris paribus*—if calories increase by 100 calories. Additionally, discuss the statistical properties of the estimated coefficient  $\hat{\beta}_1$  and the meaning of the *Adjusted R-squared*.
16. (5 points) OLS estimates can suffer from omitted variable bias. State the two conditions that need to be fulfilled for omitted bias to occur.
17. (5 points) Discuss potential confounding variables that may cause omitted variable bias. Given the dataset above how can some of the confounding variables be *controlled for*?