Exercise sheet 2 – October 22, 2021

Please submit your solution electronically until October 29, 2021, 23:59
Send your script-file to jerke@soziologie.uzh.ch

Notes:

- **Please sufficiently comment your script and structure it according to the different tasks.**
- **Whenever a task asks for an explicit answer, please write down your answer directly in the script within a comment.**
- **Make sure that you fully document your solution in your script. If you give an answer, but there is no code to clearly reconstruct how the answer was determined, the answer cannot be counted.**
- **The tasks vary in difficulty. For some of them you may have to combine commands in a new way or have to look in the documentation of the respective libraries.**
- **Most exercise sheets will contain bonus questions, providing the possibility to obtain extra points.**
- **The solution of this sheet will be published on OLAT after the submission deadline expires.**

1. Download the data set "Voting_DE_2017" and load it with Python. [Background: It is the same data set that we used in the last session on logistic regression. The data set contains results from a post-election survey after the German federal election in 2017. Besides socio-demographic information, it contains the attitudes of the respondents towards politically relevant topics and their voting behavior.]

2. Let's do some data cleaning:
   a. Copy the data cleaning part from the py-script of the last course session (on logistic regression) into your script. Make sure to add the variable `intdauer_ges` (interview length) to the list of variables to be selected for the subset.
   b. We want to remove cases that might not have taken the survey seriously. Therefore, drop observations that have more than ten missing values across the variables (hint: check the Pandas documentation of the `dropna`-command).
   c. [Bonus] Further, drop all observations that were suspiciously fast with answering the survey. Therefore, remove the 10% fastest respondents with respect to interview length.

3. Our variable of interest is the Left-Right self-assessment (`q32`) of the respondents. Before running an explorative regression analysis, we want to inspect the variable and its properties.
   a. Plot a histogram of the distribution of that variable.
   b. Get an overview of the descriptive statistics.

4. We are interested in the relationship between the Left-Right self-assessment and several socio-demographic variables:
   a. What is the correlation between age and the Left-Right self-assessment? Further, illustrate the relationship in a scatter plot. How would you assess the relationship?
   b. What is the mean of the Left-Right self-assessment for male and female participants?

   c. What is the mean of the Left-Right self-assessment for Western and Eastern Germany participants?
   d. Do you think there is a relationship between education and the Left-Right self-assessment?

5. Run a first regression analysis in which you regress the socio-demographic variables from the task above on the Left-Right self-assessment.
   a. Run the regression and print the regression summary.
   b. Provide an interpretation for all estimated coefficients in the regression model.
   c. How would you assess the R square from that regression model? What is the proportion of explained variance of the Left-Right self-assessment?
   d. Is there an interaction between the age and gender? Integrate the respective interaction term into the regression model and interpret the resulting coefficient and its significance.

6. It seems that the socio-demographic characteristics cannot really explain much of the variation in the Left-Right self-assessment. Therefore, let's have a look on several political attitudes.
   a. What is the relationship between the Left-Right self-assessment and the respondents fears with respect to the refugee crisis, global warming, international terrorism, globalization, political developments in Turkey and the use of nuclear power (`q73a-f`)? Therefore, calculate the pairwise correlations of these variables with the Left-Right self-assessment.
   b. Include these variables in the regression model from before and run another regression. Interpret the effects of the new variables.
   c. How does the R square change and what do you conclude for the explanatory power of that model?
   d. Save the predicted values and the residuals for each of the observations into two new variables.
   e. [Bonus] Regression diagnostics: Check whether the residuals approximately follow a normal distribution with mean 0.

7. [Bonus] We will further analyze the satisfaction with the current situation in the country and its effect on the Left-Right self-assessment.
   a. Corruption (`q10`): Create a new dummy (0-1) variable for the belief about the level of corruption. The value 1 should imply the belief of *(quite/very) widespread* corruption.
   b. Government performance (`q11`): Create a new dummy (0-1) variable for the opinion on the governmental performance. The value 1 should imply the opinion *(very) bad* governmental performance.
   c. Add the dummy variables to the regression model from before. Does the explanatory power of the model improve?
   d. Interpret the new regression coefficients.

8. Export the final data frame to a csv file.