EXAMINATION QUESTION PAPER - Course paper

GRA 60365
Data Analytics with Programming

Department of Economics

Start date:	23.04.2021	Time 09.00
Finish date:	07.05.2021	Time 12.00

Weight:	40% of GRA 6036
Total no. of pages:	8 incl. front page
No. of attachments files to question paper:	1
To be answered:	In groups of 1 - 3 students.

Answer paper size:	Length defined in each assignment. excl. attachments
Max no. of answer paper attachment files:	0
Allowed answer paper file types:	pdf

All information about what should be submitted is part of the question paper.

BI Norwegian Business School

Home group exam: GRA 6036 - Data analytics with programming
Made available:
Due data:
Permitted aids: Lecture notes, books, Google, etc.
Impermissible aids: People outside the group.
Format for your answer: A .pdf-file with text and math.

None of the questions on this exam are meant to be 'trick questions', and we have tried to make them as clear as possible. If you are unsure about the interpretation of a question, send us an email. Of course, understanding what questions mean requires some technical understanding, and we will not answer questions we judge as originating mainly from this. For the questions we do answer, we will post both the question and the answer on Itslearning so everyone has the same information at all times.

While the home exam is a group project, it is strictly forbidden to collaborate between groups, or in any other way get help with the problems from people outside of your group. Any suspicion of cheating will be reported and may result in oral exams. The consequences of being caught cheating are very harsh.

Draconian practicalities

You will be severely penalized for breaking the following rules:

- On R. Throughout the home exam, you are strongly encouraged to use the R-commands we have covered in the course. If you use other commands than those covered in the course (and not mentioned in the exam), you have to comment on what these commands do, and why you are using them.
- Note that if you are asked to program something in R, and you instead find some package or script that does this for you, this is not considered a solution and will yield zero points.
- On the page limit. You are to turn in a report with two parts: A main part, which has strict page limits specified in detail in each assignment, and an appendix with R-code. The appendix shall only include R-code and has no page limit. The code found in your appendix is to be complete, in the sense that one can run it and reproduce your entire analysis. You are free to include comments in your R-code.
- Formatting. You must use the Digiex formatting guidelines. You can simply copy-paste in the R-code directly into an appendix at the end of your document. Also write the answers to the mathematical problems in a word processor, such as Microsoft Word.

Assignments

- All assignments are weighted equally. All tasks within an assignment are weighted equally.
- In most assignments, you will be asked to only include very specific objects. These objects are indicated in the page limit description given at the start of each task.
- In most tasks you are not to include R-code in the main part of your report. The few exceptions will explicitly state that this is required.

Assignment 1

In this exam project, we will use (an extract of) a dataset originally collected and used by Benny Geys to study the relation between the physical appearance of politicians and their success/failure in the market for extra-parliamentary activities (https://onlinelibrary.wiley.com/doi/full/10.1111/geer.12041). The exam assignment does *not* require you to have read the article, though reading the article may benefit your understanding of some of the variables in the dataset. The dataset covers information on 617 German politicians, and will be provided on its learning in .csv format at the start of the exam (in a file named "Beauty_Germany.csv"). You can read it into R by running the following command:

Looks <- read.csv("Beauty_Germany.csv")

Below is a list of variable names and descriptions:

mp_id:	Unique identifier for each politician in the dataset
party:	Political party of each politician in the dataset
sex:	Dummy variable equal to 1 if female, 0 if male
age:	Age in years
highedu:	Dummy variable equal to 1 if university/college degree, 0 otherwise
phd:	Dummy variable equal to 1 if PhD degree, 0 otherwise
Terms:	Number of terms in office
business:	Dummy variable equal to 1 if economics/business degree, 0 otherwise
law:	Dummy variable equal to 1 if law degree, 0 otherwise
leadingpos:	Dummy variable equal to 1 if the politicians hold an important office (such as minister)
committees:	Number of committee memberships
job:	Number of extra-parliamentary positions maintained by the politician
money:	Amount of income from extra-parliamentary positions per month (in euro)
beauty:	Assessment of politician's physical appearance or attractiveness
competence:	Assessment of how competent people find this person
likability:	Assessment of how nice, pleasant and agreeable people find this person
trustworthiness:	Assessment of how ethical, honest and responsible people find this person
intelligence:	Assessment of how intelligent people find this person

The last five variables are based on evaluations of the politicians' official pictures (taken from their Bundestag website) by 28 independent individuals. As not all individuals assessed all politicians' pictures, the data rely on 7.5 evaluations per picture on average. The question asked was: 'Based on the picture provided, what do you think of this person – compared to people living in your country – in terms of […]'. This question was repeated for all five traits (beauty, competence, likability, trustworthiness and intelligence), and respondents replied on a five−point scale, where 1 means 'very positive' and 5 means 'very negative'. The data provided are the average score across all evaluations of a given picture.

When you start solving the problems given below, we assume that you have successfully read the dataset into R. (Make sure you set the working directory to the folder location of your data file when using the command above!)

(A) (Page limit: The requested barplot and variable values. No text.)

We start with some descriptive statistics to get a better feel for the data we have available. The first thing we will investigate is the number of politicians with a given number of extra-parliamentary positions.

(I) Calculate the **number** of politicians with 0, 1, 2, 3, 4, 5, 6 and '7 or more' extra-parliamentary positions, and save the result in a new variable called 'jobs'. Then display the values of this new variable 'jobs' using the barplot command. Provide sensible labels for both x axis and y axis, and also make sure the values on the x axis are informative.

(II) Calculate the **proportion** of politicians with more than five extra-parliamentary positions for each of the main three German parties (i.e. CDUCSU, SPD and FDP). Save the result in a new variable called 'fivejobs', and display this variable.

Hint 1: You can, but need not, use a loop for these calculations.
Hint 2: You can define the values on the x axis of a barplot by adjusting the following barplot option to your purpose: names.arg=c("a","b").

(B) (Page limit: The requested information. No text.)

Let us now turn to the perceptions of politicians' perceived traits: beauty, competence, likability, trustworthiness and intelligence. (Recall that 1 means 'very positive' and 5 means 'very negative'.)

(I) Find the observation that has missing information about politicians' beauty (i.e. 'NA', for 'not available'). Display all available information about this politician.

(II) Remove this politician from the dataset. Now calculate the average and standard deviation for the five perceived traits for all politicians in the sample. Present the results in a table with the average values in the first row and standard deviations in the second row.

Hint 1: The function is.na(some_data) indicates which elements are missing.
Hint 2: To create the table, you can first create an empty data frame called 'MeanTraits' with the correct proportions, and then fill in the calculated numbers in the appropriate places within the table. The code below creates a table with one row and two columns (where all values are 'NA', for 'not available'), and adds column names. Adjust this code to your purpose.

```
Traits.all <- data.frame(matrix(NA,1,2))
names(Traits.all) <- c("column 1", "column 2")
```

(C) (Page limit: The requested regression output. Explain the regression output in maximum five lines.)

Use the OLS regression framework to test whether there is a significant difference in the average perceived beauty of male and female politicians.

(D) (Page limit: The requested plots. No text.)

Beauty has been argued to give you many advantages. It seems natural to suspect that politicians who are perceived to be more beautiful, are also perceived to be more likable. We will not go into that here, but we will address how looks relate to perceptions of intelligence and trustworthiness.

First, generate three histograms displaying the distribution of perceived intelligence for three sets of politicians: those with an average beauty rating i) strictly below 2.5, ii) between 2.5 and 3.5, and iii) strictly above 3.5. Do the same for the distribution of perceived trustworthiness for those same sets of politicians. Display all six plots, and give each plot a short – but comprehensible – title.

Hint 1: Name the three subgroups of politicians 'belavg' (below average), 'avg' (average) and 'abvavg' (above average). Make sure you assign these labels appropriately given the nature of the scale used in the dataset.

(E) (Page limit: The requested regression output. Explain the regression output in maximum eight lines.)

Next, let us analyse the relation between beauty and trustworthiness in some more detail. Run an OLS regression model where politicians' perceived trustworthiness is the dependent variable. Include politicians' gender and perceived beauty as explanatory variables. Also include the interaction between both variables. Display the regression output and briefly explain the key lessons we learn from this model. Focus on the interpretation of the coefficient estimates (and their statistical significance).

(F) (Page limit: The requested regression output. Explain your results in maximum eight lines.)

Now we turn to the determinants of politicians' perceived trustworthiness. Run an OLS regression model where politicians' perceived trustworthiness is the dependent variable. Include politicians' age and age squared as the main explanatory variables. Add politicians' sex and partisan affiliation as additional variables in the model. Display the regression output and briefly explain the key lessons we learn from this model. Include the interpretation of the coefficient estimates, their statistical significance as well as the model's explanatory power.

Hint 1: Note that party affiliation is a factor variable. When included in a regression model using the lm() command, R automatically converts them into indicator variables. Check carefully what R uses as the reference category when you discuss your findings.

(G) (Page limit: The requested plots and numbers. Add additional text only when explicitly requested.)

Finally, we study the determinants of politicians' income from extra-parliamentary positions.

(I) Plot politicians' income from extra-parliamentary positions against their number of terms in office. Add a smoothing curve to the plot.

(II) Remove the person with a monthly income from extra-parliamentary positions in excess of 12000€. Also ***keep only*** those politicians who actually earn positive income from extra-parliamentary positions (i.e. strictly larger than 0). Hence, the remainder of this analysis is going to be on a very selected subsample (N=168 politicians). Replicate the plot from (I) with this sample.

(III) Using your selected subsample, run an OLS regression model where politicians' income from extra-parliamentary positions is the dependent variable, and terms in office as well as its squared value are the main explanatory variables. Add the predicted values from this regression model to the plot created in (II), and add a legend. Discuss your key finding in maximum five lines.

(IV) Calculate after how many terms in office the income obtained from extra-parliamentary positions reaches its maximum. Show this number.

(V) Calculate how much income from extra-parliamentary positions a politician earns when (s)he is in office for the exact number of terms you calculated in (IV). Show this number.

(VI) Add the indicator variable for politicians with an important office to the regression model specified under (III) and look at the regression output. Discuss the main message you obtain from this regression model in maximum three lines.

(VII) Finally, do you think the observed relationship for politicians holding an important office is likely to be causal? Why (not)? Discuss this in maximum three lines.

Assignment 2

(A) (Page limit: The requested calculations, minimal text.)

Consider two groups with n_1, n_2 observations in each group, and let $N = n_1 + n_2$
We assume that
$y_i = \mu_1 + \epsilon_i$ for $i = 1,2,\ldots,n_1$ and that
$y_i = \mu_2 + \epsilon_i$ for $i = n_1 + 1,\ldots,N$,
where $\epsilon_i \sim N(0, \sigma^2)$

Show that the least squares estimator for μ_1, μ_2 are

$$\widehat{\mu_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i$$

$$\widehat{\mu_2} = \frac{1}{n_2} \sum_{i=n_1+1}^{N} y_i$$

(B) (Page limit: The requested estimates, the output and P-value for the test, conclusion of test (max 2 lines))

We now assume that $n_1 = 80, n_2 = 50, \mu_1 = 1, \mu_2 = 2, \sigma^2 = 3$. Simulate data for these two groups, and use the R command *lm* to estimate μ_1, μ_2. Now, do a test to check whether the means are inequal (use the R command *t.test* for this). Are there some arguments in *t.test* you need to set? If yes, why? Can you conclude that the means are not equal for the two groups?

Hint: An example of a function argument is main="Log-wage" *in plot(x, y, main="Log-wage")*

(C) (Page limit: The calculations for deriving the estimates, minimal text.)

Define the indicator variable
$x_i = 1$ for $i = 1, 2, \ldots, n_1$
$x_i = 0$ for $i = n_1 + 1, \ldots, N$.
Define the linear regression model $y_i = a + b x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

Derive the least squares estimators for a, b in this model. You are expected to derive the least squares estimators from scratch, and use the fact that x_i is an indicator variable to help you get to the solution.

Conclude that

$\hat{a} = \overline{y_{n_1+1:N}}$ and that
$\hat{b} = \overline{y_{1:n_1}} - \overline{y_{n_1+1:N}}$

Recall that $\overline{y_{m_1:m_2}} = \frac{1}{m_2 - m_1 + 1} \sum_{i=m_1}^{m_2} y_i$

Use $\hat{a}, \hat{b}$ to get the estimates $\widehat{\mu_1}, \widehat{\mu_2}$ from (A)

Hint: Since x_i is an indicator variable, $x_i^2 = x_i$.

(D) (Page limit: The regression output plus 6 lines of text.)

Estimate the parameters of the linear regression model defined in (C) using R. Do you get the same conclusion as in (B)? Give a short argument for why (or why not) the model in (C) is doing the same thing as the test in (B).