Introduction and instructions

Time allowed – 2 days (expected length of tasks: 3 hours + 1 hour upload-time)

READ THESE CAREFULLY

This take-home exam has four questions. You should answer all four questions. The total possible number of marks is 100. The number of marks for each question is stated in squared parentheses after each question and sub-question. You can solve the questions in any order you want to; however, please note that sub-questions might build on previous sub-questions.

This is an 'open book' examination. You are free to use any written materials you find useful, including your own notes and annotations. Nevertheless, this needs to be your own work, so you should not request help from anyone else. In addition, you will be expected to use Rstudio for Question 4, so having that programme open and ready might be helpful. For said question, you'll be provided with a skeleton R-script.

Throughout the exam, you will need to look at tables and figures. These were derived from the analysis of data from the European Social Survey for Great Britain. When reporting numbers, please use at least 2 decimal points where applicable. You will NOT need to submit the R-script you used for Question 4, as it is only a means to answer questions in this take home exam.

For the purposes of this exam, imagine that you have been hired as an intern for Crest Advisory. You are joining a project which aims to understand what underpins trust in the police and trust in the justice system. On your first day on the job, as a test, your new boss asks you to make sense of a few pages of results and some notes left to you by a previous intern.

Question 1 [as a whole: 40 marks]

It seems that the previous intern was assessing the association between people's ancestral identification, their emotional attachment to Europe, and their trust in the police.

trstplc	trust in the police, 0-10 scale with 0='No trust at all' and 10='Complete trust'
ancest	ancestral identification with being 0=British, 1=English, 2=Scottish, 3=Welsh, 4=Northern Irish, 5=Irish, 6=Other European, 7=Outside of Europe
atcherp	emotional attachment to Europe, 0-10 scale with 0='Not at all attached' and 10='Very emotionally attached'

1(a) First, consider the correlation plot (Figure 1) presented below. What do you think about the approach taken by the intern? How would you interpret the association between these three variables? **[6 Marks]**

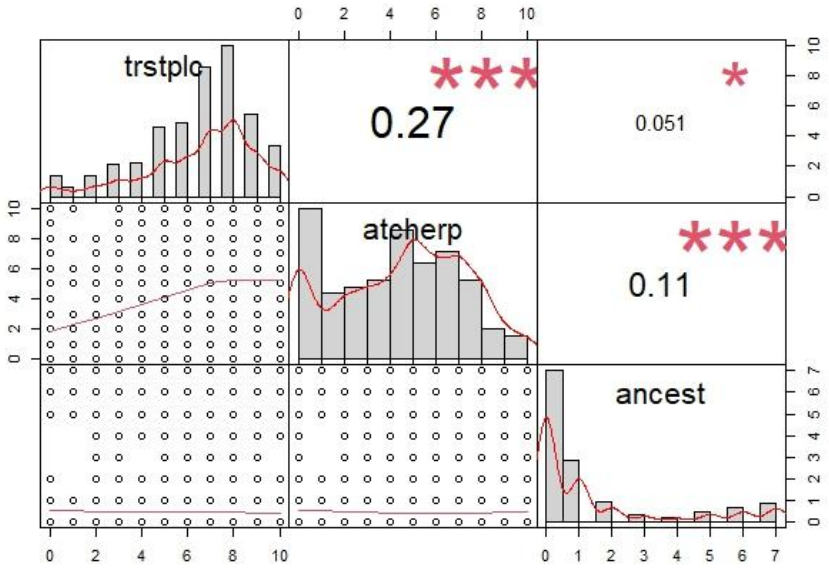

Figure 1

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.59983    0.07042  93.721  < 2e-16 ***
ancestf1    -0.36428    0.13091  -2.783  0.00544 **
ancestf2     0.02086    0.21021   0.099  0.92094
ancestf3     0.13602    0.33508   0.406  0.68482
ancestf4     0.07664    0.41503   0.185  0.85350
ancestf5    -0.19983    0.28425  -0.703  0.48214
ancestf6     0.77517    0.24423   3.174  0.00153 **
ancestf7     0.22412    0.21217   1.056  0.29094
```

Model 1

1(b) The previous intern decided to embark on an analysis as shown by Model 1. They had these three a priori hypotheses: (1) people identifying as English will be less trustworthy of the police compared to people identifying as British; (2) people identifying as other European nationals will be more trustworthy of the police compared to people identifying as British; (3) people outside of Europe will be less trustworthy of the police compared to people identifying as British. Interpret the corresponding variables in the model including the point estimate and the significance of the results. Do these results meet the expectations of the intern? Please explain why they do (not). **[12 Marks]**

1(c) The intern wanted to add the below table (Table 1) to their final model. Based on Model 1, please add the missing numbers to the table. **[4 Marks]**

	Average (mean) trust in the police according to ancestral identification
British	
English	
Other European	
Outside of Europe	

Table 1

1(d) Your colleague wrote down the following reminder next Model 1: 'DO NOT FORGET to change the reference/baseline category. NB: this could change the model fit and the substantive interpretation of the results!!!!'. What do you think about this note? Were they right to be concerned about these? **[4 Marks]**

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.62314    0.10713  52.490   <2e-16 ***
ancestf1    -0.22993    0.12701  -1.810   0.0704 .
ancestf2     0.07551    0.20351   0.371   0.7106
ancestf3     0.04330    0.32325   0.134   0.8935
ancestf4     0.04167    0.40028   0.104   0.9171
ancestf5    -0.14038    0.27425  -0.512   0.6088
ancestf6     0.27143    0.24077   1.127   0.2597
ancestf7     0.08563    0.20496   0.418   0.6761
atcherp      0.21633    0.01805  11.983   <2e-16 ***
```

Model 2

1(e) As the next step, the intern fitted Model 2, now adding the variable regarding people's emotional attachment to Europe as an explanatory variable. Focussing on the ancestral identification variable, how did the results change? Please use both statistical and substantive (i.e. common sense) explanations why this might be the case **[6 Marks]**

1(f) The intern was so inspired by the results that they left the following note at the bottom of the page: 'This is a very clear example how statistics can help with hypothesis generation.' Do you agree with this statement? Please explain why you do (not). **[4 Marks]**

1(g) Unrelated to any of the above, what do you think about the following quote from John Steinbeck regarding scientific discovery: “We knew that what we would see and record and construct would be warped, as all knowledge patterns are warped, first by the collective pressure and stream of our time and race, second by the thrust of our individual personalities. (…) Let us consider that factor and not be betrayed by this myth of permanent objective reality.” Do you agree with this? In your own words, describe why you do (not). **[4 Marks]**

Question 2 [as a whole: 35 marks]

As a continuation of the modelling described in Question 1, your colleague considered a series of models where he included several additional explanatory variables. These included the individual and their experience of victimisation and the gender and age of the respondent:

crmvct anyone in the household a victim of burglary in the last five years (1=yes; 2=no)

gndr gender of the individual (1=male, 2=female)

agea age of respondent (in years)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.822566   0.331555  14.545  < 2e-16 ***
ancestf1    -0.202565   0.129429  -1.565  0.11772
ancestf2     0.062967   0.203717   0.309  0.75728
ancestf3     0.040306   0.322988   0.125  0.90070
ancestf4     0.009877   0.400154   0.025  0.98031
ancestf5    -0.123234   0.274197  -0.449  0.65316
ancestf6     0.261148   0.246441   1.060  0.28941
ancestf7     0.105072   0.207824   0.506  0.61320
atcherp      0.212302   0.018161  11.690  < 2e-16 ***
crmvct       0.368562   0.136157   2.707  0.00685 **
gndr         0.159788   0.100477   1.590  0.11192
agea        -0.002104   0.002851  -0.738  0.46061
```

Model 3

2(a) First, look at model 3, and interpret the association between the newly added variables (i.e. victimisation, gender, and age) and the outcome variable (i.e. trust in the police). **[9 marks]**

2(b) The previous intern made note of the correlation coefficients between emotional attachment to Europe and the newly added variables (victimisation: r=0.04; age: r=-0.07; gender: r=0.06) and wrote down the following: "Weak correlations = unchanged results from Model 2 to Model 3. Makes perfect sense". Does it really make sense? Please explain why it does (not). **[4 marks]**

	Adjusted R^2	AIC	BIC
Model 1	1.1%	9924.67	9970.11
Model 2	7.2%	9612.96	9680.94
Model 3	7.6%	9705.24	9756.31

Table 2

2(c) Your predecessor made note of the changes in the model fit indices as shown in Table 2, with a handwritten note: "what's going on here?!". Based on the competing model fit indices, explain what might be going on. **[8 marks]**

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.960e-16  2.086e-02   0.000  1.00000
ancestf1.z  -3.495e-02  2.233e-02  -1.565  0.11772
ancestf2.z   6.628e-03  2.144e-02   0.309  0.75728
ancestf3.z   2.632e-03  2.109e-02   0.125  0.90070
ancestf4.z   5.190e-04  2.103e-02   0.025  0.98031
ancestf5.z  -9.523e-03  2.119e-02  -0.449  0.65316
ancestf6.z   2.305e-02  2.175e-02   1.060  0.28941
ancestf7.z   1.092e-02  2.159e-02   0.506  0.61320
atcherp.z    2.498e-01  2.137e-02  11.690  < 2e-16 ***
crmvct.z     5.740e-02  2.121e-02   2.707  0.00685 **
gndr.z       3.339e-02  2.099e-02   1.590  0.11192
agea.z      -1.620e-02  2.195e-02  -0.738  0.46061
```

Model 3.1

2(d) The previous intern derived the standardised estimates for the model (Model 3.1) but did not have time to write down anything regarding them. Explain why it makes sense to use standardised estimates and demonstrate their utility by discussing the statistically significant associations found in Model 3.1. **[6 marks]**

2(e) Model 1-3 are slightly different in the number of observations included in each of them (Model 1: 2167; Model 2: 2153; Model 3: 2133). Under what circumstances is it acceptable to use complete case analysis (as was done by the previous intern)? **[4 marks]**

Question 3 [as a whole: 9 marks]

Your predecessor has started sketching out some ideas for the introductory chapter of the report. They wrote down some key sentences but did not go into detail about either of them. Could you please fill in the gaps and elaborate what they meant by each statement?

3(a) "The European Social Survey is a probabilistic survey that is representative of the population of Great Britain." What do 'probabilistic' and 'representative' mean in this context? **[3 marks]**

3(b) "The code will be made open access thus, making the results reproducible." What is reproducibility? Why is it important? **[3 marks]**

3(c) "All our hypotheses were pre-registered to avoid p-hacking." What is p-hacking and why is it important to avoid it? **[3 marks]**

Question 4 [as a whole: 16 marks]

The previous intern was about the embark on an analysis before their departure. You found a sanitised dataset and a draft of an R code. They were planning to fit the same model to trust in the justice system as they did to trust in the police (Model 3). The only new variable is the trust in the justice system:

trstlgl trust in the legal/justice system, 0-10 scale with 0='No trust at all' and 10='Complete trust'

Showing how well prepared they were, the intern wanted to start by carrying out the necessary steps for regression diagnostics and further diagnostics

For 4(a)-4(d), select ALL the answers that you deem to be correct (and at least one of them). Choosing the wrong answer will result in 1 mark deduction for that question. You can get a maximum of 4 marks and a minimum of 0 mark for each question (i.e. getting negative marks is impossible).

4(a) Run the model in the R script. Which of the following statements are true regarding the associations between the explanatory and outcome variables? **[4 marks]**

(a) All else being equal, people identifying as English instead of British did not differ significantly in their trust of the legal system.
(b) Controlling for all else, males had significantly worse views about the legal system than females.
(c) Ceteris paribus, age did not have a significant association with trust in the justice system.
(d) It is clear from the results that, holding all else constant, past experience of victimisation had the strongest association with trust in the legal system.
(e) All of the above.

4(b) Staying with the previous model, which of the following statements are correct? **[4 marks]**

(a) The F-test is significant which indicates that we can reject the null-hypothesis of a well-fitting model.
(b) Around 1.13% variation in the outcome variable is explained by the set of explanatory variables included in the model.

(c) The model included all observations, i.e. there were no missing values on the pertinent variables in the data.
(d) Based on the output of the model, it would be impossible to calculate the confidence intervals.
(e) None of the above.

4(c) Now consider the regression diagnostics. Which of the following statements are true? **[4 marks]**

(a) The homoskedasticity assumption is likely to be violated, especially for very low and higher values of the outcome variable.
(b) The normality of residuals assumption is likely to be violated both in the middle of the distribution but also at the tails.
(c) The linearity assumption is not required for the victimisation and gender variables.
(d) The significant Tukey-tests for emotional attachment to Europe and age mean that the linearity assumption holds.
(e) The independence assumption was not scrutinised by any of the tests.

4(d) Finally, consider the further diagnostics. Which of the following are correct? **[4 marks]**

(a) There is no sign that multicollinearity would be a problem for the model.
(b) There are fewer outliers than expected, indicating that there should not be any problem with this aspect of the modelling.
(c) The added variable plots did not reveal any serious violations.
(d) None of the above.
(e) All of the above.