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 geo-data from London (UK) and the Public Attitudes Survey, a representative survey of residents living in London. 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 final project.

For the purposes of this exam, imagine that you have been hired as an intern for the Mayor's Office for Policing and Crime. On your first day on the job, as a test, your new boss asks you to make sense of a few pages of notes and some data left to you by a previous intern.

Question 1 [as a whole: 31 marks]

It seems that the previous intern was assessing the association between crime rates and some other area-level variables across London. They focused on a single year and embarked upon the analysis by deriving the correlation coefficients for the following five variables:

pbame	proportion of adults from ethnic minorities in the area
pnukb	proportion of adults born outside of the UK in the area
pengfl	proportion of adults with language other than English as their first language in the area
pnoquals	proportion of adults with no educational qualifications in the area
crimerate	number of crimes committed in the area divided by the population*100,000

1(a) First, consider the correlation plot (Figure 1) presented below. How would you interpret the association between crime rate and the other variables? **[6 Marks]**

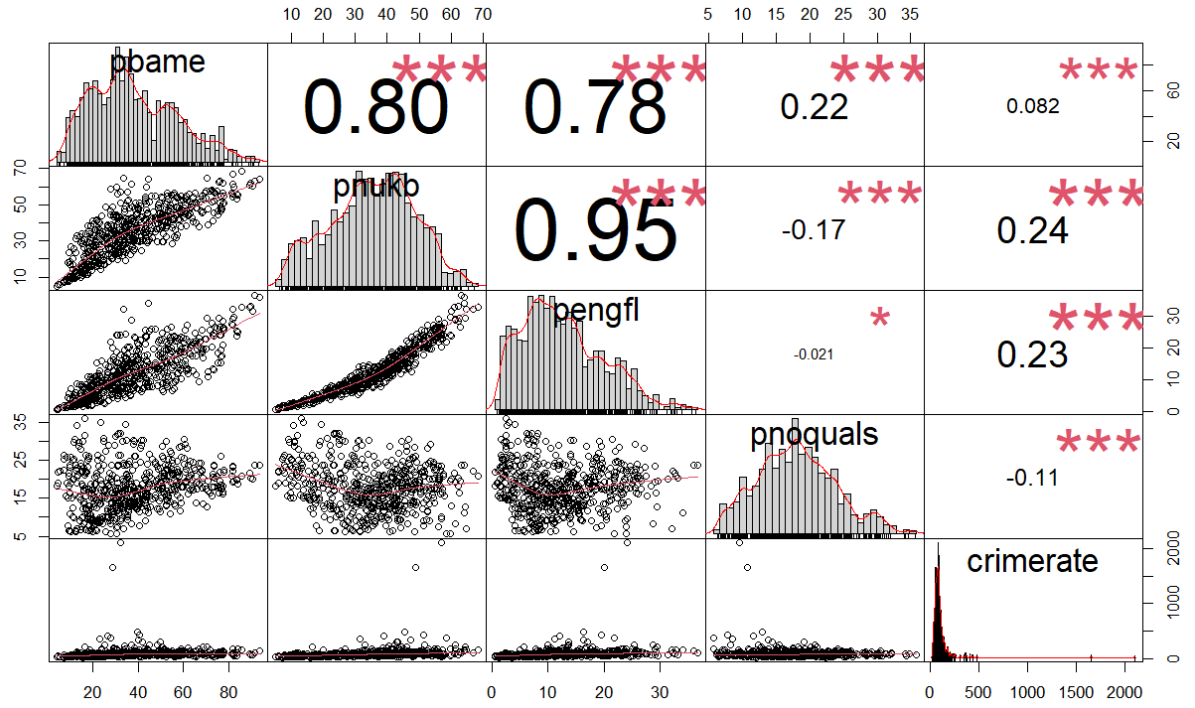

Figure 1

1(b) The previous intern made a note next to Figure 1: 0.6%-5.8%. What do these percentages stand for? **[4 Marks]**

1(c) You predecessor left another note below Figure 1. It said: 'population density-crime rate: r=0.007, p>0.05. Makes sense.' What does this mean? Were they correct asserting that this does make sense? **[4 Marks]**

1(d) As the next step, your colleague fitted a multilinear regression model (Model 1) with crime rate as the outcome variable and other constructs from the correlational analysis (Figure 1) as the explanatory variables. Notably, the proportion of adults not born in the UK was not included in this model. Was it the right decision to exclude this variable? Explain why it was (not). **[4 Marks]**

```
Call:
lm(formula = crimerate ~ pbame + pengfl + pnoquals)

Residuals:
    Min      1Q  Median      3Q     Max
-107.61  -37.49  -12.49   14.01 1909.41

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 92.98926    3.73624  24.888  < 2e-16 ***
pbame   ???  -1.32931    0.09614 -13.827  < 2e-16 ***
pengfl       6.24132    0.24090  25.908  < 2e-16 ***
pnoquals    -1.05921    0.18837  -5.623 1.92e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 112.1 on 11622 degrees of freedom
  (1203 observations deleted due to missingness)
Multiple R-squared:  0.07683,   Adjusted R-squared:  0.0766
F-statistic: 322.4 on 3 and 11622 DF,  p-value: < 2.2e-16
```

Model 1

1(e) After deriving Model 1, the previous intern wrote three question marks next to the 'pbame' line on the output. First, provide proper interpretation to the relationship between this explanatory and outcome variable. Second, please explain to the best of your ability why your predecessor might have been so surprised by the findings. Would you have shared their bewilderment? Explain why (not). **[7 Marks]**

	Effect size	t-value
Constant	91.91	41.19
pbame	-1.42	-24.88
pengfl	6.66	46.21
pnoquals	-1.06	-9.44

Table 1

1(f) As a final step, your predecessor broadened their analysis from a single year to four years. They wrote down the results for this new model in a table (Table 1) focussing on the point estimates and t-values. Compare the results of Table 1 and Model 1. How do these compare to each other? How would you explain the similarities and differences? How does this influence the interpretation of the findings? **[6 Marks]**

Question 2 [as a whole: 27 marks]

As a continuation of the study described in Question 1, your colleague considered whether the regression model they fitted meets the underlying assumptions of linear regression. It seems that they did not have time to interpret the results, but your new boss is confident that this will be a straightforward task for you.

Please read the description of the variables at the beginning of Question 1 and consider Model 1. All figures below are related to this particular model.

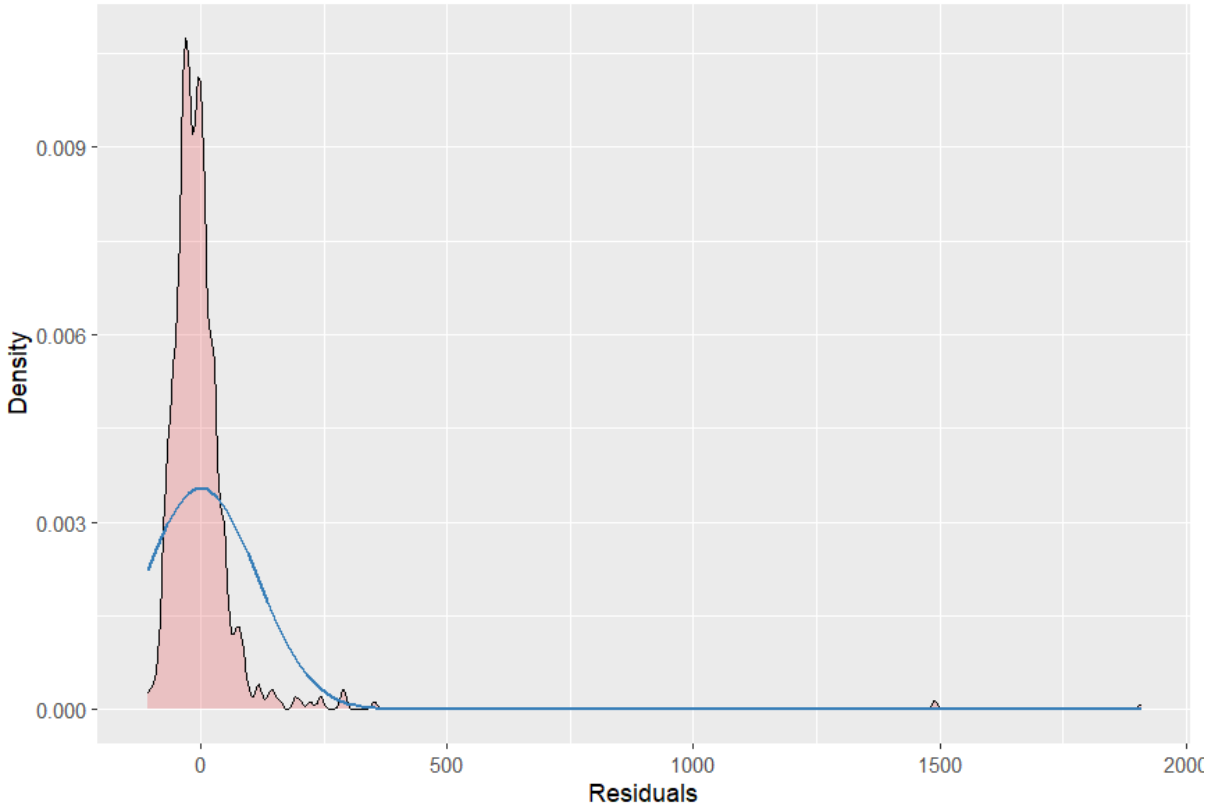

Figure 2.1

2(a) What does Figure 2.1 depict? How would you interpret this figure? What are the consequences of this interpretation for the point estimates and standard errors? **[7 marks]**

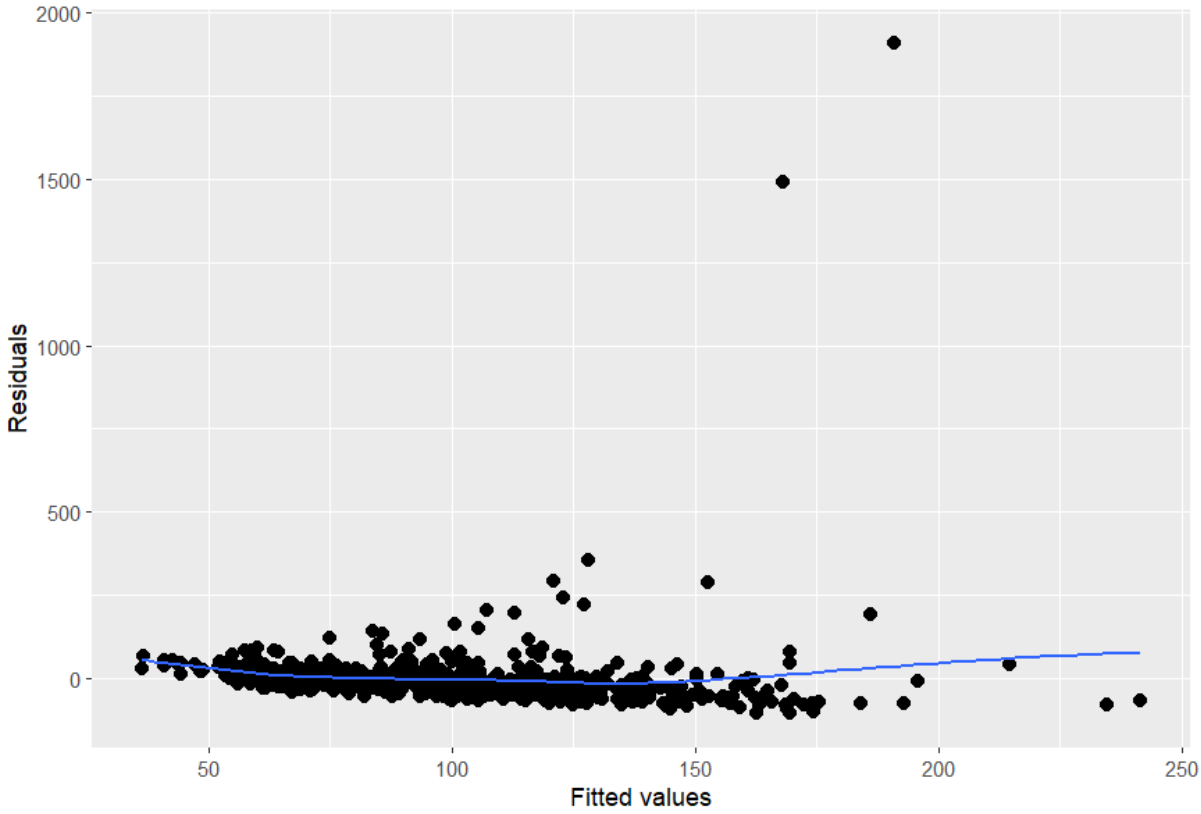

Figure 2.2

2(b) What does Figure 2.2 represent? How would you interpret this? What does this mean for fitting a linear regression? **[6 marks]**

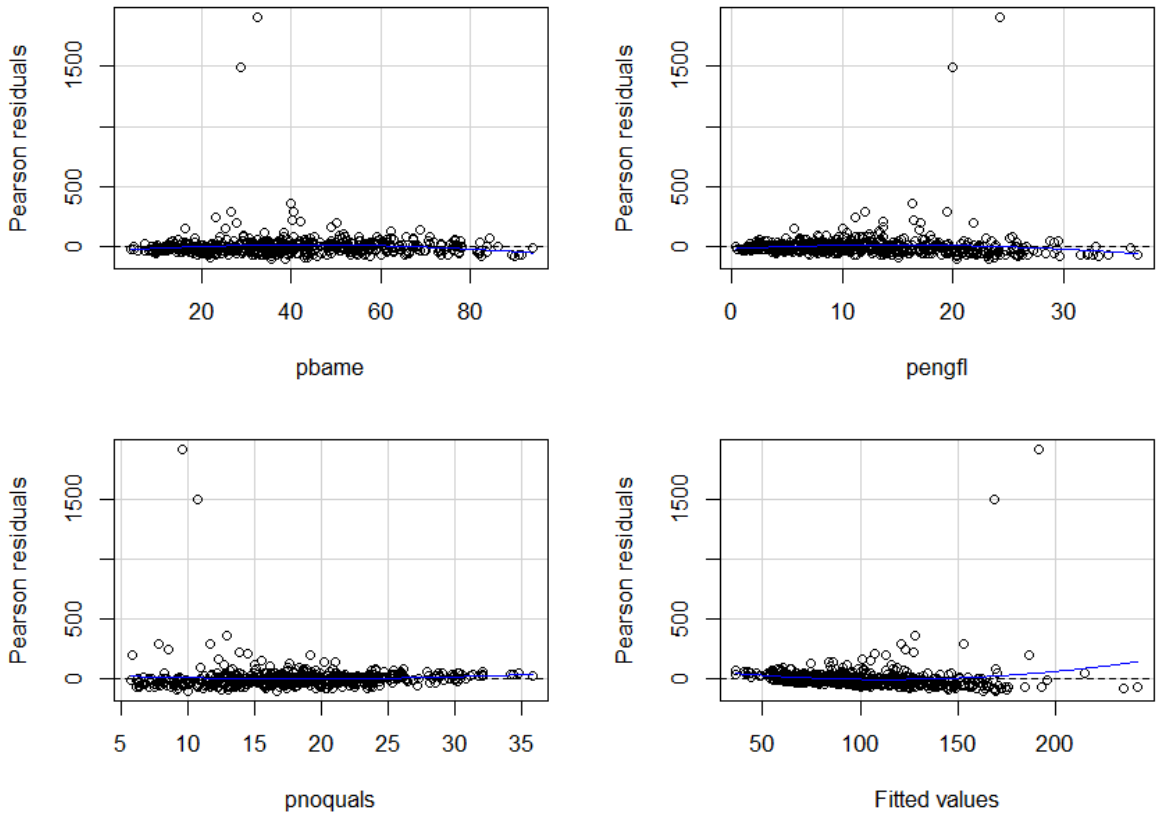

Figure 2.3

2(c) Now consider Figure 2.3. What does this group of four figures represent? How would you interpret the results? **[8 marks]**

2(d) Below these figures your predecessor made a note arguing: ‘quantitative methods and statistics are the only way to get an appropriate understanding of our world’. Challenge this notion from an epistemological perspective and explain why this is not always true. **[4 marks]**

2(e) Finally (and unrelated to any of the above): does God play dice with the universe? **[2 marks]**

Question 3 [as a whole: 26 marks]

The second project your predecessor worked on was an analysis of the Public Attitudes Survey (PAS). The PAS is a representative survey of residents living in London. They were considering the association between citizens' awareness of how they can contact the police, their sex, ethnicity, whether they were born in the UK, their physical disability, their student status, and victimisation experience. The full list of variables is available below.

contact	know how to contact the police if needed (0 = no; 1 = yes) – outcome variable
victim	victim of crime in the past two years (0=no; 1=yes)
female	sex of the individual (0=male; 1=female)
disability	1 = living with physical disability, 0 = no physical disability
born_UK	1 = born outside of UK, 0 = born in the UK
student	1 = current student, 0 = not a student
ethnic	ethnicity of the respondent
ethnic1	1 = White, 0 = other (dummy variable, recoded from ethnic)
ethnic2	1 = Black, 0 = other (dummy variable, recoded from ethnic)
ethnic3	1 = Asian, 0 = other (dummy variable, recoded from ethnic)
ethnic4	1 = Mixed, 0 = other (dummy variable, recoded from ethnic)
ethnic5	1 = Other, 0 = other (dummy variable, recoded from ethnic)

```
                                5 %        95 %
(Intercept) -0.5687483 -0.6007816 -0.5368096
victim       0.4356082  0.3408483  0.5302227
```

Model 3.1

```
                              5 %       95 %
(Intercept) 0.5662338 0.5483829 0.5846104
victim      1.5459030 1.4061399 1.6993107
```

Model 3.2

3(a) First, the previous intern fitted Model 3.1 and Model 3.2. Notice that they used 90% confidence intervals. What do you think about this decision? **[2 marks]**

3(b) How do the results of Model 3.1 and 3.2 and compare to each other? Using both outputs, interpret the association (i.e. effect size and confidence intervals) between victimisation and knowing how to contact the police. **[10 marks]**

3(c) What is the policy relevance of the findings? **[2 marks]**

3(d) Based on their notes, the previous intern was considering changing the model with victimisation as the outcome variable and knowing how to contact the police as the explanatory variable. How would this change affect the association between the two variables (Beta coefficient), the overall model fit (Hosmer-Lemeshow-test), and the comparative fit indices (AIC, BIC)? **[4 marks]**

		5 %	95 %
(Intercept)	0.5597937	0.5203567	0.6020581
victim	1.5361548	1.3933831	1.6932940
female	0.9390479	0.8824672	0.9992748
disability	1.1894724	1.0807214	1.3086659
born_UK	1.0526670	0.9813333	1.1292751
student	0.9482275	0.8340992	1.0766064
ethnic2	1.0404566	0.9222312	1.1729492
ethnic3	1.0297790	0.9398971	1.1279424
ethnic4	0.9938832	0.8167199	1.2059908
ethnic5	0.8921699	0.7671747	1.0356977

Model 3.3

3(e) As the next step, the previous intern fitted Model 3.3, adding the rest of the explanatory variables to the model, keeping knowing how to contact the police as the outcome variable. The mayor is considering embarking on a public awareness campaign with a focus on informing the public how they can reach out to the police should they need to. Based on Model 3.3, which populations should they focus on? **[4 marks]**

3(f) When reviewing the modelling carried out earlier, you notice that while in the case of Models 3.1 and 3.2 there were only three missing observations, in the case of Model 3.3 the number of missing cases increased to 657. The previous intern made a note that he found 'irrefutable evidence' that the data was missing completely at random so, this should not be an issue. What evidence could he think of? Was he right? **[4 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 a model to victimisation (as seen in Question 3) as the outcome variable. The full list of variables can be found below:

victim	victim of crime in the past two years (0=no; 1=yes)
crimeworry	worried about crime in the area (1-4, not at all worried-very worried)
vandalism	vandalism is a big problem in the area (1-4, not a problem at all-very big problem)
victsupp	the Met supports victims and witnesses of crime (1-7, does not support-completely supports)
fairtreat	the police treat everyone fairly in your area (1-5, strongly disagree-strongly agree)

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 two models. Which of the following explanatory variables had a significant association with the outcome variable in both models? **[4 marks]**

(a) Worry about crime
(b) Attitudes towards vandalism
(c) Perceived support of the Met of victims and witnesses
(d) Subjective fair treatment by the police
(e) All of them

4(b) Based on the two models you ran, which of the following statements are correct? **[4 marks]**

(a) Based on the binary logistic regression model, a one unit increase in subjective fair treatment by the police is associated with an increase of 88.9% in the odds of being a victim of crime, all else held constant.
(b) All partial associations in the logistic regression model are significant on the 1% level.
(c) Both models use complete case analysis.
(d) In the multiple linear regression, after controlling for all else, being very worried about crime instead of not at all worried increases victimisation by 0.153.
(e) The constant terms in both regression models are significant, implying that they are different from zero.

4(c) Now consider the model fit estimates. Which of the following statements are true? **[4 marks]**

(a) The F-test indicates that the model fits the data well, while the Hosmer-Lemeshow-test implies the opposite.
(b) The R-squared value of the linear regression model implies that 40.7% or 40.3% (adjusted R-squared) of the variation is explained by the explanatory variables.
(c) The R-squared value for the binary logistic regression model is larger, indicating a better fit of that model compared to the linear regression model.
(d) Based on the comparative fit indices, the binary logistic regression model provides a better fit than the linear regression model.
(e) None of the above.

4(d) Which of the following interpretations would you agree with based on the modelling? **[4 marks]**

(a) Controlling for all else, becoming a victim of crime makes people more likely to be worried about crime in general.
(b) Holding all else constant, there is a positive partial association between the seriousness of vandalism in one's area and victimisation.
(c) All else being equal, those who were victims of crime in the past two years, have a more negative opinion of victim and witness support compared to those who were not victims of crime.
(d) Ceteris paribus, by perceiving the police as treating people less fairly, you are more likely to become a victim of crime.
(e) In the two regression models, the results regarding the relationship between fair treatment of the police and victimisation are significantly different from each other.