The following can be considered as material for a partial mock of a representative Epiphany practical exam. You can think of the material below as representing **18** out of 30 marks of the practical exam (a further 6 marks would be obtained through the general multiple choice questions at the beginning of the test, and the final 6 marks come from the material on PCA, which would normally not have been taught by the time students tackled this sheet).

You will work through this sheet in the computer class on the 8th of March. You will **not** submit all answers. You will receive solutions after the computer class.

I have set up a very brief "practice" submission portal for the sheet, to remind you of how these work. You will need to create and upload the figure from Question (2)(a) for submission, but nothing else. I just wanted to give you a reminder of what the submission process looks like. The deadline for submission to the practical portal is 9am BST on Thursday 10th March.

Your **actual** practical exam will take place on Wednesday 16th of March, starting at 1pm, and will have the same style as in Michaelmas. In particular, the pratical exam will be available online (students wishing to sit the exam in a room on-campus should contact me as soon as possible), and the same rules regarding submission times (including additional time for those entitled to it) will be in force.

Question 7.1CC: For computer class on 08/03/22

We consider data from the Munich rental guide 1999. We are having available a sample of $n = 175$ flats, with certain characteristics of these flats provided by the following variables:

`rent`	net rent (DM) paid for a specific flat
`rentsqm`	net rent (DM per m^2) paid for a specific flat
`area`	living area (m^2)
`yearc`	year of construction
`location`	prime location=3, medium=2, poor=1
`kitchen`	specially equipped kitchen: yes=1, no=0.
`cheating`	central heating: yes=1, no=0.
`bath`	tiled bathroom: yes=1, no=0.

Preliminaries

The data can be found in the Epiphany Week 9 Practical Materials folder.

Before you start working on the data, please construct a new variable `age`, and transform `location` into a factor, as follows

```
munich$age<- 1999-munich$year
munich$location<-as.factor(munich$location)
```

(1) **Regression and Analysis of variance**

(a) Fit a linear model with `rentsqm` as response, and `age, area, location, kitchen, cheating, bath`, as well as an interaction term for `area` and `bath`, as predictor vari-

ables. Provide the model `summary` table and the sequential ANOVA table, with variables entering in the given order, and interpret the latter briefly. From any of these two outputs, extract the estimated error variance s^2 and save into an object `var`.

(b) Carry out backward elimination to simplify the model. Also try forward selection. If the results disagree, decide for the smaller of the selected models. In any case, fit your finally selected model, and save the fitted model object into `mfit`.

Hint: Use the object `var` created earlier for argument `scale` in `step(...)`.

(2) **Diagnostics and Influence**

(a) For model `mfit`, in a 2×2 split-table, provide plots of

i. residuals versus fitted values;
ii. residuals versus age;
iii. a QQ-plot of studentised residuals;
iv. leverage values versus studentised residuals.

Interpret each plot carefully. In particular, discuss what the last plot seems to tell about the existence of actually *influential* values.

(b) An ingenious Statistician proposes an alternative measure to Cook's distances in order to detect influential observations. The new criterion, E_i, is given by

$$E_i = \frac{n}{ps}|\hat{\epsilon}_i|[\exp(h_i) - 1].$$

Write a function `E.distance` which computes the values of E_i from a given fitted linear model (no other function arguments are needed!)

(c) For the model `mfit` fitted in part (1c), provide plots of

i. E_i versus i;
ii. Cook's distances D_i versus i;
iii. D_i versus E_i.

Does it appear that the two measures provide approximately equivalent information?

(d) The Statistician proceeds with developing an appropriate rule of thumb to classify cases as "influential" according to this new measure. Their line of thought is roughly as follows: Through a first order Taylor expansion, $\exp(h_i)$ can be approximated by $1+h_i$, so that the term in the squared bracket is $\approx h_i$. Then, consider a hypothetical case "at the edge" of being simultaneously outlying (i.e. $|\hat{\epsilon}_i|/s \approx 2$) and potentially influential (i.e., $h_i \approx 2p/n$). Give the value of E_i for such a case. Using now this value as rule–of–thumb criterion for the detection of influential cases, identify all influential observations for the regression problem at hand. Compare this result with the conclusions that would have been drawn using Cook's distances. Give your judgement on the suitability (for detecting influential cases) of the new measure E_i in general, and the suggested rule–of–thumb in particular.