MAE4000 Data Science 2018 – 2019 (resit)

Course coordinator: Professor dr. Johan Braeken
Contact:

Format: 4 hour written exam (inspera)

Support Material

You are allowed to bring and consult your own a priori self-made "cheat sheet". The cheat sheet can contain whatever course contents you find useful for yourself, but the sheet does need to fulfill the following requirements

- *1 page A4-format, double-sided, hand-written contents, and inside a plastic cover*

Anything outside the requirements (e.g., a typed page, 2,5 pages or even if it is a copy of the original), will not be permitted and taken away.

No calculators or other support material allowed.

Exam Instructions:

Questions mainly target understanding so answer short, to the point and correct. You are provided about double the space a concise and fully correct answer would fit in.

Before answering, REREAD THE QUESTION CAREFULLY

You have the default four hours of exam time. Do use your time wisely

- Start with questions you feel comfortable with & finish these;
- For remaining questions, prioritize questions with more points;

Best of luck!

Saskia, Fredrik, & Johan

Total score: 80 marks
Threshold values: A ≥ 67; B ≥59; C ≥ 52; D ≥ 45; E ≥ 41; F < 41

Problem 1.

1. What is printed as output to the R console if you would run the R-syntax below

```
x = c(1, seq(0, 10, 3) )
i = 2
while( x[i] <= 7 ){
   print( x[i]*2 )
   i = i+1
}
```

(15w) Marks 3

[1] 0, [1] 6, [1] 12 all on their own line

Sufficient to give the actual numbers in right order.
1 point per correct number on right spot
Wrong numbers but right idea gets max 2 points

2. A colleague has made exactly the same syntax as in the question above, except that he forgot the 5th line with the i = i+1; His Rstudio seems to now need ages to run and finish this code. Explain what is going on here. *(100w) Marks 2*

If line 5 is gone, the i variable that is used as index will always remain at the same value 2, as such the logical condition in the loop will remain correct forever; Thus you recreate an infinite loop and running the code will never get to be "finished".

Implied never-ending procedure / Infinite loop = 2 points
Notice Index i remains the same, but lack the above insight = 1 points

3. What is meant by the Garbage In Garbage Out (GIGO) principle in the context of data science? *(140w) Marks 3*

You can do all kind of fancy statistical analyses but the inferences and conclusion you draw from these will still be worth nothing if the data that goes into them is low quality, meaningless, full of errors, not representative for the study population, etc.

Importance of Data Quality & Study Design = 1 point each
Analysis cannot miraculously "save" you issues with the former = 1 point

Problem 2. The table below summarizes the success (S) rate of two treatments (T) in a medical trial for a number of patients suffering from one of three conditions (C).

	Treatment A		**Treatment B**	
	Patients	Success rate	Patients	Success rate
Condition 1	2000	90%	1000	80%
Condition 2	0	.	2000	70%
Condition 3	3000	80%	2000	60%

2.1 Provide a probability tree representation of the information provided on the left that summarizes the results of the medical trial. *(80w) Marks 7*

Tree

Applicants 10000

A 5000
- C1 2000
  - S 1800
  - \ S 200
- C3 3000
  - S 2400
  - \ S 600

B 5000
- C1 1000
  - S 800
  - \ S 200
- C2 2000
  - S 1400
  - \ S 600
- C3 2000
  - S 1200
  - \ S 800

Branching system 4 points: 1 per hierarchical branching variable

Frequencies correct 3 points:

Each sub branching needs to sum up to total branch before

2.2 Compute the probabilities requested below. Note that there is no need for calculators as all numbers are rounded and it suffices to answer with fractions of whole numbers (e.g. 800/1000), no decimal values required. *Marks 4*

Pr(\Success) = 2400 / 10000

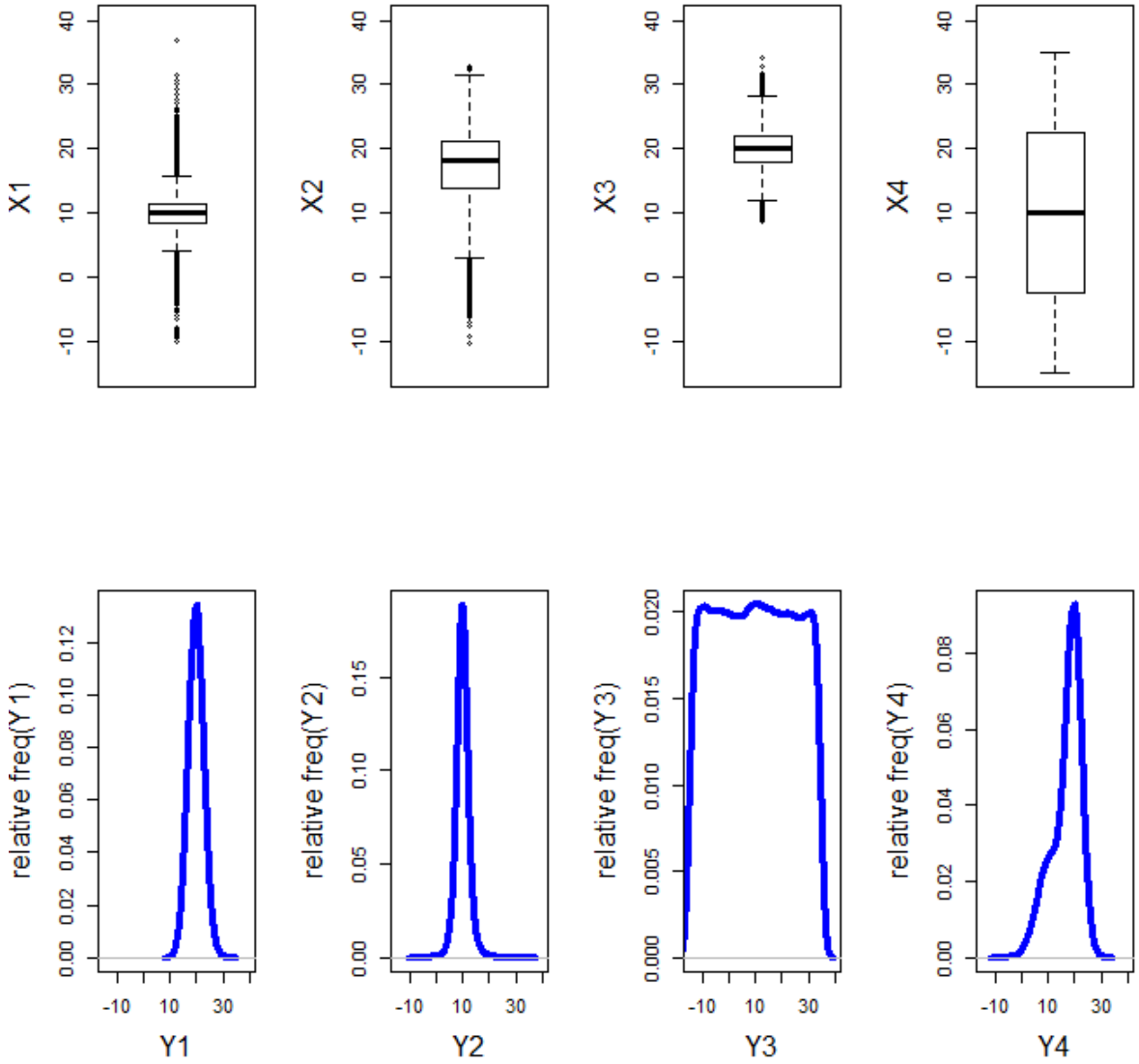
Pr(Success | Treatment A) = 4200 / 5000

Pr(Success | Condition 2 ∩ Treatment B) = 1400 / 2000

Pr(\ condition 2 U \Treatment B) = 8000 / 10000

Note. Correct fraction but not the obvious whole numbers, is correct as well!

Problem 3.



Variable	Skewness	Excess Kurtosis
Z1	0.04	14.20
Z2	-0.01	0.00
Z3	-1.22	0.25
Z4	0.00	-3.20

3.1 On the left you see boxplots of 4 X-variables and 4 figures depicting the distribution of 4 Y-variables (the height on the vertical axis of the blue line corresponds to the relative frequency of occurence in the sample of the value on the horizontal axis).

The X-variables and Y-variables are in fact the same set of variables! Every X-variable coresponds to a single Y-variable. Which X corresponds to which Y? Indicate briefly why they correspond to each other by referring to descriptive statistics that lead you to this conclusion. *(150w)* *Marks 8*

*2 points for each "couple", with 1 for correctness and 1 for use of proper indicators

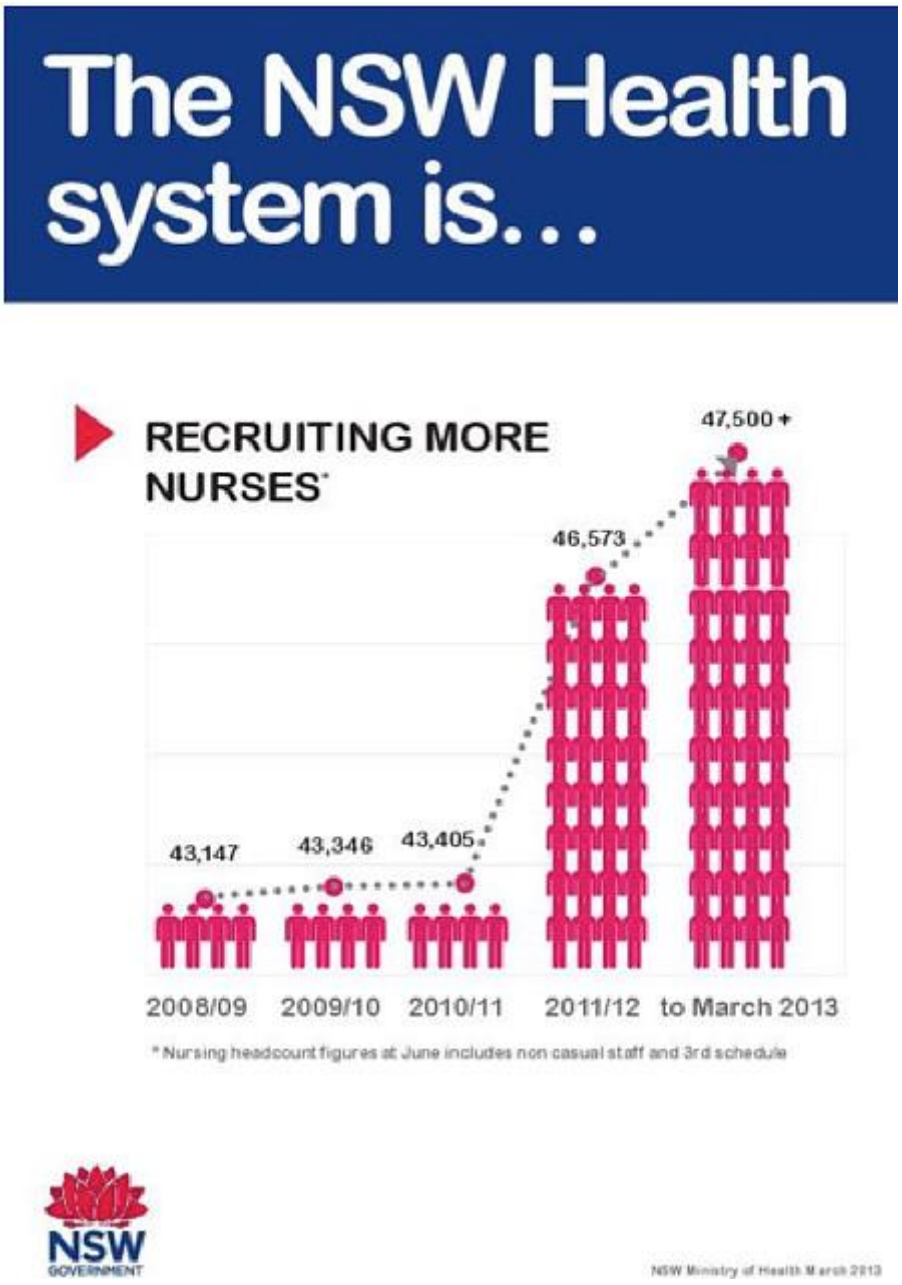
- X2 and Y4 correspond as they are both skewed in the same direction whereas other distributions are all symmetric.
- X1 and Y2 as they have both central tendency around 10 and long symmetric tails with lots of more extreme observations
- X3 and Y1 as they have both mean at 20 and most observations tightly but symmetrically clustered around the central tendency with low spread.
- X4 and Y3 correspond as they both have mean at 10 but high spread as seen in the interquartile range and quite uniformly flat distribution across the scale. These two also are the only ones with observations far past -10 at the lower end of the scale.

3.2 The table on the left reports on skewness and excess kurtosis of 4 Z-variables. Also here every Z-variable in fact corresponds to a single X-variable. *Marks 4*

Autoscored X1-Z1 , X2-Z3, X3-Z2, X4-Z4

Problem 4. The figure on the left appeared in a report of a public health institute and represents the headcount of nurses working in their facilities across a given period of time.

Evaluate the figure and discuss its merits & flaws according to the criteria of good graphical design for data representation. *(400)* *Marks 10*



Criteria: 2 points per main theme that is discussed with supporting evidence. Exact wording/terminology of themes is not essential as long as it is clearly implied.

1. Gestalt principles & illusions +2
   e.g., row of humans not easily separated visibly, horizontal grouping obvious though, + clear impression of difference between left and right

2. Keep it simple: Decoding & Operations +2
   e.g., Not complicated but still only five points in different representations

3. Less is more: Chartjunk & data-ink ratio +2
   e.g., No need for the little humans, numbers imply seemingly accuracy that is not there?, trendline + barchart of same data is redundant, strange gray stuff on top of last bar

4. Graphical data integrity & lie factor +2
   e.g., liefactor with tower of people, time scale is not proportional to spacing, title is guiding to specific conclusion, 47500+ unclear whether actual number or projected number or comparable to other numbers (note doesn't help)

5. Annotation & stand-alone Readability +2
   e.g., no vertical axis which is also not named so no clue what the numbers represent, x-axis scale same problem, the title and note compensate a bit for the lack of this information, but only implicit guidance

Problem 5

Researcher F.Ar performs a null hypothesis significance test of the difference in outcome between group A and group B. The resulting p-value is .74. Researcher F.Etched concludes therefore that there is no difference in outcome between the two groups.
Do you agree with the latter conclusion? Why / why not?
Explain what the .74 stands for in your answer. *(260w) Marks 6*

Long: There clearly is some difference between the two groups in the sample, because the p-value is not exactly 1. *The p-value of .74 would lead us to conclude that the probability of observing such a sample difference or more extreme for a similar sample is rather large, given that the null hypothesis of no difference would be true in the population*. Hence, we have no empirical contra-evidence that would lead to rejection of the null hypothesis of no difference; but that does not imply we have evidence supporting it, the latter would be a logical fallacy. So if the researcher's conclusion is for the sample at hand, it is clearly false; if it is for the population, you could argue we are undecided about the difference and hence the conclusion is also false.

- Short: p-value definition 3
- Short: not reject null hypothesis does not mean accept null hypothesis 2
- Short: observed sample difference is there, does not disappear by significance testing 1

How can residuals tell us anything about the tenability of the assumptions of a regression model for a given dataset? Aren't residuals supposed to be errors, and hence always wrong anyway? *(50w each) Marks 5*

Residuals represent non-systematic part of data not captured by systematic model part; 2.5
As an implication residuals are expected to be distributed as uniform white noise ("error") and not show any systematic pattern among themselves or with systematic elements of the model (e.g., predictors, predicted values, …). 2
When residuals systematically deviate from such random noise pattern, then the model assumptions are necessarily violated. 1

The following formula provides the asymptotic sampling distribution of the mean

$$\lim_{n\to\infty} \bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

(50w each) Marks 3

a) Explain which parameters in the formula determine the precision of the estimated mean and how.

Higher sample size (1point) and/or Lesser variation in X (1point), leads to more precision

b) If your current sample size is 9 how much does it need to change to half the standard error of the estimated mean, assuming that all other parameters remain constant?

SE.new = SE/2 = (sigma/sqrt(n)) / 2 = sigma / (sqrt(n)*2) = sigma / sqrt(n *2^2)
Thus quadruple sample size (n.new = 9*4 = 36)!
(1point for four times larger or just the number n = 36)

Problem 6. Researchers F. Ar & F. Etched are interested in studying whether neuroticism is to some extent linked to the person's extraversion and impulsivity. They have gathered data for a random sample of juvenile delinquents. These personality characteristics are operationalized within the Eysenck framework: Extraversion (variable epiE), Impulsivity (variable epiImp), Neuroticism (variable epiNeur), and a scale intended to measure how much the person tends to pretend/lie about her/himself (variable epilie), with on all variables higher values being indicative of more of that personality trait. Selected output of a data-analysis in R is given below:

```
> sapply(data,desc)
                    id    bdi   epiE epiImp epilie epiNeur
M               116.00   6.71  13.33   4.37   2.38   10.41
SD               66.83   5.80   4.14   1.88   1.50    4.90
skewness          0.00   1.29  -0.33   0.06   0.66    0.06
excess.kurtosis  -1.20   1.54  -0.04  -0.60   0.27   -0.48
min               1.00   0.00   1.00   0.00   0.00    0.00
max             231.00  27.00  22.00   9.00   7.00   23.00
n.eff           231.00 231.00 231.00 231.00 231.00  231.00
n               231.00 231.00 231.00 231.00 231.00  231.00
There were 24 warnings (use warnings() to see them)
> m = lm(epiNeur~1+epiE+epiImp+epilie,data)
> summary(m)

Call:
lm(formula = epiNeur ~ 1 + epiE + epiImp + epilie, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9521 -3.2956 -0.0002  3.1894 11.6666

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.7244     1.2540  13.337  < 2e-16 ***
epiE         -0.4165     0.1231  -3.385 0.000839 ***
epiImp        0.3540     0.2716   1.303 0.193752
epilie       -0.9700     0.2098  -4.624 6.32e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.61 on 227 degrees of freedom
Multiple R-squared:  0.1263,    Adjusted R-squared:  0.1148
F-statistic: 10.94 on 3 and 227 DF,  p-value: 9.706e-07

> regCoef(m,data,digits=2)
             b   SE     t    p [b.l,  b.u]   b_Z  Tol   Ryx Ryx.x Ry(x.x) R2yx R2y(x.x)
(Intercept) 16.72 1.25 13.34 0.00 14.25 19.20  0.00   NA    NA     NA      NA   NA        NA
epiE        -0.42 0.12 -3.38 0.00 -0.66 -0.17 -0.35 0.36 -0.18 -0.22   -0.21 0.03     0.04
epiImp       0.35 0.27  1.30 0.19 -0.18  0.89  0.14 0.35 -0.07  0.09    0.08 0.01     0.01
epilie      -0.97 0.21 -4.62 0.00 -1.38 -0.56 -0.30 0.94 -0.25 -0.29   -0.29 0.06     0.08
```

6.1 Based on the model results, give an interpretation (direction, size, generalizability, & meaning) of the intercept and the regression coefficient of epiE in model m. *(200w)*
Marks 10

Intercept: max 4

Base = Expected epiNeur score for someone with 0 score on all other epi variables.
2.5

- Generic interpretation without reference to actual variables max 1.5

Generalizability +1

- e.g., P-value/Null hypothesis 0 in pop rejected OR quite wide CI so pretty uncertain/imprecise estimate

Putting numeric value in context +1

e.g., lowest value on epiE in sample is apparently 1, so such a person does not exist in study sample; comparing value to descriptive statistics summaries, mean epiNeur is 10 and intercept slightly more than 1Sd above …

Slope: max 6

Base = Expected difference in epiNeur score for persons that are one score unit apart in epiE but have similar score on all other epi traits. 3.5

- Generic interpretation without reference to actual variables max 2.5
- Causal version max 2.5

Generalizability +1

- E.g., P-value indicates that such a finding or more extreme in a sample of the current size can be regarded an extreme unlikely result under the null hypothesis that this difference would be 0 in population. Based on this counterevidence we therefore can reject this null hypothesis. OR CI points …

Direction +1

- E.g., this implies a negative relation between epiE & epiNeur within context provided by our other predictors.

Putting numeric value in context +1

- E.g., reference to population, descriptive summary statistics, -.4 is only half a pointbut 1SD epiE is 4 so then 1.6 half a point = 1/3 SD on epiNeur so not that huge effect, also see semi-partial squared correlation of .04 → small "effect", in comparison epilie seems more crucial predictor .

Round up for final mark

6.2 Discuss why you agree / disagree with each of the following conclusions that the researchers have formulated. You can argue about contents or imprecision of the formulation.

a) Being extravert reduces the risk for an individual of acting neurotic.

b) Because epilie has a tolerance of .94, it would be good to remove it from the model. Removing it would also hardly influence the model results.

(250w) *Marks 10*

a) Max 5

Causal statement not warranted by study design. 2.5

Nothing in the data about individual risk for neurotic actions! Only response on personality survey! Yes, negative relation between trait scores but small and is at group level, not individual actions. 2.5

b) Max 5

Tolerance is % of variation in epilie not accounted for by other predictors, so this actually implies that epilie is almost not correlated with the other predictors. This does not make it bad for the model in contrast it contributes rather independent unique information to explain away variation in epiNeur. 2.5

Removing it from the model would have a big impact, as it is is a “significant” predictor in the model contributing about 8% to the R-square of 12.63%! The explanatory power of the model would decrease a lot, leading to an even bigger RMSE. 2.5

We even have a case of Enhancement here as both epiE and epilie explain more variation in the presence of the other predictors than on their own! +1.5

Basically +1 for comments that make sense, but don’t credit when contradicted by later comment.

Problem 7. Consider a continuous outcome variable Y and two categorical predictors variables X1(with categories A and B) and X2 (with categories male and female). X1 is recoded into a dummy variable dummy D1 that takes a value 0 when X1 = "B" and a value 1 when X1 = "A". X2 is recoded into a dummy variable dummy D2 that takes a value 0 when X2 = "female" and a value 1 when X2 = "male".

The following linear model is formulated: $Y = \beta_0 + \beta_1 D1 + \beta_2 D2 + \beta_3 D1 \times D2 + \varepsilon$. What is the interpretation of the regression coefficients β_1 and β_3 in terms of category group means? *(50 & 80w)* *Marks 3 each*

b1: *Expected difference in outcome variable Y between female persons of group A compared to females of group B 3

regression coefficient of first-order term of predictor D1 in model with interaction between D1 and D2 +1

expected mean difference in Y between persons one unit apart on D1 but with value of 0 on D2 +2

b3: *expected difference between male and female in expected mean difference in Y between groups B and A (and vice versa) 3

regression coefficient of crossproduct interaction term between D1 and D2 +1

if b3 positive then value of the expected difference in Y between B and A for males is higher compared to that for females (or vice versa for male-female difference and A vs B) 2