MAE4000 Data Science 2018 – 2019

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

Problem 1. You are handling a dataset in R that consists of exam scores for 189 students in grade 8 or 9 of 3 different schools in Oslo municipality and has the following structure

```
> str(data)
'data.frame':   189 obs. of  8 variables:
 $ School   : chr  "Norstrand" "Sagene" "Blindern" "Norstrand" ...
 $ Grade    : int  8 9 8 8 8 9 8 8 9 9 9 ...
 $ Name     : chr  "solveig-sundbo" "elin-rodum" "grunde-kreken" "per-amundsen" ...
 $ AgeInYears      : num  13.6  14.7 13.5 14.1 ...
 $ Norwegian: num  84 87 85 87 84 ...
 $ Math     : num  69 64 71 73 63 ...
 $ History  : num  78 71 80 79 80 ...
 $ Gender   : chr  "F" "F" "M" "M" ...
```

1.1 What does the following R-code do? *(50w)* *Marks 3*

```
> Anonymized  = data[sample(1:189,189),-3]
```

It takes the original dataset, shuffles the rows and deletes the third variable 'Name', and stores this modified data.frame in the object 'Anonymized'.

One point per action shuffle/delete/store

1.2 Is the action as coded in the R syntax sufficient to create a new anonymized dataset that can be publicly shared without the risk of individual students being easily identified? Briefly explain your answer. *(100w)* *Marks 3*

At least it does not contain the students' names anymore, but because we still know the school name, gender, and so much more about the students, we could in principle still easily figure out who is who based on specific value combinations as the number of students is quite low. Thus, it is only to some extent anonymous, but not sufficient.

Names are out 1 point
but still so much left: school = 1 point + combo other variables = 1 point

1.3 What principle or actions can be taken to (further) improve the anonymity of the persons in this sample? *(75w)* *Marks 2*

We need to loose detail to make it more difficult to identify specific individuals. For instance by replacing the school name by a code value; Forming age groups, transforming exam scores to grade point average, …

School = 1 point, grouping = 1 point,
other valid solution (e.g.,delete school, name, etc) can compensate for lack of one of these

Problem 2. The table below summarizes admission information for two departments split up by male and female applicants.

	Male		**Female**	
	Applicants	Admitted	Applicants	Admitted
DepartmentA	2000	60%	1000	80%
DepartmentB	1000	30%	1000	40%

2.1 Provide a probability tree representation of the above information. *(50w) Marks 7*

Tree

Applicants 5000

A	3000		
	M	2000	
		admitted	1200
		\ admitted	800
	F	1000	
		admitted	800
		\ admitted	200
B	2000		
	M	1000	
		admitted	300
		\ admitted	700
	F	1000	
		admitted	400
		\ admitted	600

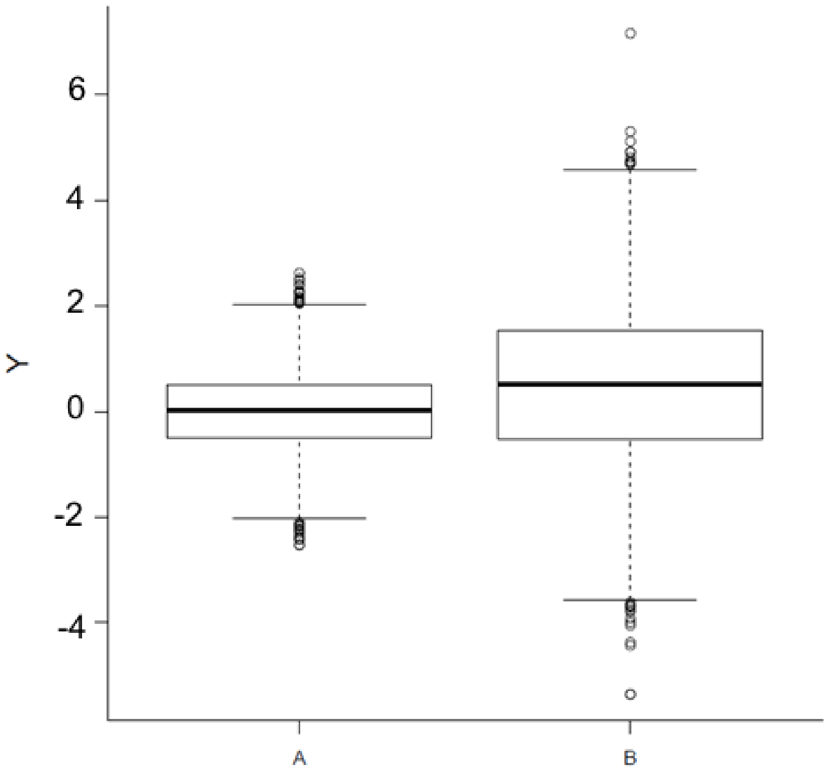
Branching system 4 points: 1 per hierarchical branching variable (Appl, AB, MF, Adm\Aadm)
Frequencies correct 3 points (each sub branching needs to sum up to total branch before)

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 ➔ ,5 per ?*

Pr(admitted) =	? / ?	2700 / 5000
Pr(admitted | Female) =	? / ?	1200 / 2000
Pr(\ admitted ∩ Male) =	? / ?	1500 / 5000
Pr(\ admitted U DepartmentB) =	? / ?	3000 / 5000

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

Problem 3. The figure represents the return on investments Y for two investment schemes A and B (Y values: positive = make profit, zero = break even, negative = make loss).



3.1 For scheme A, verbally summarize and describe the data represented in terms of known descriptive statistics as used in the boxplot. It is sufficient to provide approximate values for statistics used in your description. *(150w)* *Marks 6*

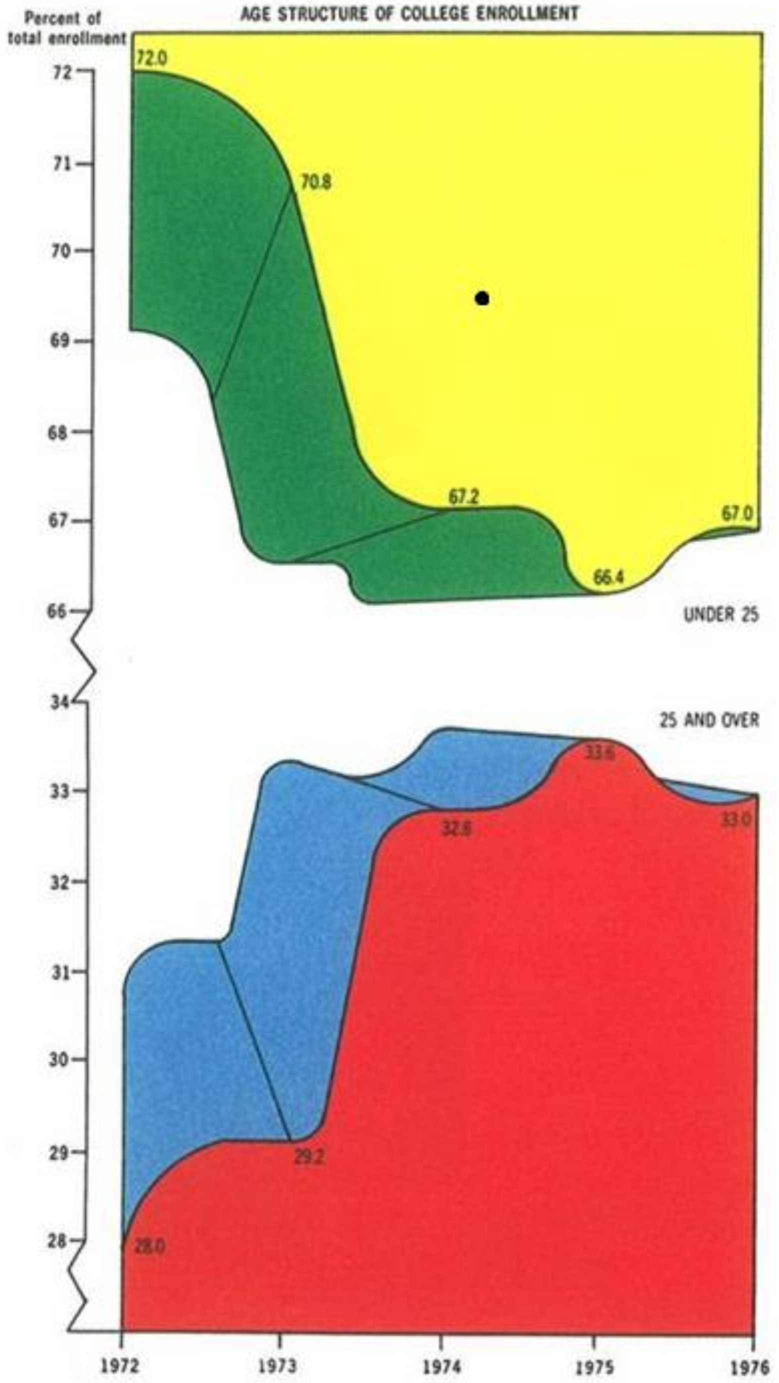
Median return of investment is slightly higher than 0, +1.5
with quartile 1 and 3 at respectively -.5 and .5, +1.5
such that about 50% of the observations fall between those two values, corresponding to an interquartile range of about 1, +1.5
such that observations with absolute values more than about 1.5 times this range outside the box (i.e., -/+2) are considered outliers. +1.5

Optional point for Pointing out the correct min/max value (i.e., corresponding to the most extreme lower and upper outlier respectively)

3.2 In which investment scheme are you better off if you want to play it completely safe and not risk losing too much money. Argue for your choice. *(150w)* *Marks 5*

The safer bet is A which guarantees to keep your losses within -2.5 boundary. +3
Whereas Scheme B gives on average more profits than scheme A, +1
it is also more variable in outcome than Scheme A +2
as you could have very high profits, but also high losses of 5 or lower. +1
Therefore Scheme B is more risky than A, and A the safer option +1
(score max out at 5 full mark, no extra points given)

Problem 4. The figure below appeared in the magazine *American Education* and is intended to represent the age structure of student enrolment to college. The vertical axis reflects the percent of total enrollment; the horizontal axis represents the year of enrollment (i.e., student cohort). 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.
Same specifics can be used to refer to multiple themes.

1. Gestalt principles & illusions +2
   e.g., A closing gap between two separate entities is implied, but it is a bit of an illusion as the two "trend lines" are perfectly interdependent

2. Keep it simple: Decoding & Operations +2
   e.g., Overcomplicated for basically 5 unique numbers

3. Less is more: Chartjunk & data-ink ratio +2
   e.g., No need for 3D and 4 colors as they code for no extra information or variable

4. Graphical data integrity & lie factor +2
   e.g., Deceiving as the 3D+curving effect implies that have continuous enrolment data across years whereas you only have 5 measurement points, two scale breaks that do not contribute, implies trend whereas we only see percentages so we do not know whether more under 25 students are attending as we have no clue about the actual total number of students enrolled (can be growing for both age groups, but not proportional) …

5. Annotation & stand-alone Readability +2
   e.g., Horizontal axis has no label, numbers on red background are hard to read, vertical axis way too small, not clear what colors represent, …

Problem 5.

5.1 Discuss why you agree / disagree with each of the following statements. *(50w each) Marks 2 each*

a) There is a 90% chance that the population statistic falls within the 90% confidence interval around the sample estimate of that statistic.

b) A 90% confidence interval is going to be wider than a 95% confidence interval.

a) Close, but does not hold for one specific sample, +1
the guarantee only holds for the procedure under repeated sampling = in the long run +1

b) 90% CI will be less wide than a 95% CI +1
as the latter needs to be wider to comply with the higher demands on the long run CI coverage of the population statistic +1

5.2 A researcher has the R code below to generate a 95% bootstrap confidence interval for the estimated standard deviation of the variable Y, but is concerned that there might be an error hiding in the code as the interval keeps changing each time he runs the code. Can you explain what's wrong / why this happens? *(75w)* *Marks 5*

```
resample <- function(y){
    n = length(n)
    index = sample(1:n, n, replace = TRUE)
    x = sd( y[index] )
    return(x)
}
replicate(13, resample(data$Y) )
quantile(bootstrapSTAT, c(.025, .975) )
```

2018-2019: bootstrapSTAT assignment on second last line was missing in INSPERA exam question so if student picks up that coding problem 4

Resampling makes use of random number generation so each time run a bit different 4

- …unless if start from same random.seed = 5
- …number of resamples is very low, 13, so the resulting CI will be extra variable 5

These two = Full marks even if random number generation not literally mentioned, it is implied by given these specific answers.

5.3 The formula below provides the asymptotically derived standard error of a regression coefficient b_j in a linear model with J predictors **X** and outcome Y. What are the ideal data/model conditions that would ensure a very high precision for the coefficient b_j. *(90w)*
Marks 5

$$se(b_j) \quad = \quad \sqrt{\frac{1 - r^2_{Y.\mathbf{X}}}{(1 - r^2_{X_j.\mathbf{X}_{-j}})(n - J - 1)}}\,\frac{s_Y}{s_{X_j}}$$

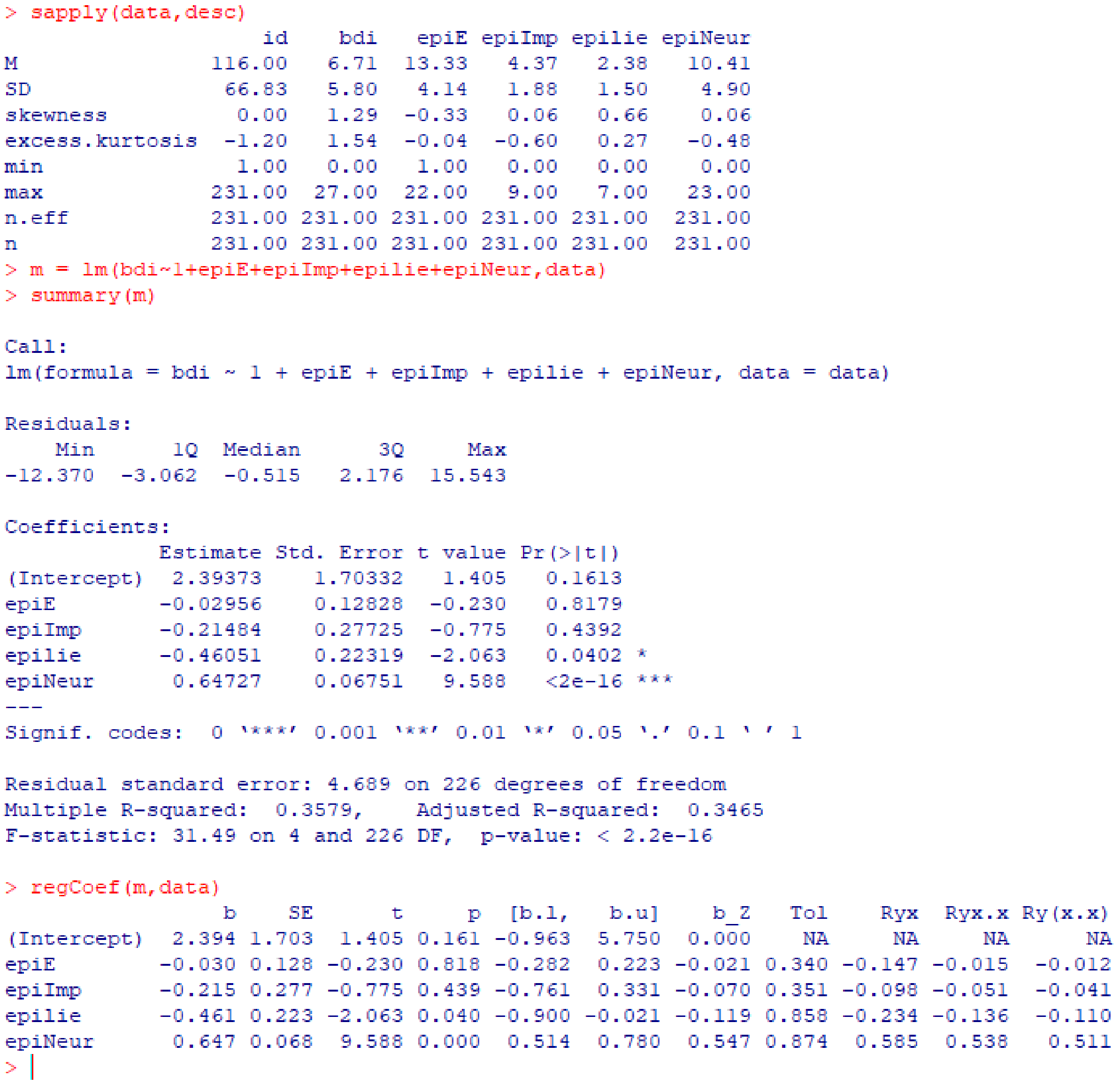
Higher precision is obtained with

1. Higher sample size
2. Lesser variation in Y (relative to variation in the predictor Xj)
3. And vice versa, More variation in Xj (relative to variation in Yj)
4. Larger reduction of prediction error of the linear model as a whole
5. Less predictors (more degrees of freedom)
6. Lesser correlated predictors / lower multicollinearity

One point per point of the above, with max at 5.

Although not directly answering the question, bringing up model assumptions for linear model also earns a point, as one can argue that SE formula only holds if model assumptions hold.

Problem 6. Researchers F. Ar & F. Etched are interested in studying whether depression is to some extent linked to the person’s personality. They have gathered data for a random sample of juvenile delinquents. Depression is operationalized as a score on the Beck Depression Inventory (variable bdi: higher values correspond to a higher self-reported depression). Personality is 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
> m = lm(bdi~1+epiE+epiImp+epilie+epiNeur,data)
> summary(m)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-12.370  -3.062  -0.515   2.176  15.543 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.39373    1.70332   1.405   0.1613    
epiE        -0.02956    0.12828  -0.230   0.8179    
epiImp      -0.21484    0.27725  -0.775   0.4392    
epilie      -0.46051    0.22319  -2.063   0.0402 *  
epiNeur      0.64727    0.06751   9.588   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.689 on 226 degrees of freedom
Multiple R-squared:  0.3579,	Adjusted R-squared:  0.3465 
F-statistic: 31.49 on 4 and 226 DF,  p-value: < 2.2e-16

> regCoef(m,data)
                 b    SE      t     p  [b.l,   b.u]    b_Z   Tol    Ryx  Ryx.x Ry(x.x)
(Intercept)  2.394 1.703  1.405 0.161 -0.963  5.750  0.000    NA     NA     NA      NA
epiE        -0.030 0.128 -0.230 0.818 -0.282  0.223 -0.021 0.340 -0.147 -0.015  -0.012
epiImp      -0.215 0.277 -0.775 0.439 -0.761  0.331 -0.070 0.351 -0.098 -0.051  -0.041
epilie      -0.461 0.223 -2.063 0.040 -0.900 -0.021 -0.119 0.858 -0.234 -0.136  -0.110
epiNeur      0.647 0.068  9.588 0.000  0.514  0.780  0.547 0.874  0.585  0.538   0.511
> |
```

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

Intercept: max 4

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

- Generic interpretation without reference to actual variables 1.5

Generalizability +1

- e.g., P-value/Null hypothesis 0 in pop not rejected OR quite wide CI so pretty uncertain/imprecise estimate

Putting numeric value in context +1

e.g., lowest value on epilie in sample is apparently 1, so such a person does not exist in study sample; comparing value to descriptive statistics summaries, …

Slope: max 6

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

- Generic interpretation without reference to actual variables 2.5
- Causal version 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. CI .5/.8 points and relate that to scale BDI …

Direction +1

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

Putting numeric value in context +1

- E.g., reference to population, descriptive summary statistics

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) "It is mainly neuroticism that is linked to depression."

b) "A significant amount of variation in depression is explained"

(200w) *Marks 10*

a) Max 5

highest r_{yx} .585, +2

still also other variables seem to contribute (e.g., ry.X^2 > ryx^2) +1

also epilie has unique contribution as b = -.46 (.22), p = .040 +2

bold generalization: +2

> e.g., reference to population was omitted in statement, reference to other predictors in model was omitted, …

b) Max 5

Put in actual R-square value +1

Reduction in prediction error is significantly different from 0, +2

yet whether you consider 35% "significant" is another question, +2

there is still RMSEA = 4.689 so quite large margin of error left! +2

Respecify sentence to be more inclusive of actual model it relates to +2

Whether you really "explain" this variation is also a bit of a causal interpretation as we at most can conclude that we found variables = other variation that relate to the variation of focus. +2

So basically + points for comments that make sense, but no credit when contradicted by later comment.

Problem 7. A categorical predictor variable X with four categories (1 to 4) is recoded as 3 dummy variables D1, D2, and D3 according to the following coding system

X	D1	D2	D3
1	0	1	0
2	0	0	0
3	1	1	0
4	1	1	1

The following linear model is formulated: $Y = \beta_0 + \beta_1 D1 + \beta_2 D2 + \beta_3 D3 + \varepsilon$.
What is interpretation of the regression coefficients β_0 and β_1 in terms of category group means? *(50 & 75w)* *Marks 3 each*

b0: intercept meaning expected Y for someone with 0 on D1, D2, and D3, +1
which is basically someone in the group defined by category 2 +1
immediate answer
expected mean on Y for individuals in category group X = 2 3

b1: coefficient of D1 meaning expected mean difference in Y between persons one unit apart on D1 but with similar values on D2 and on D3. +1
Coding scheme is stairs ordered as 2-1-3-4, +1
Immediate answer:
b1: expected mean difference in Y between individuals in category group X=3 and group X=1 3