MAE4000 Data Science 2019 – 2020

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 ≥ 65; B ≥58; C ≥ 51; D ≥ 44; E ≥ 41; F < 41

Problem 1.

1. RZYcode

What does the following R-syntax do? *(140w) Marks 4*

```
data$Z1 = (data$Y==”a”)+(data$Y==”b”)*(-1)
data$Z2 = (data$Y==”c”)+(data$Y==”b”)*(-1)
```

It creates two variables/columns, Z1 and Z2 *+1*
based on the values a third variable Y takes (in an object named data). *+1*
literal description of operations that go into Z1/Z2 *+1*
insight that 3 values are assigned:
Z1 becomes 1 if Y equals “a”, -1 if Y equals “b”, and 0 otherwise. *+1*
Z2 becomes 1 if Y equals “c”, -1 if Y equals “b”, and 0 otherwise. *+1*

2. Merge

A colleague comes to you with the question to merge two datasets that contain data on the same sample of research units. Each dataset has the wide format with 444 rows, with one row corresponding to one research unit.

Dataset1 has the variables IDnumber, Gender, ExamScore 1, and ExamScore2.
Dataset2 has the variables ExamScore2, ExamScore3, and FirstName and LastName.

Explain what question(s) you need to ask your colleague before you can get started to solve their merging problem, and why you would ask the question(s). *(180w) Marks 4*

First of all I would need to verify whether there is a unique identifier variable that I can use to link the two datasets, based on the information it seems that such a variable is lacking. Hence, this would make a correct merger of the two datasets impossible. We do not know which row in the first dataset corresponds to which row in the second dataset. 2

It can be solved if

- the colleague has somewhere a document that links the dataset 1 IDnumber to the dataset 2 First/LastName variables 2
- if they are really really sure that the research units on each row of the two datasets exactly correspond, then the two datasets can just be binded next to eachother 2
- If ExamScore2 has one unique value for each individual 2

Problem 2. Students telling the truth or lying: Believing excuses for missing homework

Mister Hector Det is a teacher in high school and claims that he has a gut feeling that can always tell when students are lying when they make an excuse about their missing homework. This claim is in fact true! Assume that when a student comes up with a homework excuse they are lying about it in 1 out of 6 six cases. Furthermore, when the student tells the truth, Mister Det believes them in 3 out of 4 cases.

3. Tree

Provide a probability tree representation of the information provided on the left that summarizes how well Mister Det's gut feeling can be trusted.

(80w) Marks 7

Tree

Applicants 2400

- Lying 400
  - \Believed 400
  - Believed 0
- NotLying 2000
  - \Believed 500
  - Believed 1500

Branching system 4 points: 1 per hierarchical branching variable

Frequencies correct 3 points:

Each sub branching needs to sum up to total branch before

4. Pr

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(Believe | Lying) = 0 / 2400

Pr(Believe ∩ \ Lying) = 1500 / 2400

Pr(Lying U \ Believe) = 900/2400

Pr(Believe) = 1500 /2400

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

Problem 3.

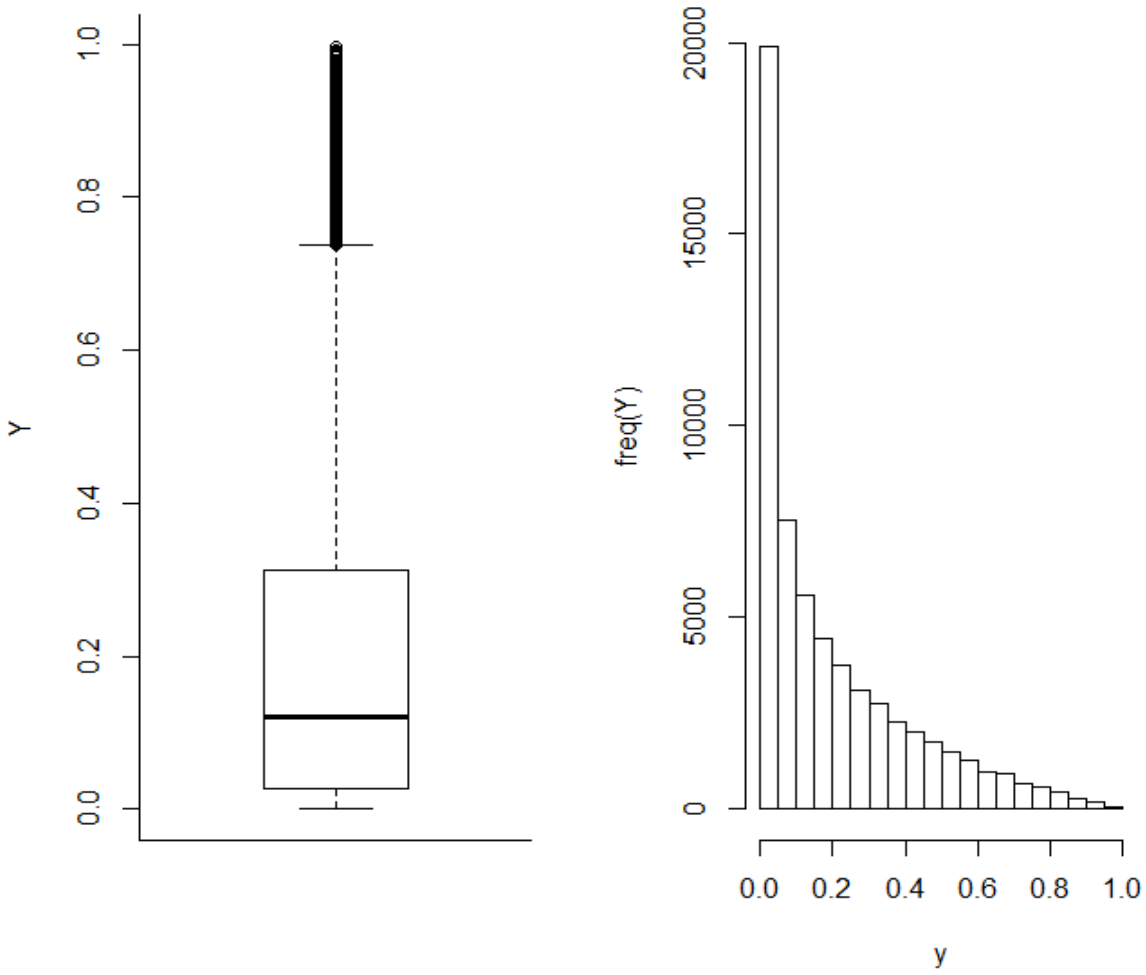


Figure 1. Distribution of variable Y.

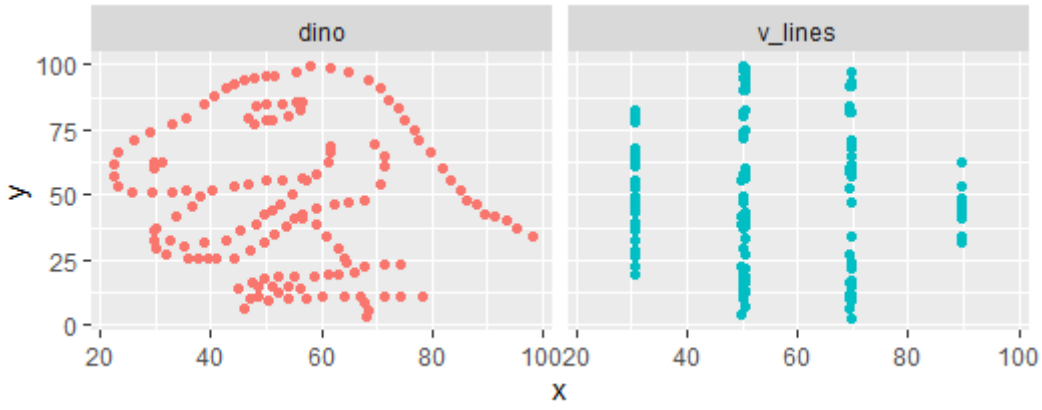


Figure 2. Scatterplots of variables X and Y for two datasets.

5. Descriptive statistics (1)

On the left in Figure 1 you see two graphical representations of the variable Y's distribution. This variable has an excesss kurtosis of either

- -.15 or
- .85

and a skewness of either

- -1.25,
- .25, or
- 1.25.

Pick out the correct value for both descriptive statistics and explain how you arrived at your two choices. *(100w)* *Marks 4*

Excess kurtosis of .85 +1
as the distribution tends to have more weight in the tails than does the normal distribution +1
Skewness of 1.25 +1
because there is a pronounced tail on the right side (higher values on Y) of the mass of the distribution of Y (located around .05 to .3). +1

6. Descriptive statistics (2)

Figure 2 on the left shows catterplots of the variable X and Y for the datasets dino and v_lines.

- These two datasets have the same means, standard deviations, and correlation for both variables?

Generally speaking, how is it possible that two datasets can be clearly different, but still have similar values for some descriptive statistics? *(50w)* *Marks 3*

Descriptive statistics make a summary of the data and while making a summary some more detailed aspects can get lost. This is what is happening here, the summary is similar, but the underlying dataset is different. 3
Likely if you compute different descriptive statistics then the ones mentioned, that code for different aspects of the data, their differences would be picked up. +1

7. Correlation

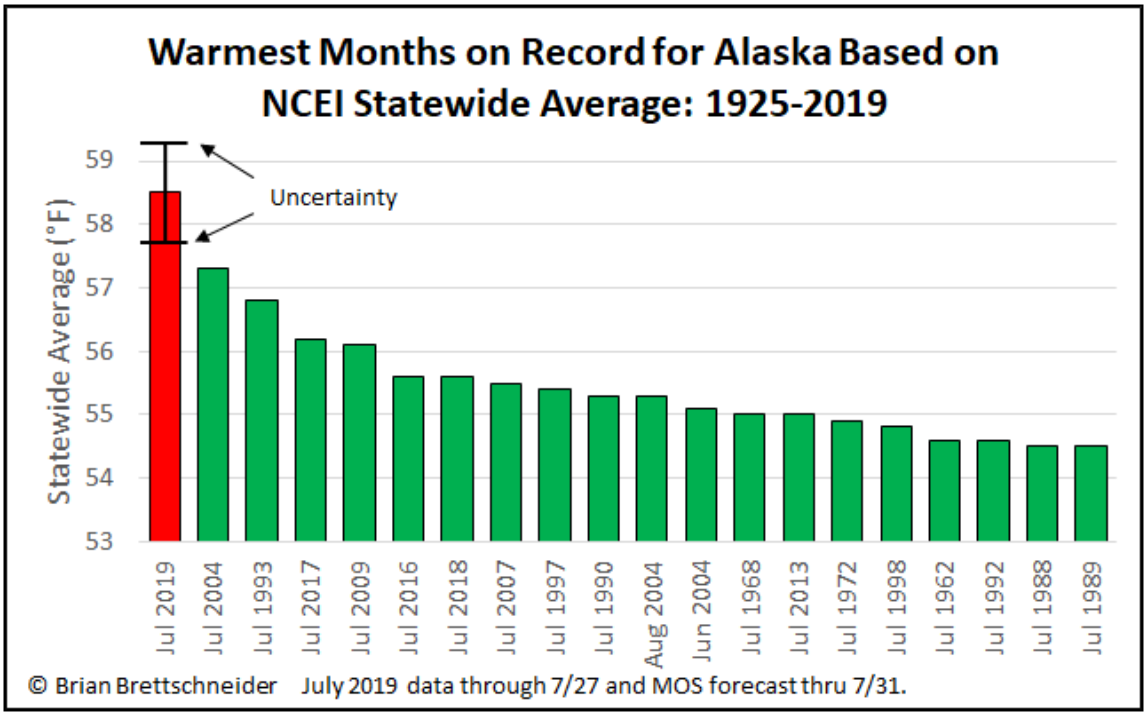
What is the difference between a Pearson linear correlation and a Spearman rank correlation.
They are both measures of covariation and have the same formula and range of values [-1,1] +1
except that the former is computed on the original scores and the latter on the rank transformed scores. +1
Pearson codes for linear relations, Spearman for more general monotone relation (higher X ~higher Y) +1
The rank transformation just orders values from small to large and looses the underlying metric. +1
This loss of unit does make the rank correlation less sensitive to the shape of the marginal distributions of the variables and more resistant against extreme outliers. +1

Problem 4.

8. Summer in Alaska

The figure on the left appeared on Twitter to illustrate that Alaska has experienced extreme heat this summer, with the July statewide average temperature apparently shattering the previous record.

Evaluate the figure and discuss its potential 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 proper supporting evidence. Exact wording/terminology of themes is not essential as long as it is clearly implied.

1. Gestalt principles & illusions +2
   e.g., Months are ordered in terms of temperature which clearly communicates that the one month is the highest, yet at the same time some type of downward trend seems to be implied, but this is not what was intended to be communicated; the color coding does clearly set apart the one measure in 2019 which seems in line with the intended message...

2. Keep it simple: Decoding & Operations +2
   e.g., Figure is relatively simple to perceive. The color coding is perhaps a bit redundant, but see above. It would have been good to supply a reference line providing a running average of prior summers' temperatures to allow for direct comparison...

3. Less is more: Chartjunk & data-ink ratio +2
   e.g., Why only have one uncertainty interval and where does it come from, what is it and what is its added value?...

4. Graphical data integrity & lie factor +2
   e.g., seems to imply a trend across time but measures themselves are not properly ordered across time on the X-axis and the distance between the different points are equal-sized although the time intervals are not constant. Not all data is displayed as range starting from 1925, but not shown in graph...

5. Annotation & stand-alone Readability +2
   e.g., proper Y-axis label including units of measurement, but X-axis label is lacking; Red/green color coding lacks legend + these colors are a bad choice given color blindness; Informative title that does also provide data source, with subnote on measure for 2019 (year not finished yet). ...

Problem 5
Discuss why you agree / disagree with each of the following statements. Marks 4 each

9. CI
A confidence interval for an estimated parameter conveys more information about the parameter than a p-value from a null hypothesis test for the same estimated parameter. *(120w)*

Agreed, it provides information on magnitude of the parameter estimate by means of the interval's location +2

and also of its corresponding uncertainty by means of the width of the interval. +1

You can even use it for an implicit significance test if you check whether the interval overlaps with the value posited in the null hypothesis (if not then reject null hypothesis). +1

If nothing except definition of the two concepts: +2?

10. Type I and II
When testing hypotheses, we can end up making the wrong decisions. Your colleague tells you that a Type-II error is worse than a Type-I error, but that simultaneously making a Type-I and Type-II error for the same hypothesis test is of course the worst. *(210w)*

- *Define what a Type I and II error are in your discussion.*

A type-I error happens when we reject the null hypothesis based on our sample estimate when it is in fact true in the population; +1

A type-II error happens when we fail to reject the null hypothesis when it is false in the population. +1

The null hypothesis cannot be both true and false at the same time, so it is impossible to make a simultaneous type-I and II error for one and the same hypothesis test. +1

Finding something that's not there (Type-I) has historically been considered the worst error of the two, but it depends on the specific field and consequences of the errors in inference we would make. (so reverse also fine when arguments given) +1

11. SE Under which conditions will the linear model parameter, for which the standard error formula is shown below, be less/more precise? Do not forget to start your explanation with stating what the parameter and the different symbols stand for. *(250w)* Marks 6

$$SE(\hat{\mu}_{Y|X=x_i}) = \sigma_{error}\sqrt{n^{-1} + \frac{[x_i - \bar{x}]^2}{SS_X}}$$

This is the standard error of the estimated mean of Y conditional on X being equal to the value that research unit *i* has on this X variable. 2

Sigma error is the standard deviation of the residual also known under the acronym the RMSE. Hence, the less error variation, the better our linear model's predictions, and the more precise the conditional average will be estimated. 1.5

If sample size n increases, and we have in other words more data, it will become possible to estimate the conditional average more precisely. 1

The further away the X value of the research unit is from the average value, in other words the more extreme this research unit *i* is on X, the less precise the conditional average can be estimated. 1.5

Problem 6. Researchers F. Ar & F. Etched have gathered BIG 5 personality questionnaire data for a random sample of individuals that got married not more than three years ago.

- O: Openess to new experiences; C: Conscientiousness; E: Extraversion; A: Agreeableness; N: Neuroticism
- Gender: M = male, F = female

Selected output of a data-analysis in R is given below:

```
> sapply(data[,-c(1:3)],desc)
                     O      C      E      A      N
M                29.27  30.83  20.56  35.19  18.87
SD                5.65   7.16   5.26   3.15   6.17
skewness         -0.16  -0.59   0.32  -0.32   0.58
excess.kurtosis  -0.50  -0.31  -0.60   0.38   0.31
min              13.00  11.00  11.00  25.00   6.00
max              42.00  42.00  35.00  42.00  40.00
n.eff           198.00 195.00 194.00 193.00 195.00
n               200.00 200.00 200.00 200.00 200.00
> m = lm(O ~ E+C*Gender,data)
> summary(m)

Call:
lm(formula = O ~ E + C * Gender, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-16.1843  -4.0822   0.5126   3.6288  13.7041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.59880    3.35826   9.707   <2e-16 ***
E           -0.09758    0.07763  -1.257   0.2104
C           -0.05791    0.09129  -0.634   0.5266
GenderM     -7.45237    3.76795  -1.978   0.0495 *
C:GenderM    0.28369    0.11851   2.394   0.0177 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.493 on 182 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.06448,   Adjusted R-squared:  0.04392
F-statistic: 3.136 on 4 and 182 DF,  p-value: 0.01595

> regCoef(m,data)
                b    SE      t     p   [b.l,   b.u]    b_Z   Tol    Ryx  Ryx.x Ry(x.x)  R2yx R2y(x.x)
(Intercept) 32.599 3.358  9.707 0.000  25.973 39.225  0.000    NA     NA     NA      NA    NA       NA
E           -0.098 0.078 -1.257 0.210  -0.251  0.056 -0.091 0.987 -0.098 -0.093  -0.090 0.010    0.008
C           -0.058 0.091 -0.634 0.527  -0.238  0.122 -0.074 0.375  0.115 -0.047  -0.045 0.013    0.002
GenderM     -7.452 3.768 -1.978 0.049 -14.887 -0.018 -0.665 0.045  0.072 -0.145  -0.142 0.005    0.020
C:GenderM    0.284 0.119  2.394 0.018   0.050  0.518  0.774 0.049  0.145  0.175   0.172 0.021    0.029
> m2 = lm(O ~ E+scale(C,scale=FALSE)*Gender,data)
> summary(m2)

Call:
lm(formula = O ~ E + scale(C, scale = FALSE) * Gender, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-16.1843  -4.0822   0.5126   3.6288  13.7041

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      30.81328    1.66308  18.528   <2e-16 ***
E                                -0.09758    0.07763  -1.257   0.2104
scale(C, scale = FALSE)          -0.05791    0.09129  -0.634   0.5266
GenderM                           1.29409    0.83733   1.545   0.1240
scale(C, scale = FALSE):GenderM   0.28369    0.11851   2.394   0.0177 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.493 on 182 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.06448,   Adjusted R-squared:  0.04392
F-statistic: 3.136 on 4 and 182 DF,  p-value: 0.01595
```

12. Coefficients

Based on the model results, give an interpretation (direction, size, generalizability, & meaning) of the intercept and the regression coefficient of GenderM in model m. *(500w)* *Marks 10*

Intercept: **max 4**

Base = Expected O score for a Female with 0 Conscientiousness and 0 E 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 any of the BIG5 is 6 in the sample, so a person with zero score does not exist in study sample; comparing value to descriptive statistics summaries, mean O is 29.27 and intercept is less than 1SD above that value …

Slope: **max 6**

Base = Expected difference in O score between Male and Female individuals that are identical in Extraversion but have a zero score on Conscientiousness 3.5

- Generic interpretation without reference to actual variables max 2.5
- Causal version max 2.5
- Missing interaction max 2.5
- Missing identical max 1.5

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 that within this type of personality group (C=0) females are more open to new experiences than males.

Putting numeric value in context +1

- E.g., reference to population, descriptive summary statistics, 7.5 points is quite a difference as it is above 1SD on O, but then again a person with 0 score on C does not exist, so this parameter is not very meaningful in this context.

Round up for final mark

13. Statements

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) After centering the Conscientiousness predictor, the p-value of the corresponding regression coefficient is no longer significant, indicating that after all Gender is not related to how Open to experience individuals are.

b) Because E has a tolerance of .987, removing it would hardly influence the model results.

(250w) *Marks 10*

a) Max 5

first-order effect of C was already not significant and remains not significant after centering +1

first-order effect of C doesn't tell anything about Gender +2

No: Conscientiousness and Gender are still involved in the interaction term which is significantly different from zero and hence has some influence on the predictions of O +3

No there is just no Gender difference that can be supported to be statistically significantly diferent from zero for individuals that are of average C and of identical E +2.5

We could agree, but other reasons; the relative reduction of the prediction error by the whole bunch of predictors including Gender is rather limited (6%), then again we are trying to predict individual differences in human behavior, which is not an easy challenge +1

b) Max 5

Tolerance is % of variation in E not accounted for by other predictors, so this actually implies that E is almost not correlated with the other predictors. + 2

We see E is also hardly correlated with O, Rxy. + 1

Removing E from the model would have not a big impact as it is neither related to the outcome nor the other predictors, and hence has no way to impact the predictions. + 2

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

14. Reading Comprehension and Gender (1) Consider that we have estimated a linear model in which a pupil's score on a Reading Comprehension assessment is the outcome and the Gender of the student is the predictor. Gender has been effect-coded: -1 for Male and 1 for Female.

ReadingComprehension = \beta_0 + \beta_1 Gender + \varepsilon

What is the interpretation of regression coefficient beta_1 in terms of category group means? If female students score on average a 8 on the assessment and male students on average score a 7 on the assessment, what is then the corresponding value for \beta_1?

(50w) *Marks 2*

Expected difference in reading comprehension score between female students and the grand mean, the latter mean being the average of the average scores for both gender groups. 1
Its corresponding value would then be 8 – (8+7)/2 = .5 1

15. Reading Comprehension and Gender (2)

If in a model predicting reading comprehension based on gender of the student, the null hypothesis that the regression coefficient of the gender predictor is equal to zero in the population would be rejected, can we then conclude that

gender has an effect on reading comprehension?

Explain why you agree / disagree with such a conclusion.

(180w) *Marks 5*

No this conclusion would imply causality as if we could ever change a person's gender to improve his/her reading comprehension. 2

Furthermore, rejecting the null hypothesis implies that we conclude that the estimated difference for our sample is significantly different from zero; it does not tell us whether this is a big or important difference. 1

We could conclude that the gender and reading comprehension are related but this could very well be a spurious correlation due to all other kind of relevant predictors that we are not taking into account. 2

Replication needed +1

Possibility of making wrong decision +1