Trial Exam ADA WS 2021/22

Part 1

1. If Missing Values of INCOME depend on AGE, they are [1]
   - ◯ Missing completly at Random
   - ◯ Not Missing at All
   - √ **Not Missing at Random**
   - ◯ Missing at Random

2. You want to test if men and woman have different income. What kind of test should yu use? [1]
   - ◯ Z-score test
   - ◯ 1-tailed t-test
   - √ **2-tailed t-test**
   - ◯ F-test

3. What does the Central Limit Theorem states? [1]
   - √ **That the distribution of sample means approximates a normal distribution as the sample size get larger, regardless of the population's distribution.**
   - ◯ That the distribution of the sample means approximates a normal distribution for only a sample size of $n < 30$, regardless of the population's distribution.
   - ◯ That the distribution of sample means approximates a normal distribution as the sample size get larger, only if the population's distribution is a normal distribution.
   - ◯ That the distribution of the sample means approximates a normal distribution for only a sample size of $n = 30$, regardless of the population's distribution.

4. What statement of the Log transformation is correct? [1]
   - ◯ The log transformation is used to prepare the data for further use. It replaces missing values with better guessed values.
   - ◯ The log transformation is only used to make left skewed distributions normally distributed. It allows to use powerful statistical procedures that only apply if the data is normally distributed.
   - ◯ The log transformation can be used to make not skewed distributions more skewed. It allows better interpretation of the data.
   - √ **The log transformation is used to make highly skewed distributions less skewed/normally distributed. It allows to use powerful statistical procedures that only apply if the data is normally distributed.**

5. You can handle missing values by deleting the subjects with mussing values. [1]
   - ◯ TRUE, but only if the missing values are nominal scaled.

√ **TRUE, but if you can't fill the missing spots with sample means.**

◯ TRUE

◯ FALSE

6. Which statement about descriptive analysis is **wrong**? 1

◯ Univariate analysis deals with only a single attribute and is mainly used for paramters of location and parameters of dispersion.

◯ The interpretability of histograms is strongly dependent of the number of bins and the width of each interval.

√ **The frequency distribution can only be represented by a histogram.**

◯ Bi-variate analysis deals with the relationship between two varaibles and is mainly used for measurements of correlation.

7. The attribute *size* has the values "small", "medium", "large". What level of measurement is represented? 1

√ ***Size* is based on an ordinal scale.**

◯ *Size* is based on an nominal scale.

◯ *Size* is based on an interval scale.

◯ *Size* is based on an ratio scale.

8. What is a hypothesis? 1

◯ A research question the results will answer.

◯ A statistical method for calculating the extend to which the results could have happend by chance.

◯ A theory that underpins the study.

√ **A belief concering a parameter that the researcher wants to test through the data collected in a study.**

9. Multiple linear regression is used to: 1

◯ Describe the relationship between one dependent variable and one independent variable.

√ **Describe the relationship between one dependent variable and multiple independent variables.**

◯ Describe the relationship between multiple dependent variables and one independent variable.

◯ Describe the relationship between multiple dependent variables and multiple independent variables.

10. The value for Model Sum of Squares (SSM) describes: 1

◯ The distribution of the residuals.

√ **The total variance of the data.**

◯ The explained variance.

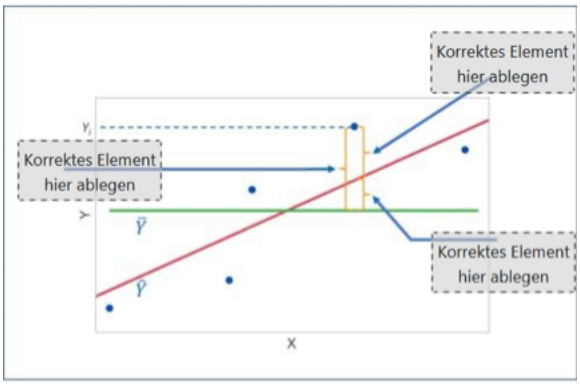
◯ The unexplained variance.

Part 2

You found an old data set from 1981 online. Since you have always paid attention in the "Applied Data Analysis" excercise, it was no problem for you to import this data set into R under the variable name **data**. In this second part of the exam, we will mainly deal with this data set. The following table shows the first 5 lines of the 100-observation data set.

gender	age	education	height	weight	IQ	ID
0	34	1	189	99	100	1
1	56	2	156	54	110	2
0	21	3	173	67	107	3
2	45	2	178	71	122	4
1	32	2	171	69	98	5

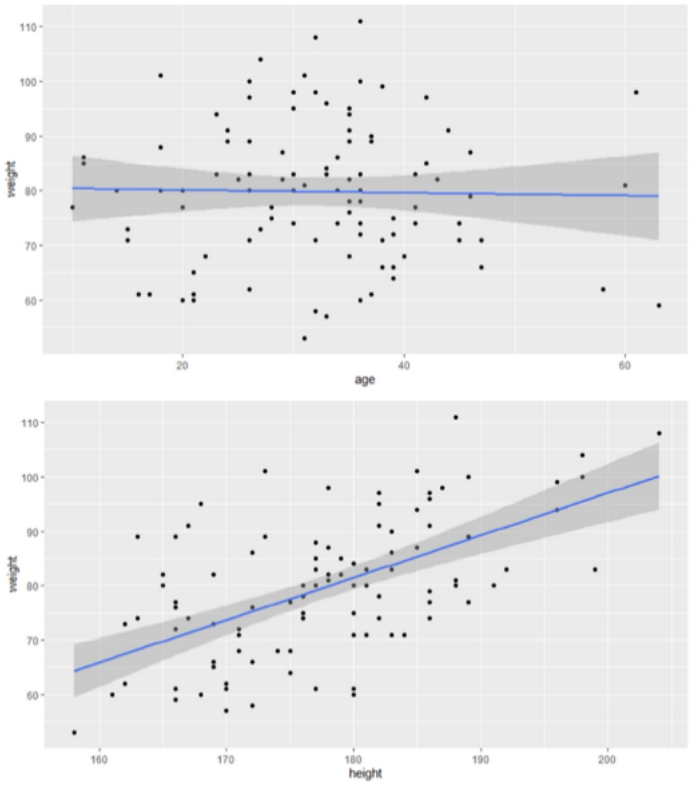
11. To check the fit of a linear model, different values are determined. The following figure shows one step of a common procedure. Assign the labels to the illustration: [1½]



- Total error
- Residual
- Explained error

Solution: Total error = Residual + Explained error

12. We want to use another numeric variable to estimate the weight of a person. We choose the height and the age as possible candidates for the independent variable. Look at the two plots and decide which variable is more suitable as estimator for the weight. Explain you choice. [3½]



Solution: Height is more suitable because slope $\neq 0 \Rightarrow$ correlation between height and weight

13. What is the purpose of R-squared? How is this value calculated abd what does is express? Describe the role played by the values "Total Sum of Squares" (SST), "Model Sum of Squares" (SSM) and "Error Sum of Squares" (SSE) in this context. Also name another measure to determine the quality of a linear model. [4]

Solution:

$$R^2 = \frac{\text{Explained Error}}{\text{Total Error}} = \frac{SSM}{SST}$$

Measure for goddness of fit, value of 1 indicates perfect fit, value of 0 indicates no fit at all. R^2 increase in multiple linear models with the number of parameters, better measures will also include the number of paramters like AIC.

14. Next we want to create linear models with R and assess some metrics to describe the fit between the models and the data. One model should use the height and the other should use the age as independet variable. Please create a short R-script to create the models and to output some metrics for these models. [7]

Solution:

```r
model1 = lm(weight ~ height, data = data)
model2 = lm(weight ~ age, data = data)

summary(model1)
summary(model2)
```

15. For the two linear models to be created in R previously, R would give the following two outputs in the console. Which output belongs to which independent variable (height or age)? Evaluate the fit of the models using the metrics. Interpret R-squared in particular. 4

1.

```
Residuals:
     Min       1Q   Median       3Q      Max
-26.8104  -8.4950  -0.0163   9.0223  31.3183

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 80.60849    4.18465  19.263   <2e-16 ***
data$age    -0.02575    0.12377  -0.208    0.836
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.84 on 98 degrees of freedom
Multiple R-squared:  0.0004413, Adjusted R-squared:  -0.009758
F-statistic: 0.04327 on 1 and 98 DF,  p-value: 0.8356
```

2.

```
Residuals:
     Min       1Q   Median       3Q      Max
-21.4616  -9.0473   0.6457   7.2849  24.9881

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -58.6751    19.9385  -2.943  0.00406 **
data$height   0.7785     0.1120   6.954    4e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.51 on 98 degrees of freedom
Multiple R-squared:  0.3304,    Adjusted R-squared:  0.3236
F-statistic: 48.36 on 1 and 98 DF,  p-value: 4.003e-10
```

Solution: output 1: weight $\sim$ age
output 2: weight $\sim$ height, better R^2 (0.3304 vs. 0.0004413 for other model) $\Rightarrow$ loking at significance for parameters, age is not significant at all, height is significant ($***$ and p-value for height $= 0$ is $4 \cdot 10^{-10}$)