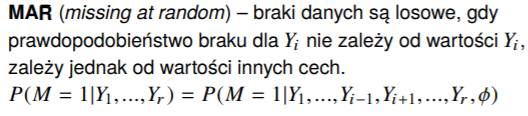
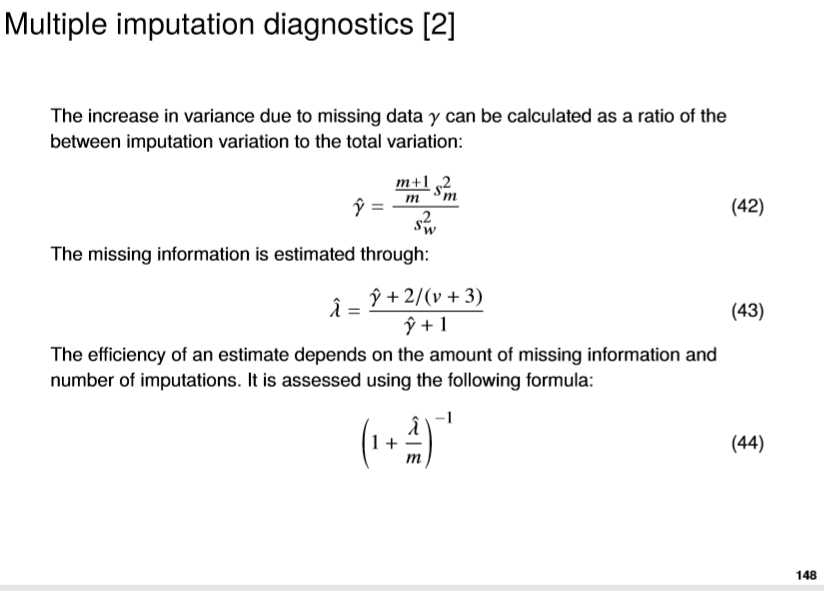
1. Explain what is predictive modeling? Give the basic criterion for assessing predictive models?
2. Describe the validation process for predictive modeling?
3. Explain what is metadata and what kind of meaning it has in terms of evaluating quality of data?
4. Name and provide formulas describing two rules for detecting outliers
5. The mechanism of formation of missing data on the variable Yi is described below by the formula (M denotes elements of the vector of missing data on Yi). Indicate the name of the missing data mechanism that corresponds to this entry. What is the practical meaning of this assumption?

****

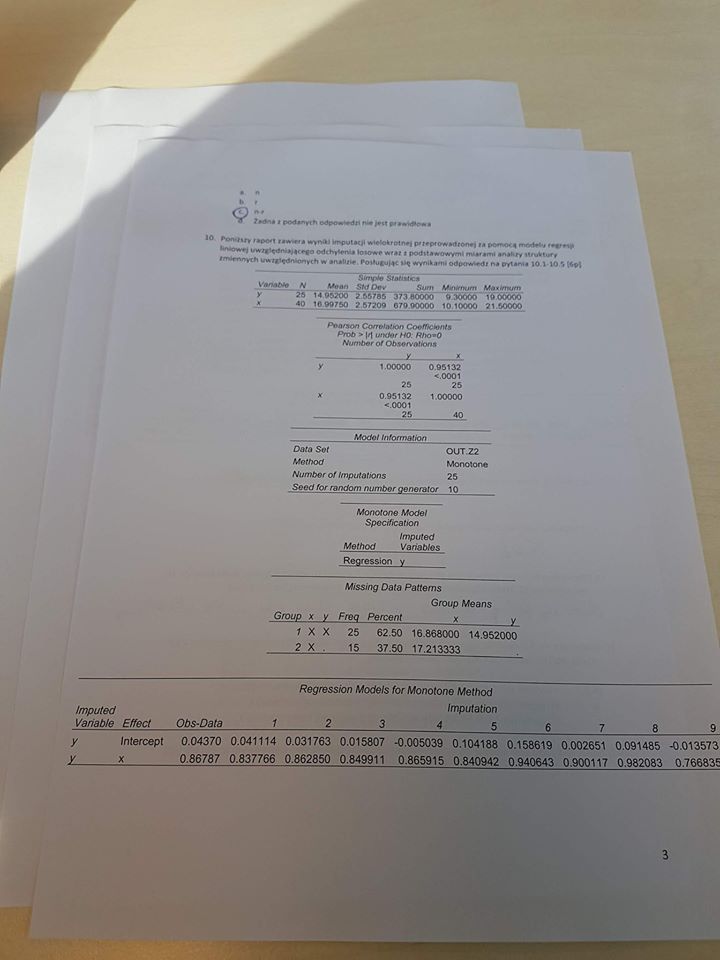
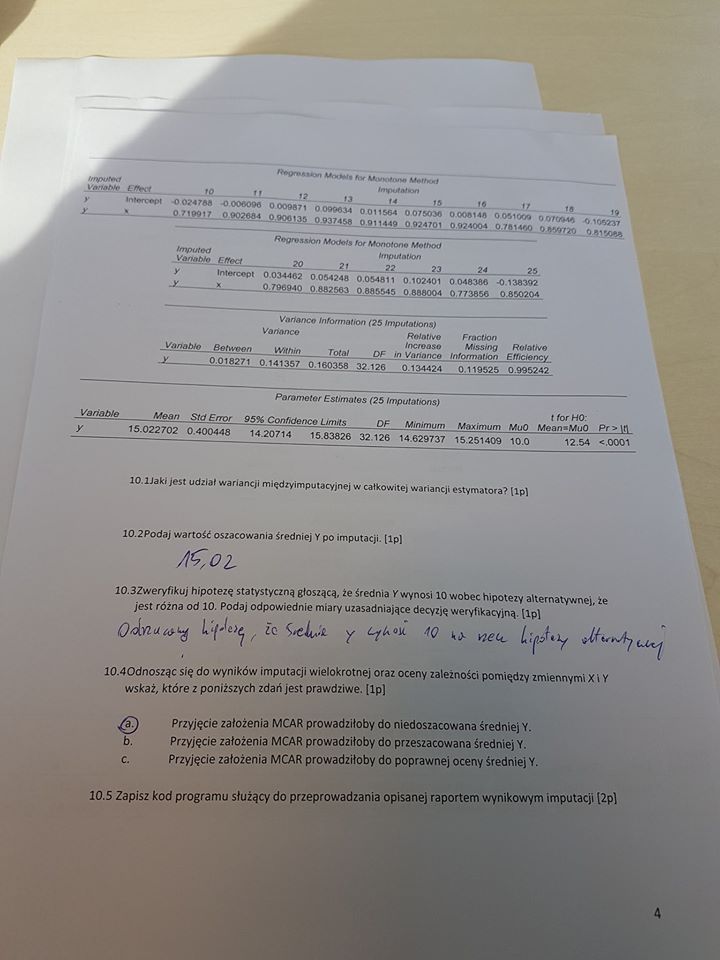
1. List and explain the two components of variance in the multiple imputation method?
2. Enter the statistics that you use to select the number of imputations in the multiple imputation method. Enter the range of variation and interpretation of this statistic?

**Relative Efficiency?**



1. Enter a name and describe the basic steps of the algorithm used to determine the most reliable estimates for incomplete data

**EM algorithm Expectation-Maximization**

1. What sample size do we take in order to conclude the statistical model whose estimation was obtained by using data supplementation with the highest reliability method? (n is the assumed size of the sample, while r <n is the number of respondents who provided answers in the survey). Choose the correct answer:
2. ***N***
3. **R**
4. **n-r**
5. **none of the above**
6.  
   1. What is the contribution of between imputation variance to the total estimator variance? [1p]

0.018271 / 0.160358 = 0.113938

10.2 Provide the estimated mean value of Y after imputation.

10.3 Verify the hypothesis that the average Y is 10 versus the alternative hypothesis that it is different from 10. Provide relevant measures justifying the verification decision

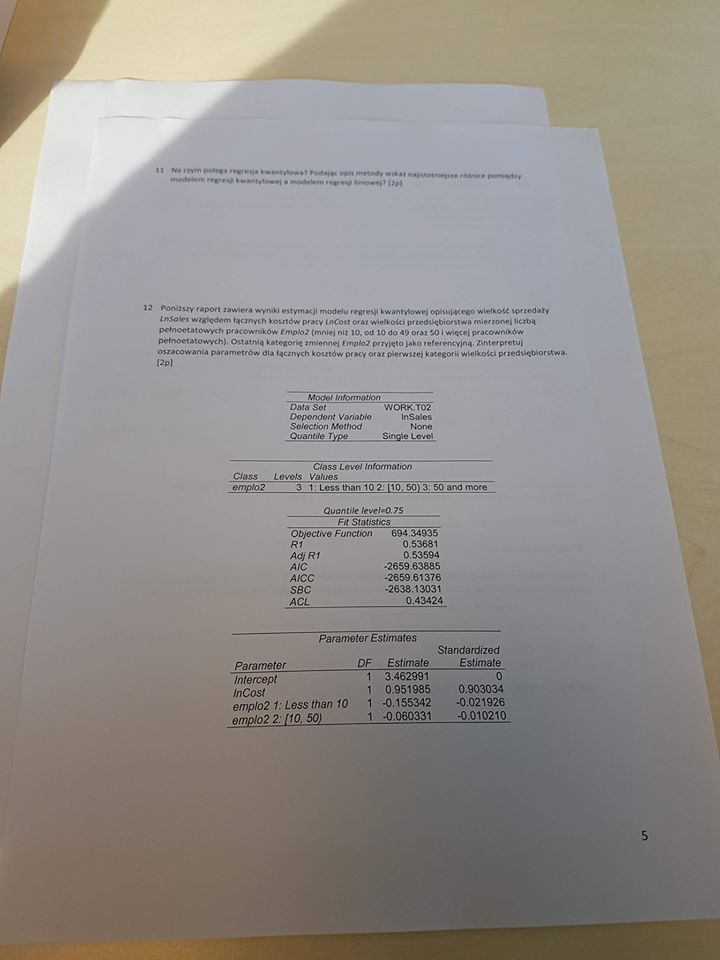
proc mi data=OUT.Z2 seed=10 mu0=10 ;

Monotone reg(y);

var x y?;

run;

1. What is quantile regression? By providing a description of the method, indicate the most significant differences between the quantile regression model and the linear regression model.
2. The following report contains the results of an estimation of the quantile regression model describing the sales volume of LnSales in relation to the total labor costs of LnCost and the size of the enterprise measured by the number of full-time Emplo2 employees (less than 10, from 10 to 49 and 50 and more full-time employees). The last category of the Emplo2 variable was taken as reference. Interpret parameter estimates for total labor costs and the first category of enterprise size.



1. Which of the assumptions is false in relation to the models estimating CLV class "always a share" [1p]

a) do not assume that the lack of activity on the part of the client means his permanent departure

b) applies in cases where the analysis should include clients who have become active

c) an example is the generalized retention model

d) an example is the “recency-frequency” migration model

1. What is the probability that a new customer of the company (bought the product or used the service of the company at time 0), will be active in the next period in the case of the transition matrix of the form below? We assume that the first state is "buy" and the second state is "no-buy".

**P= (0,2 0,8)**

**(0,1 0,9)**

1. **0,2**
2. **0,8**
3. **0,3**
4. **0,9**
5. Two retention models were built estimating the CLV value: a simple and generalized model. The simple model gave a more optimistic (higher) estimate than the generalized model. Indicate what this difference may result from. Will the estimates obtained with the simple model usually be higher?
6. Among the types of retention and migration models discussed, give an example of one deterministic and one probabilistic model estimating the value of CVL. [2p]

a) deterministic model: customer rent model

b) probabilistic model: generalized retention model

1. A segmentation model was built using PROC FASTCLUS. The following statistics were obtained in the results. Indicate which value corresponds to the model error. Justify your choice?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1. **Statistics for Variable** | | | | |
| **Variable** | **Total STD** | **Within STD** | **R-Square** | **RSQ/(1-RSQ)** |
| **x** | 2.06854 | 0.89098 | 0.823609 | 4.669219 |
| **y** | 1.02113 | 1.00352 | 0.039039 | 0.040683 |
| **OVERALL** | 1.63119 | 0.93959 | 0.669891 | 2.029303 |

Overall Within STD

1. Write the program code that generates the non-parametric Kaplan-Meier retention model, and saves the estimated survival function values to an output file called outsurv in the WORK library. The dataset called retention is in the WORK library, the variables mean the length of the relationship duration is rel\_lenght, and the censorship variable is censor, where 1 means no censorship, and 0 means censored observations. In addition, in the code, include a test on the significance of the impact of the source quality variable (customer acquisition source) on survival values

proc lifetest data=work.retention outsurv=outsurv;

time rel\_lenght\*censor(0);

freq count;

strata source;

run;

1. After performing the test on the significance of the impact of the *source* quantitative variable, the following results were obtained. Interpret the estimates below, indicating the null and alternative hypothesis

|  |  |  |  |
| --- | --- | --- | --- |
| **Test of Equality over Strata** | | | |
| **Test** | **Chi-Square** | **DF** | **Pr > Chi-Square** |
| **Log-Rank** | 25.4037 | 3 | <.0001 |
| **Wilcoxon** | 19.4331 | 3 | 0.0002 |
| **-2Log(LR)** | 33.9343 | 3 | <.0001 |

Ho: Survival value is the same for all groups, regardless of the source of the customer (on the value of the source variable).

H1: Survival value varies depending on the source of the customer acquisition.

Assuming the significance level α = 0.05, we reject the null hypothesis (because the values of Pr> Chi-Square for all tests <α) in favor of the alternative hypothesis, i.e. the survival value varies depending on the source of customer acquisition.