QUALITY DATA ANALYSIS

09/02/2024

General recommendations:

- Write the solutions in CLEAR and READABLE way on paper and show (qualitatively) all the relevant plots.
- Avoid (if not required) theoretical introductions or explanations covered during the course.
- Always state the assumptions and report all relevant steps/discussion/formulas/expression to present and motivate your solution.
- When using hypothesis tests provide the numerical value of the test statistic and the test conclusion in terms of p-value.
- Exam duration: 2h
- **For multichance students only: you can skip Exercise 2, point 4), Exercise 3, question 2).**

Exercise 1 (14 points)

The measured diameters of the shafts produced over two shifts are reported in `diameter_phase1.csv`. The columns of the table report a sequential index ('idx' column), the measurements in mm ('diam' column) and the shift ('shift' column) at which the data were collected.

1) Find an adequate model to fit the data.
2) Design the appropriate control charts to monitor the diameter of the shafts (use K = 3). *Note: in case of violations of control limits, assume no assignable cause was found.*
3) Using the control chart(s) designed in point 1 (phase 1), check if the data collected for the following 20 shafts produced during shift 2 (stored in `diameter_phase2.csv`) are in control. Report the index of the OOC points, if any.

Exercise 2 (15 points)

A German company operating in the aerospace sector needs to monitor the manufacturing process for a new type of titanium bracket. The quality characteristic of interest is the Brinell hardness. Four hardness measurements are performed in four pre-defined locations of the component, and parts are randomly picked up from the shop floor and inspected every two hours. Data to be used for control chart design are reported in the file `AERO_phase1.csv`. Each column refers to one location where the hardness measurement is performed. *Assume the measurements within each sample were performed following the same order shown in the provided table.*

1) Check the assumptions and discuss the result.
2) Design a statistical test to check if the hardness in location 1 is statistically higher than the hardness in location 2. Discuss the result.
3) Design a suitable univariate control charting method for these data, using K = 3. *In case of violation of control limits, assume no assignable cause is known.*
4) Using the control chart designed in point 3) determine if the new Brinell hardness measurements in `AERO_phase2.csv` are in control or not. Discuss the result.

Exercise 3 (4 points)

Question 1)

The Shewhart control chart for the mean of a process, where the data are known to be normally distributed, sets the control limits (i.e., LCL and UCL) to be K standard deviations away from the center line, for some constant, $K > 0$. If a user will decide to replace K by $K_1 > K$, then which of the following statements will be valid for the control chart performance:

a) The false alarms will increase.
b) The false alarms will decrease.
c) The out-of-control detection power will increase.
d) We cannot tell from the above information only.

Question 2)

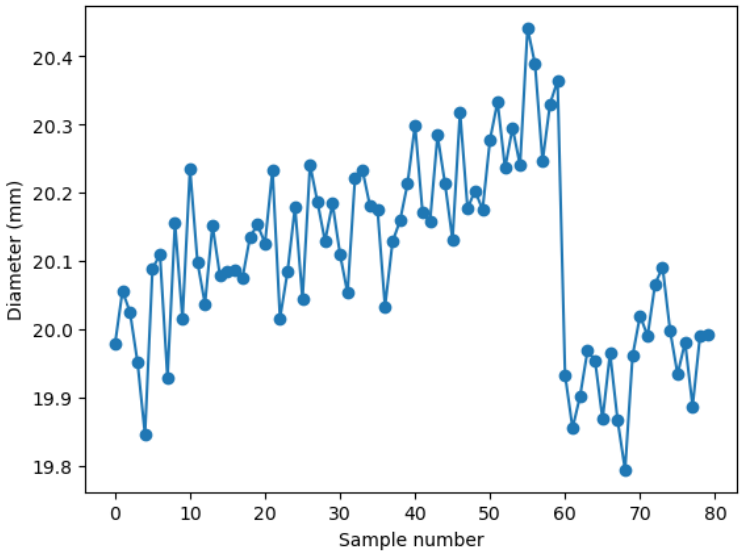
On a data set, we run the linear model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$ and we derive the ANOVA table for the fitted model. If the overall model's F-ratio has a p-value that is smaller than the predetermined level of significance (alpha) for this problem, then which of the following statements will be valid?

a) All model coefficients $(\beta_1, \beta_2, \dots, \beta_k)$ are statistically significantly different from zero.
b) All model coefficients $(\beta_1, \beta_2, \dots, \beta_k)$ are not statistically significantly different from zero.
c) At least one of the model coefficients $(\beta_1, \beta_2, \dots, \beta_k)$ is statistically significantly different from zero.
d) At most one of the model coefficients $(\beta_1, \beta_2, \dots, \beta_k)$ is statistically significantly different from zero.

Exercise 1 solution

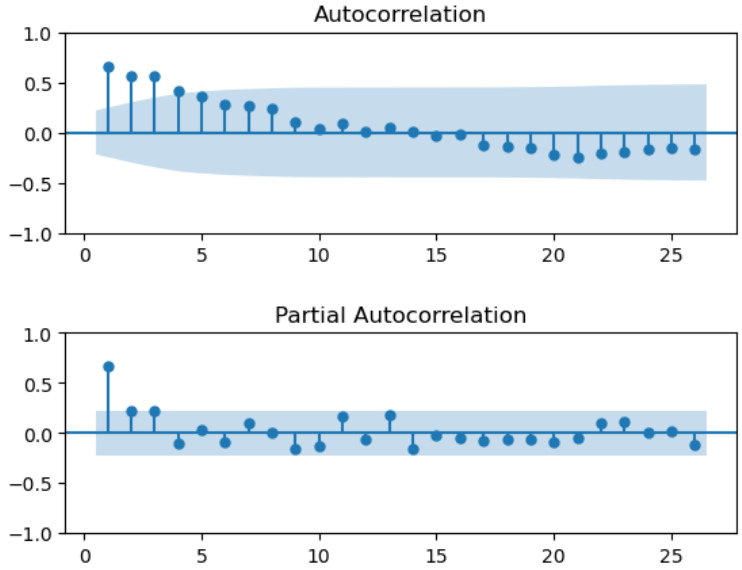
1)

Import the data and plot them.



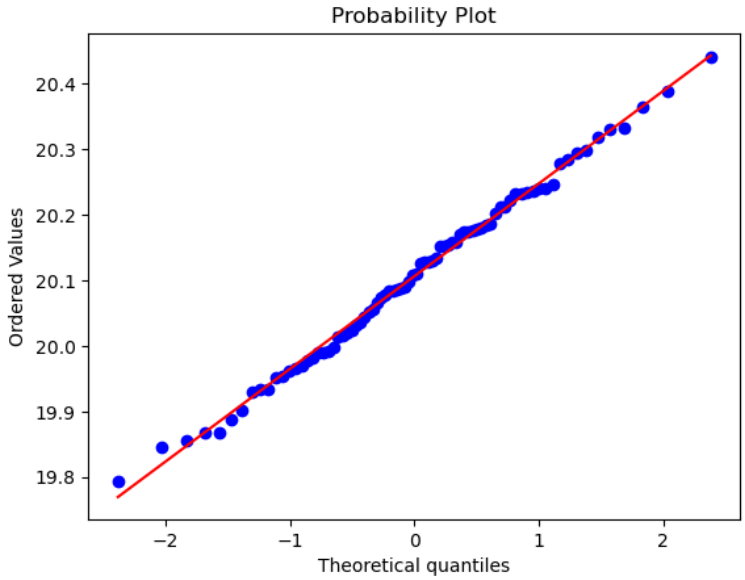
A trend and a shift in the mean seems to be present in the data. The change corresponds to the switch from shift 1 to shift 2.

Let's check the randomness.



The runs test returns a p-value of 0.000, and the sample ACF shows a linearly decaying trend, which is typical of non-stationary time series. Based on ACF analysis, we can state that the process is non-stationary.

Let's check the normality.



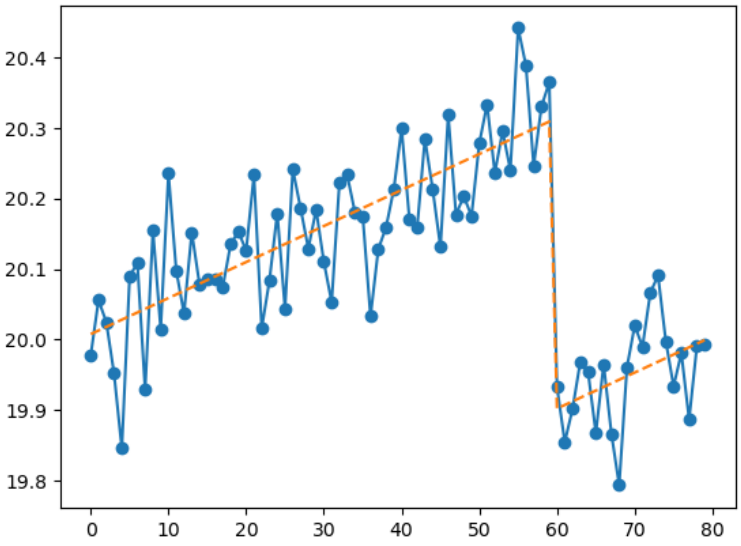
The Shapiro-Wilk test returns a p-value of 0.947, we cannot reject normality hypothesis.

Let's use the index and the shift as regressors and try creating a model:

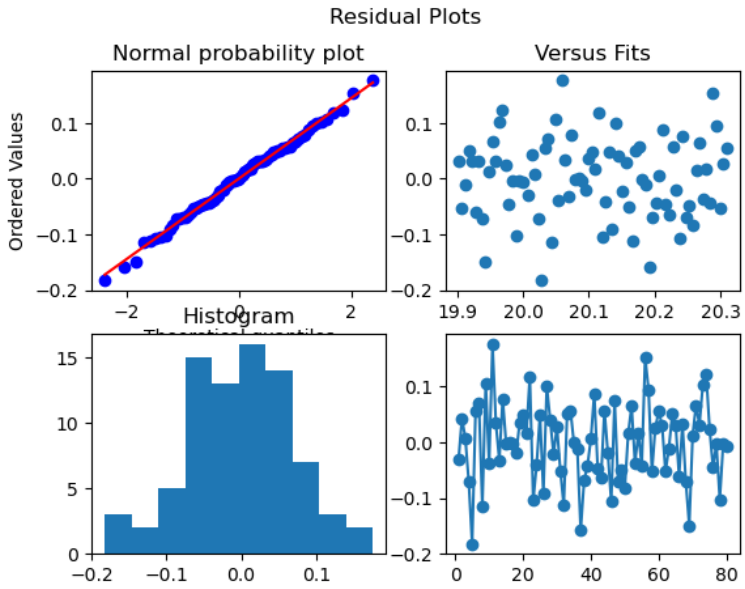
```
REGRESSION EQUATION
-------------------
diam =  + 20.414 const + 0.005 idx -0.411 shift

COEFFICIENTS
------------
 Term    Coef  SE Coef  T-Value     P-Value
const 20.4140   0.0251 811.8836 3.5828e-153
  idx  0.0051   0.0005   9.6542  6.6269e-15
shift -0.4112   0.0282 -14.6030  6.8483e-24

MODEL SUMMARY
-------------
     S   R-sq  R-sq(adj)
0.0721 0.7382     0.7314

ANALYSIS OF VARIANCE
--------------------
    Source   DF    Adj SS    Adj MS     F-Value     P-Value
Regression  2.0    1.1297    0.5648    108.5523  3.9136e-23
     const  1.0 3429.8021 3429.8021 659155.0434 3.5828e-153
       idx  1.0    0.4850    0.4850     93.2026  6.6269e-15
     shift  1.0    1.1096    1.1096    213.2472  6.8483e-24
     Error 77.0    0.4007    0.0052         NaN         NaN
     Total 79.0    1.5303       NaN         NaN         NaN
```

The runs test p-value on the residuals is 0.653 (OK) and the normality test p-value on the residuals is 0.993 (OK), so both normality and randomness on the residuals are verified.

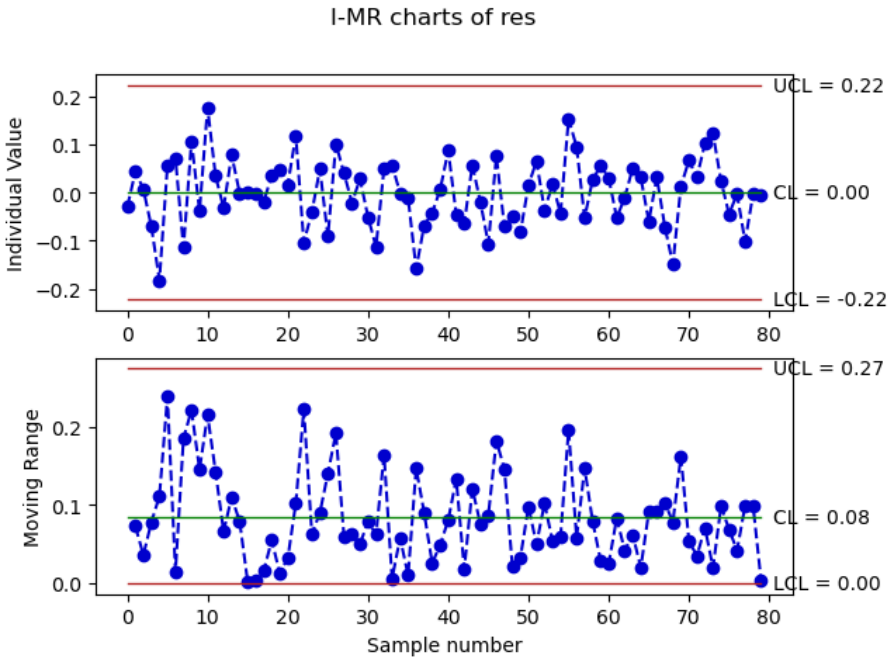
The fitted values seem to match closely the datapoints (R2_adj = 0.7314).



No particular trends are found in the residuals.



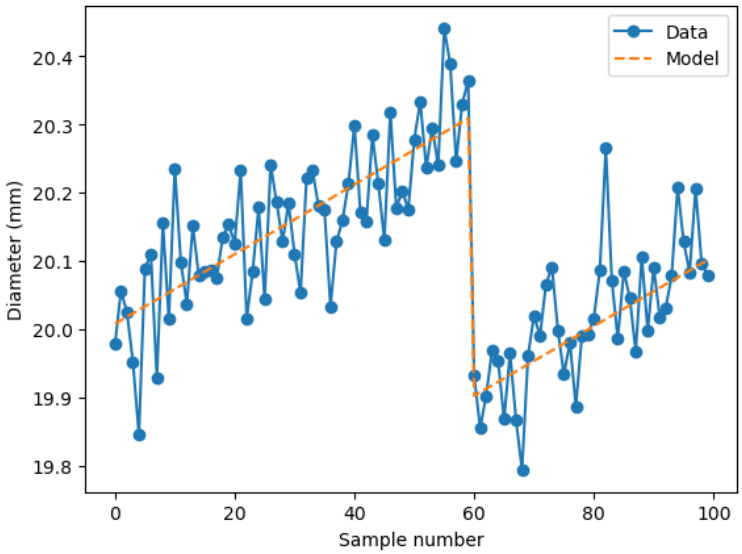
2) We can design a special cause control chart, using the residuals computed from the regression model to design an IMR control chart. K is set to 3.



No OOC in the control chart. The design phase (phase 1) is concluded.

3)

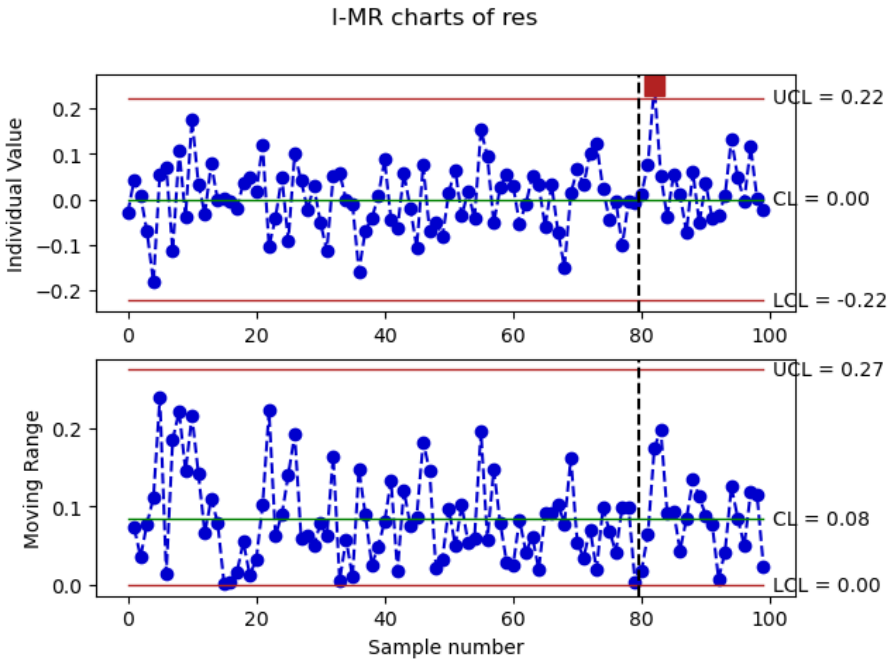
Let's use the model fitted in phase 1 to model the new data.



The model fitted in phase 1 seems to be a good fit for the phase 2 datapoints as well.

The residuals are still normal (SW test p-value = 0.730) and random (runs test p-value = 0.685).

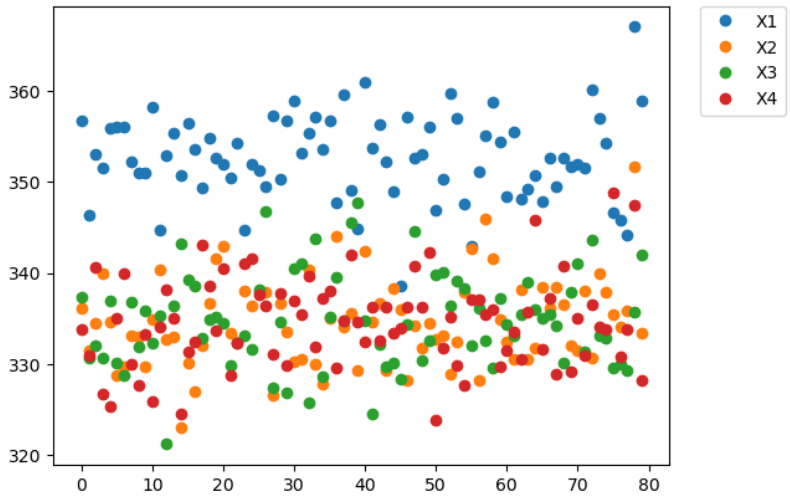
Let's apply the CC. One out-of-control is present in the phase 2 dataset (idx = 83).


I-MR charts of res
Individual Value
UCL = 0.22
CL = 0.00
LCL = -0.22
Moving Range
UCL = 0.27
CL = 0.08
LCL = 0.00
Sample number

Exercise 2 solution

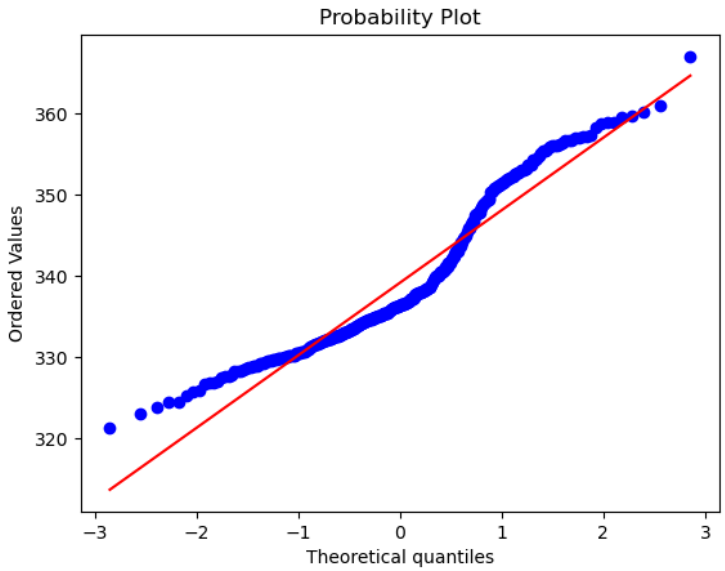
1)

Let's have a look at the data.



Hardness in location 1 is systematically higher than hardness in other locations.

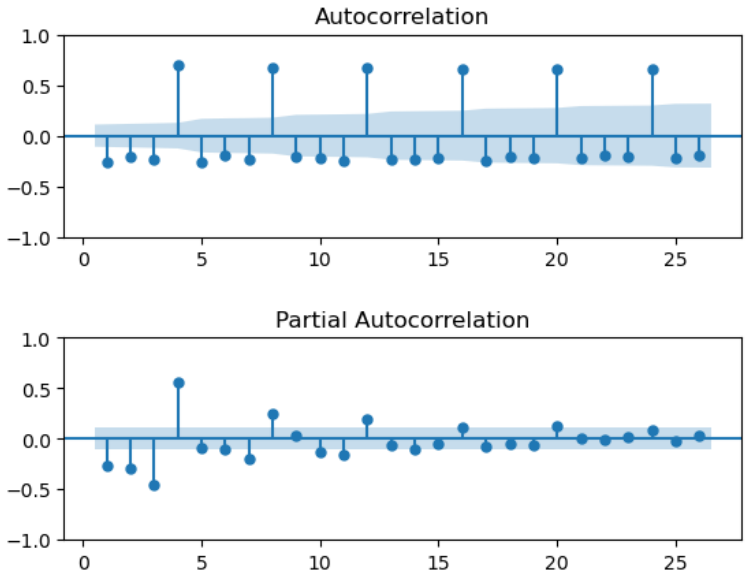
This affects the normality as shown below.



p-value of the Shapiro-Wilk test: 0.000

Since we know the time order of the data, we may also stack the data and check their autocorrelation pattern over time.

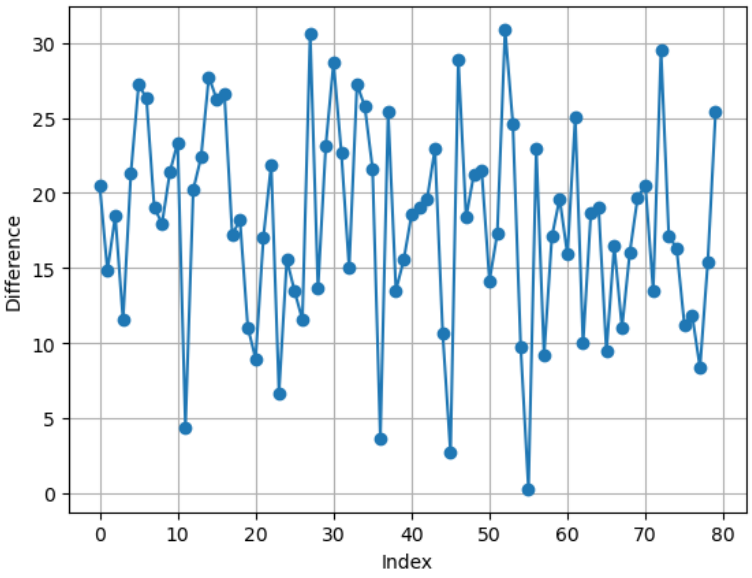
Runs test p-value = 0.000



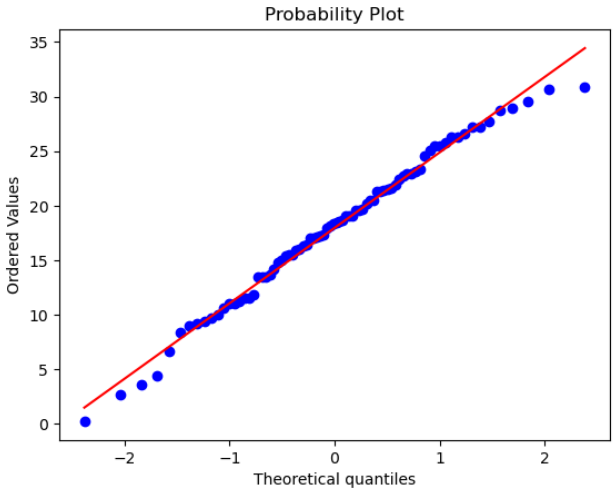
The lag 4 autocorrelation reflects the systematic pattern affecting location 1.

2)

We can perform a paired t-test.

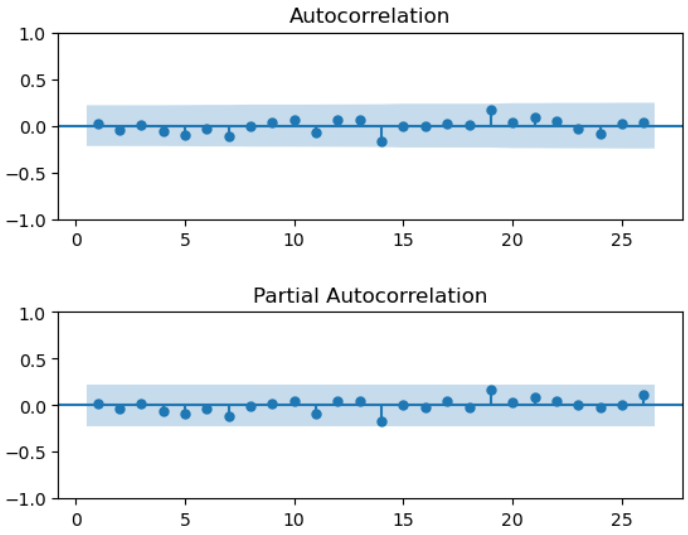


Let's check the normality of the difference with the Shapiro-Wilk test and the independence with the runs test and the ACF/PACF functions.



p-value of the Shapiro-Wilk test: 0.565

Runs test p-value = 0.982



Now that we know that the data are normally distributed, we can use the t-test to evaluate the following hypothesis:

$$H_0: \mu_d = 0$$

$$H_1: \mu_d > 0$$

The t-test statistic is:

$$t_0 = \frac{\bar{d}}{s_d/\sqrt{n}}$$

Where $\bar{d}$ is the sample mean of the difference, s_d is the sample standard deviation of the difference and n is the number of observations.

We get:

t-statistic: 23.554

p-value: 0.000

There is statistical evidence that the hardness in location 1 is statistically greater than the hardness in location 2.

3)

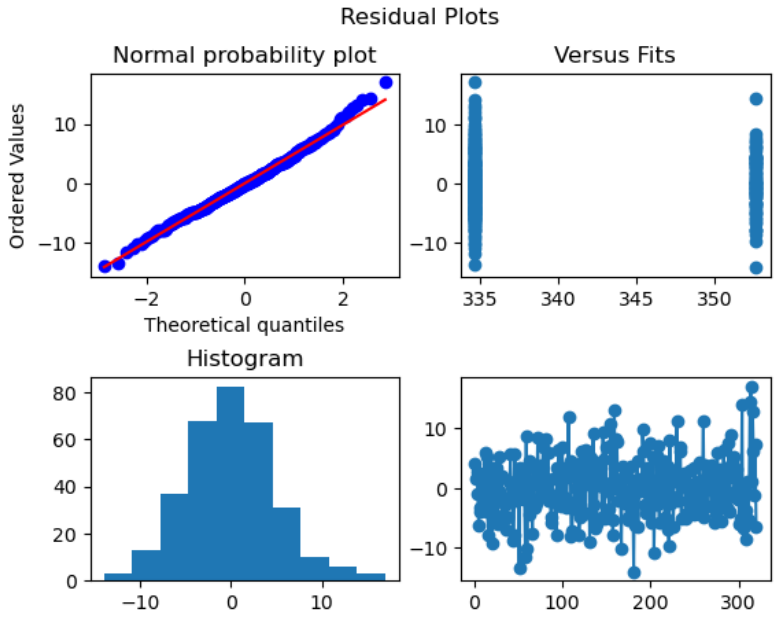
Due to the violation of the normality assumption caused by the statistically different hardness in location one, the most suitable approach would consist of fitting a model with a dummy regressor, and design a control chart for the residuals of the model. The dummy variable can be defined such that its value is 1 for location 1 and 0 for all other locations.

We get:

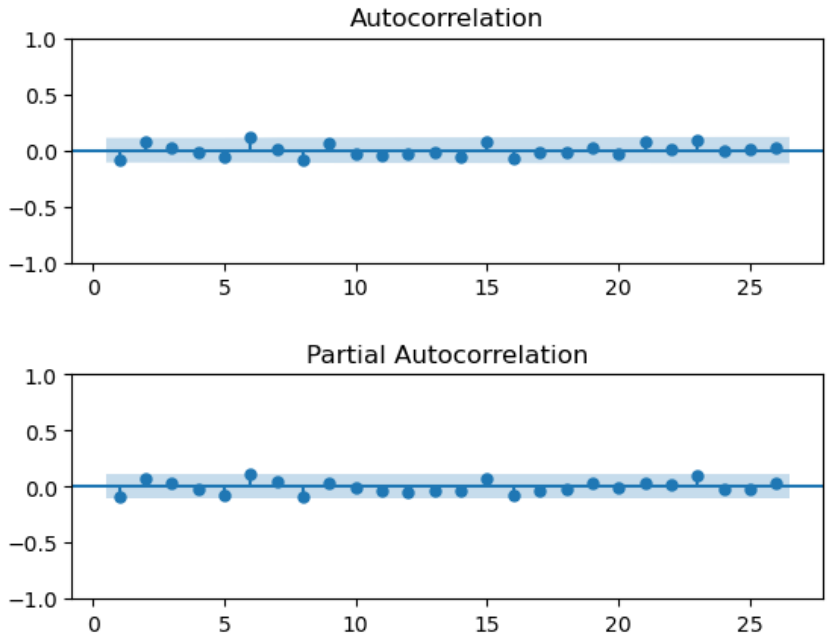
```
REGRESSION EQUATION
-------------------
data =  + 334.650 const + 17.979 loc

COEFFICIENTS
------------
 Term     Coef  SE Coef   T-Value    P-Value
const 334.6495   0.3183 1051.4040 0.0000e+00
  loc  17.9787   0.6366   28.2428 1.1257e-88

MODEL SUMMARY
-------------
     S  R-sq  R-sq(adj)
4.9309 0.715     0.7141

ANALYSIS OF VARIANCE
--------------------
    Source    DF     Adj SS     Adj MS    F-Value    P-Value
Regression   1.0 1.9394e+04 1.9394e+04 7.9765e+02 1.1257e-88
     const   1.0 2.6878e+07 2.6878e+07 1.1055e+06 0.0000e+00
       loc   1.0 1.9394e+04 1.9394e+04 7.9765e+02 1.1257e-88
     Error 318.0 7.7318e+03 2.4314e+01        NaN        NaN
     Total 319.0 2.7126e+04        NaN        NaN        NaN
```

Let’s check assumptions by analysing model residuals:

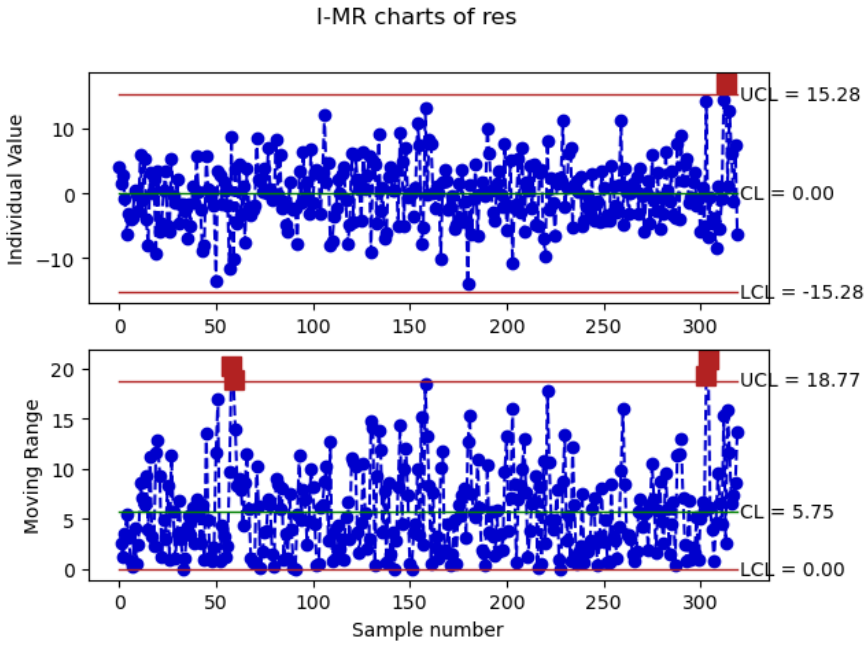
Shapiro-Wilk test p-value on the residuals = 0.198



Runs test p-value on the residuals = 0.140



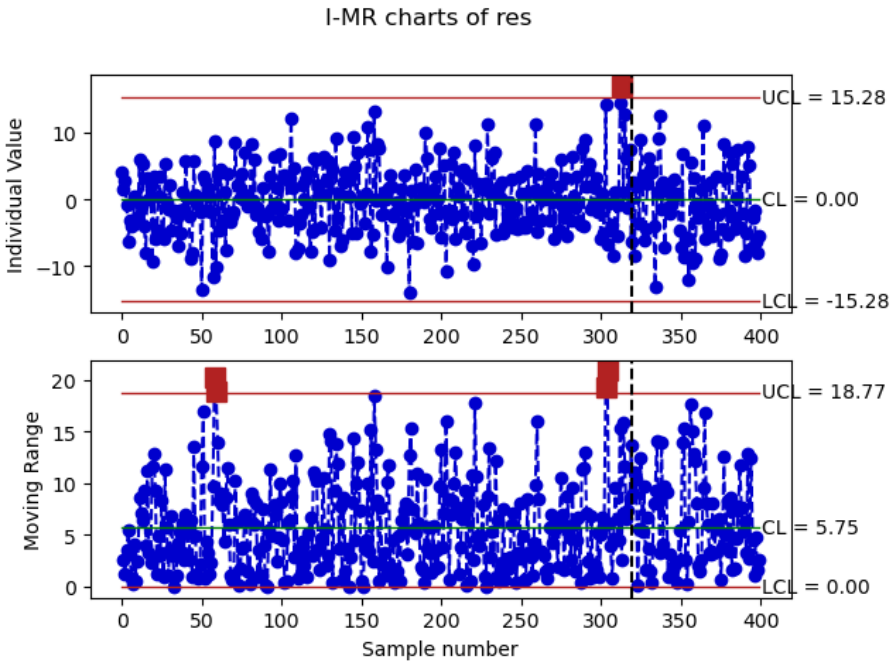
Let’s design the I-MR control chart on the residuals.



There is one violation of control limits in the I chart and two violations in the MR chart. Assuming no assignable cause for them, the control chart design is over.

4)

Phase 2 control chart. Let's fit the same model to the phase 2 data and apply the previously designed I-MR chart on model residuals of new data.



The new data are in control.

Exercise 3) Solution

Question 1)

Answer: b
Explanation: The area outside of the control limits is the type I error, i.e. describes the false alarm rate, which for the choice of K will correspond to the value α. Increasing K to K_1 will decrease the type I error to $\alpha_1 < \alpha$ and as a result the false alarm rate will decrease, so (b) is valid and therefore (a) and (d) will be invalid.
We also know that as type I error, α, decreases, then the type II error, β, will increase and given that power $= 1 - \beta$, we will have that power will decrease, so (c) is not valid as well.

Question 2)

Answer: c

Explanation: The ANOVA table for the linear model performs the following hypothesis testing: $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ versus $H_1: not\ H_0$.

As the p-value < alpha we can reject the null hypothesis H_0 and thus (b) is invalid. The alternative hypothesis H_1 is the complement of H_0, i.e. at least one of the model coefficients is not equal to zero, thus (c) is valid. Both (a) and (d) are invalid as none of them is expressing the alternative hypothesis.