QUALITY DATA ANALYSIS

19/07/2023

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 d) and Exercise 3, Question 2.**

Exercise 1 (14 points)

The oxygen concentration inside the build chamber of a metal 3D printer is measured at the end of each layer. The data of the first 100 layers is reported in the file `*oxygen_phase1.csv*`.

a) Find a suitable model for the data.
b) Predict the oxygen concentration at the end of the next layer.
c) Design a suitable control chart to monitor the oxygen concentration with a Type I error $\alpha = 0.0029$. Please show on the solution the formula of the control limits and their numerical values. *Note: in case of violations of the control limits, assume the existence of assignable causes.*
d) Using the control chart designed in point 3 (phase 1), test it on the data of the next 10 (stored in `*oxygen_phase2.csv*`). Please report all the values plotted in the control chart in Phase 2 against limits. Report then the index of the out-of-control observations (if any).

Exercise 2 (15 points)

In a process for the production of an energy drink, three quality variables are monitored, denoted by x_1, x_2 and x_3. Under in-control conditions they are known to be iid and to follow a multivariate normal distribution with $\mu_1 = 25.4$, $\mu_2 = 23.7$, $\mu_3 = 28.5$, and $\sigma_1 = 1.483$, $\sigma_2 = 1.435$, $\sigma_3 = 1.530$. It is also known that the correlation between each pair of quality variables is $\rho_{12} = 0.799$, $\rho_{13} = 0.705$, and $\rho_{23} = 0.683$.

a) Design three univariate I control charts for the three distinct quality characteristics such that the familywise type I error is at most $\alpha = 0.01$. In the use phase of the control chart, the dataset included in *exe_pca.csv* is collected. Using the designed chart, determine if these new measurements are in control or not.
b) Estimate and draw the operating characteristic curve of the control chart for x_1 designed in point a) in the presence of a shift $\Delta\mu_1$ of the mean of x_1 with $\Delta\mu_1 \in [0, 10]$, and report the value of the Type II error for $\Delta\mu_1 = 3$.
c) The head of the quality department decides to use the Principal Component Analysis to monitor these data. The first principal component (PC) allows capturing at least 65% of the data variability and it is such that:

$$\boldsymbol{u}_1 = [-0.589, -0.561, -0.582]^T, \lambda_1 = 5.402$$

Design a univariate control chart to monitor the first PC with a type I error $\alpha = 0.01$. Using this control chart, determine if the new observations in *exe_pca.csv* are in control or not.

d) Estimate and draw the operating characteristic curve of the control chart designed in point c) in the presence of a shift $\Delta\mu_1$ of the mean of variable x_1 with $\Delta\mu_1 \in [0, 10]$, and report the value of the Type II error for $\Delta\mu_1 = 3$. Compare this result with the one in point b) and discuss the difference.

Exercise 3 (4 points)

In the following questions select one of the four possible choices as your answer and provide a short justification of your choice. Answers **without** justification will **not** receive any credit.

Question 1 (2 points):
In the stepwise regression we need to define two levels of significance: Alpha-to-Enter and Alpha-to-Remove. Which of the following choices is **not** a valid selection?

a) Alpha-to-Enter ≤ Alpha-to-Remove.
b) Alpha-to-Enter = Alpha-to-Remove.
c) Alpha-to-Enter < Alpha-to-Remove.
d) Alpha-to-Enter > Alpha-to-Remove.

Question 2 (2 points):
In a fitting the linear regression model of the continuous variable **Y** on the five predictors {$\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$, $\mathbf{X_4}$ and $\mathbf{X_5}$} we obtained the following p-values:

Predictor	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$
p-value in the full model	0.015	0.241	0.001	0.087	0.047

If we will apply on the above model the backward elimination procedure using the value of Alpha-to-Remove = 0.05, which is the first predictor that will be removed?

a) X_1
b) X_2
c) X_4
d) We cannot tell from the above output only.

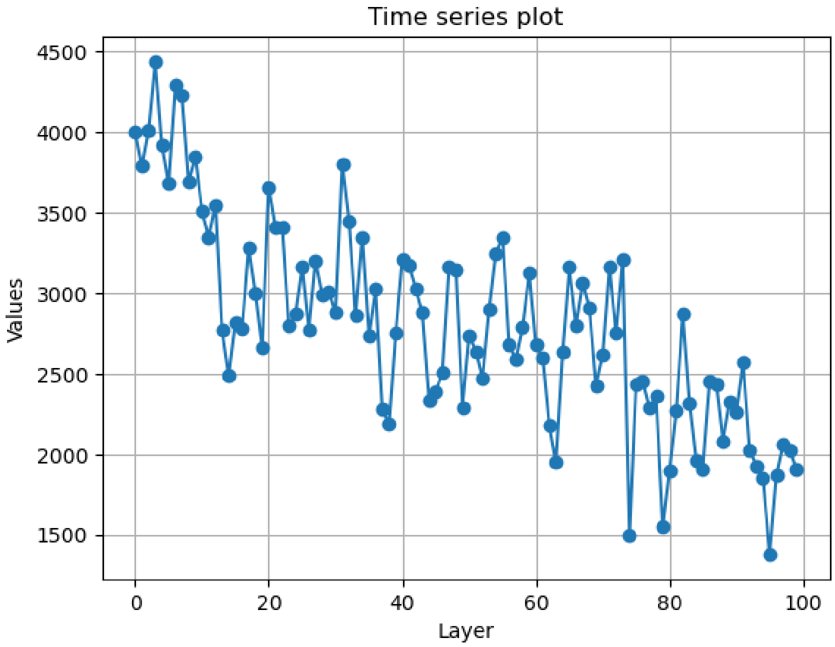
What is the predictor to be possibly eliminated on a second step of the backward elimination algorithm?

a) X_1
b) X_2
c) X_4
d) We cannot tell from the above output only.

Exercise 1 solution

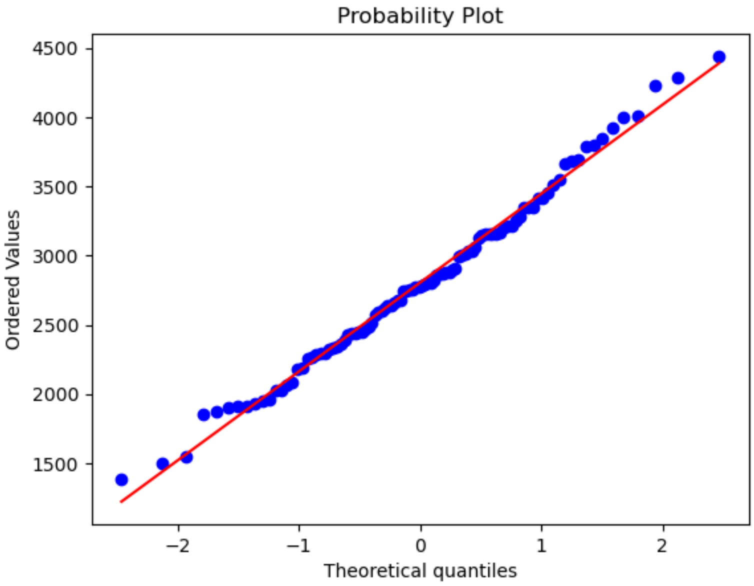
a)

Since the time order of the measurements is known, we can plot the time series as follows:



The pattern is clearly not random as it appears that the temperature is decreasing as the number of layers increases.

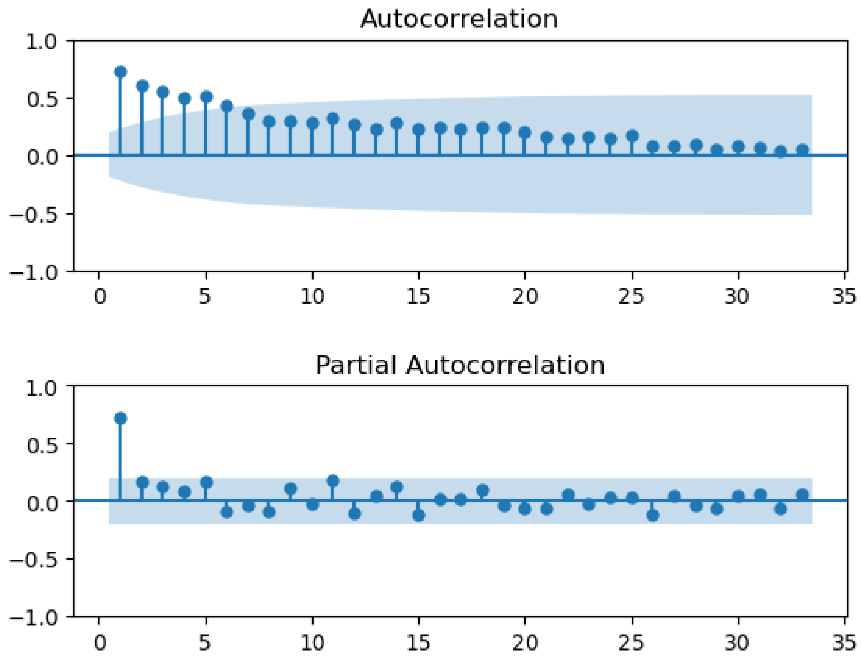
The normality assumption is met (Shapiro-Wilk’s test p-value = 0.663).



Since we know the time order of the data we can check the randomness and autocorrelation.

The randomness assumption is NOT met (Runs test p-value = 0.000).

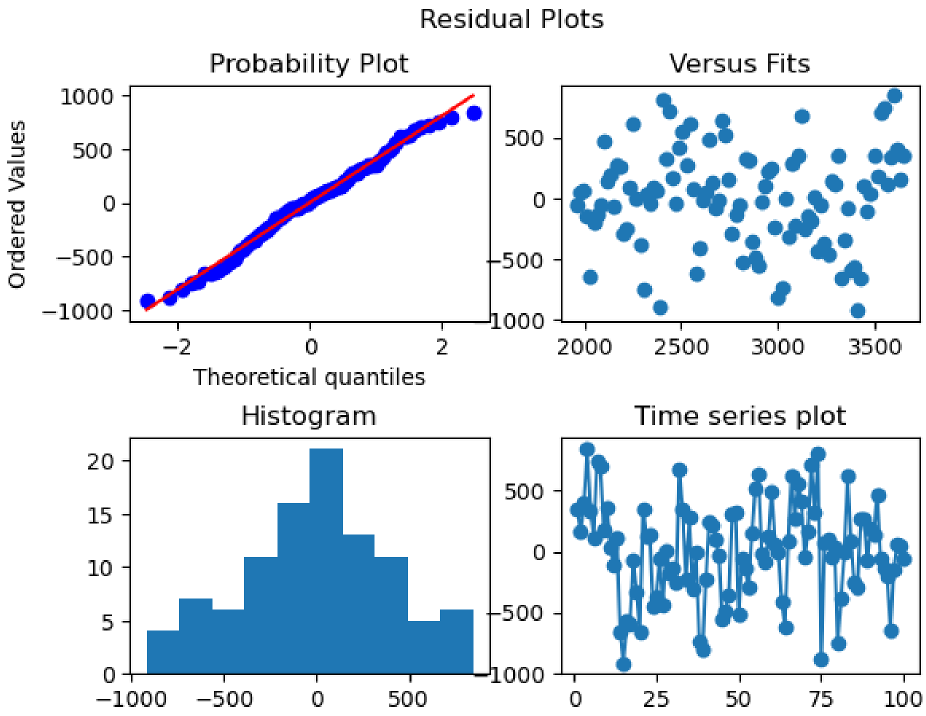
The sample ACF shows the typical pattern of a non-stationary time series (linear decay).



We can try to fit a trend model using the layer number as regressor.

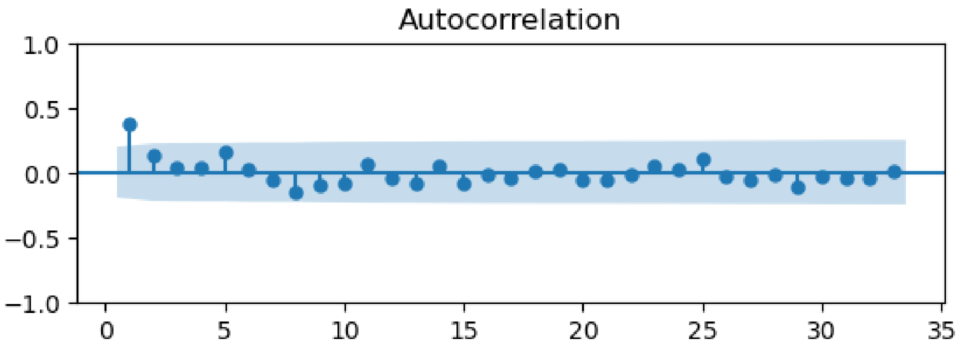
```
REGRESSION EQUATION
-------------------
val =  + 3663.885 const -16.998 layer

COEFFICIENTS
------------
 Term      Coef  SE Coef  T-Value    P-Value
const 3663.8848  81.2907  45.0714 2.5144e-67
layer  -16.9977   1.3975 -12.1628 2.7090e-21

MODEL SUMMARY
-------------
       S   R-sq  R-sq(adj)
403.4091 0.6015     0.5975

ANALYSIS OF VARIANCE
---------------------
    Source   DF     Adj SS     Adj MS   F-Value    P-Value
Regression  1.0 2.4074e+07 2.4074e+07  147.9331 2.7090e-21
     const  1.0 3.3059e+08 3.3059e+08 2031.4282 2.5144e-67
     layer  1.0 2.4074e+07 2.4074e+07  147.9331 2.7090e-21
     Error 98.0 1.5948e+07 1.6274e+05       NaN        NaN
     Total 99.0 4.0023e+07        NaN       NaN        NaN
```

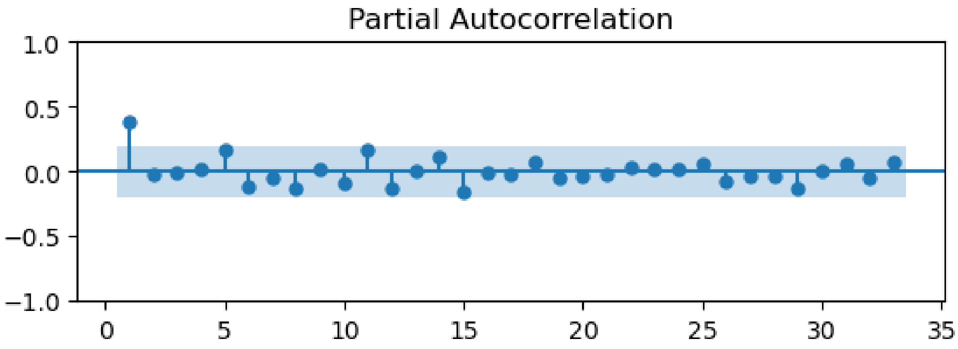
Let’s check the assumptions on the residuals.



The normality assumption is met (Shapiro-Wilk's test p-value = 0.449).

The randomness assumption is NOT met (Runs test p-value = 0.003).



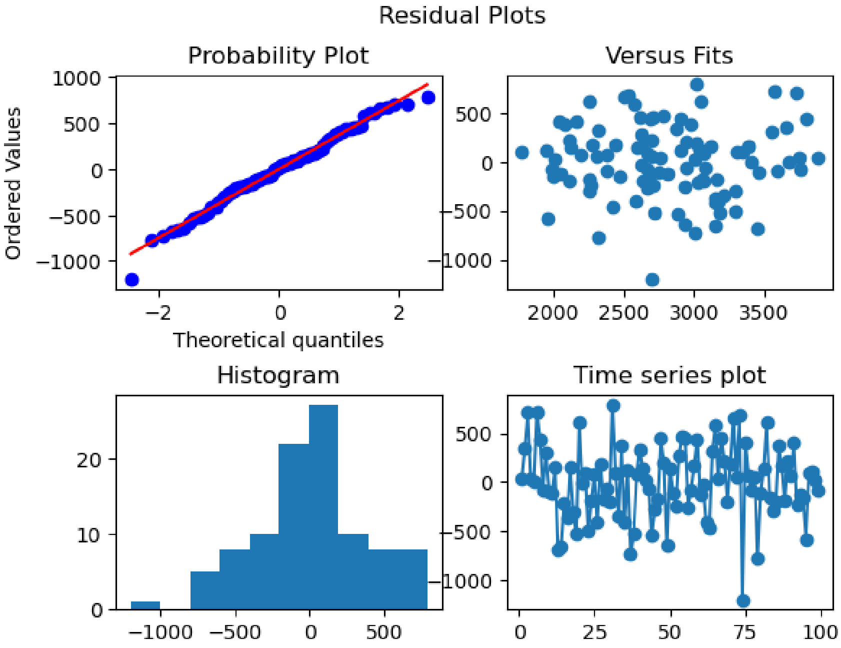


From the sample ACF and PACF it seems that the time series of the residuals is now stationary but still autocorrelated. The ACF seems to follow a "geometric" decay while only the first lag of the PACF is high. These patterns suggest to use an AR(1) model for the data.

Let's add an autoregressive term ("lag1") to the model.

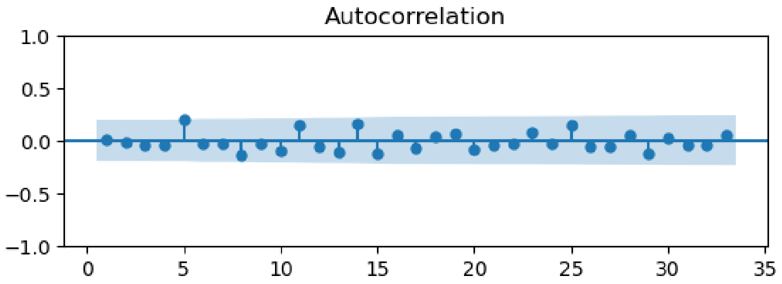
```
REGRESSION EQUATION
-------------------
val =  + 2255.631 const -10.355 layer + 0.379 lag1

COEFFICIENTS
------------
 Term      Coef  SE Coef  T-Value    P-Value
const 2255.6311 354.4765   6.3633 6.7234e-09
layer  -10.3548   2.0706  -5.0009 2.5719e-06
 lag1    0.3787   0.0940   4.0280 1.1254e-04

MODEL SUMMARY
-------------
     S   R-sq  R-sq(adj)
375.44 0.6493      0.642

ANALYSIS OF VARIANCE
---------------------
    Source   DF     Adj SS     Adj MS  F-Value    P-Value
Regression  2.0 2.5050e+07 1.2525e+07  88.8578 1.4409e-22
     const  1.0 5.7075e+06 5.7075e+06  40.4912 6.7234e-09
     layer  1.0 3.5251e+06 3.5251e+06  25.0086 2.5719e-06
      lag1  1.0 2.2869e+06 2.2869e+06  16.2244 1.1254e-04
     Error 96.0 1.3532e+07 1.4096e+05      NaN        NaN
     Total 98.0 3.8582e+07        NaN      NaN        NaN
```

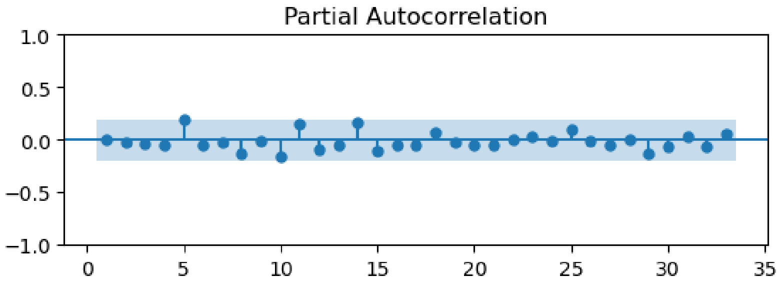
Let’s check the assumptions on the residuals.



The normality assumption is met (Shapiro-Wilk’s test p-value = 0.514).

The randomness assumption is met (Runs test p-value = 0.631).





All the terms are significant, including the new autoregressive term. The R-sq(adj) has also improved from the first model.

All assumptions are met. The model is adequate.

b)

The oxygen concentration at the end of the next layer can be computed from the fitted regression equation:

```
val =  + 2255.631 const -10.355 layer + 0.379 lag1

const = 1
layer = 101
lag1 = 1910

val = 1933.1333
```

c)

The special cause control chart is suitable for this situation. We can design an I-MR control chart on the residuals of the last model using the following formulas to compute the control limits.

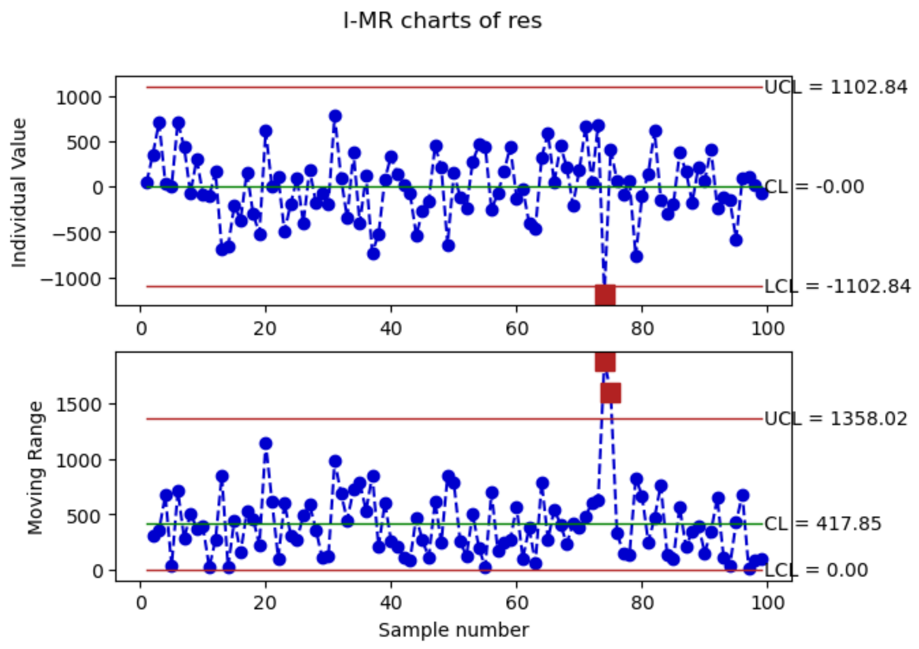
I chart:

- $UCL = \bar{x} + 3\left(\frac{\bar{MR}}{d_2}\right)$
- $CL = \bar{x}$
- $LCL = \bar{x} - 3\left(\frac{\bar{MR}}{d_2}\right)$

MR chart:

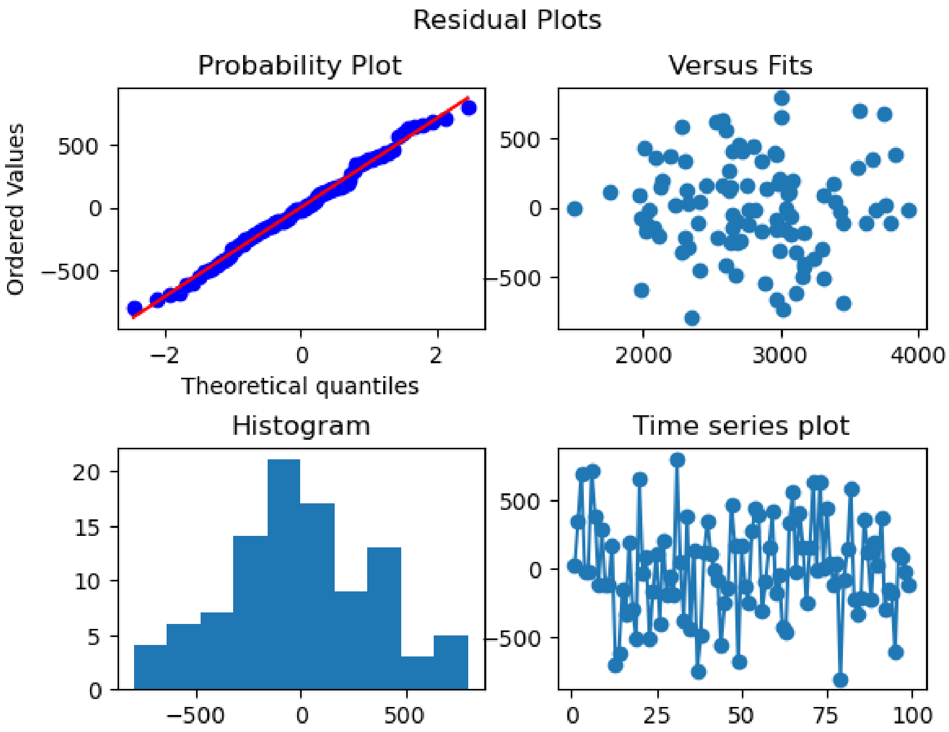
- $UCL = D_4\bar{MR}$
- $CL = \bar{MR}$
- $LCL = 0$

The type I error α must be equal to 0.0029, therefore the K we need to use to compute the values of the factors d2 and D4 is equal to 2.978.



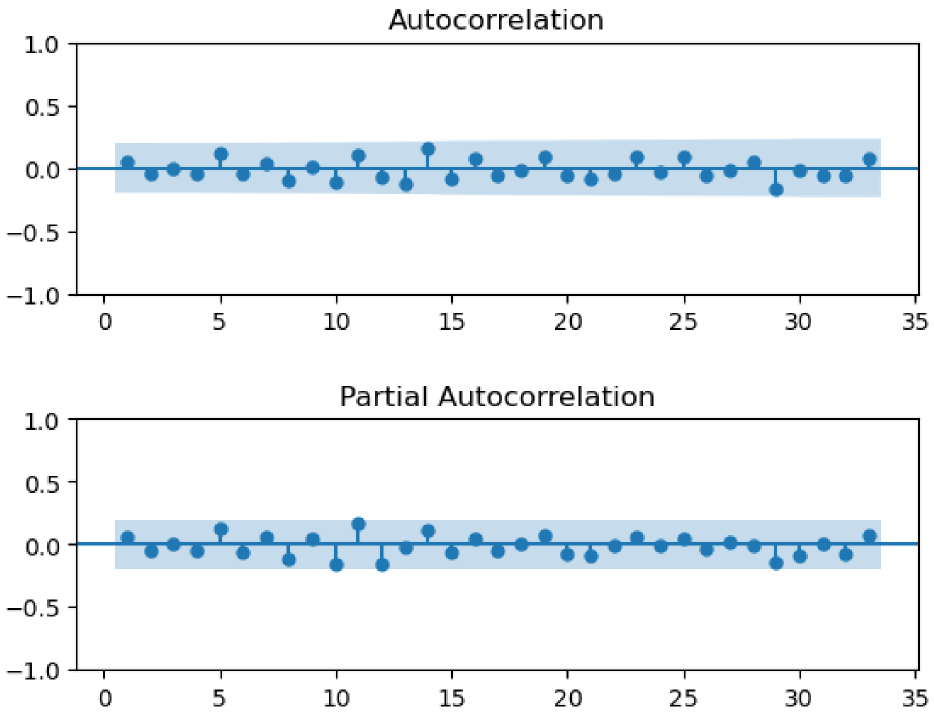
At index 74 there is a violation of the control limits in the I cc. This violation seems to affect the MR cc as well. Since we are assuming that an assignable cause was found, we need to remove the OOC points from the data and re-estimate the control limits of the CC. To do so, we need to re-fit the model using a dummy variable as regressor (= 0 for all layers except for the OOC layer).

```
REGRESSION EQUATION
-------------------
val =  + 2014.440 const -8.896 layer + 0.443 lag1 -1267.733 dummy

COEFFICIENTS
------------
 Term       Coef  SE Coef  T-Value    P-Value
const  2014.4399 342.9623   5.8736 6.2616e-08
layer    -8.8960   2.0063  -4.4340 2.4842e-05
 lag1     0.4425   0.0910   4.8652 4.5320e-06
dummy -1267.7329 366.3758  -3.4602 8.1015e-04

MODEL SUMMARY
-------------
       S   R-sq  R-sq(adj)
355.6633 0.6885     0.6787

ANALYSIS OF VARIANCE
---------------------
    Source   DF     Adj SS     Adj MS  F-Value    P-Value
Regression  3.0 2.6564e+07 8.8548e+06  70.0006 5.6583e-24
     const  1.0 4.3641e+06 4.3641e+06  34.4997 6.2616e-08
     layer  1.0 2.4869e+06 2.4869e+06  19.6600 2.4842e-05
      lag1  1.0 2.9941e+06 2.9941e+06  23.6698 4.5320e-06
     dummy  1.0 1.5145e+06 1.5145e+06  11.9730 8.1015e-04
     Error 95.0 1.2017e+07 1.2650e+05      NaN        NaN
     Total 98.0 3.8582e+07        NaN      NaN        NaN
```



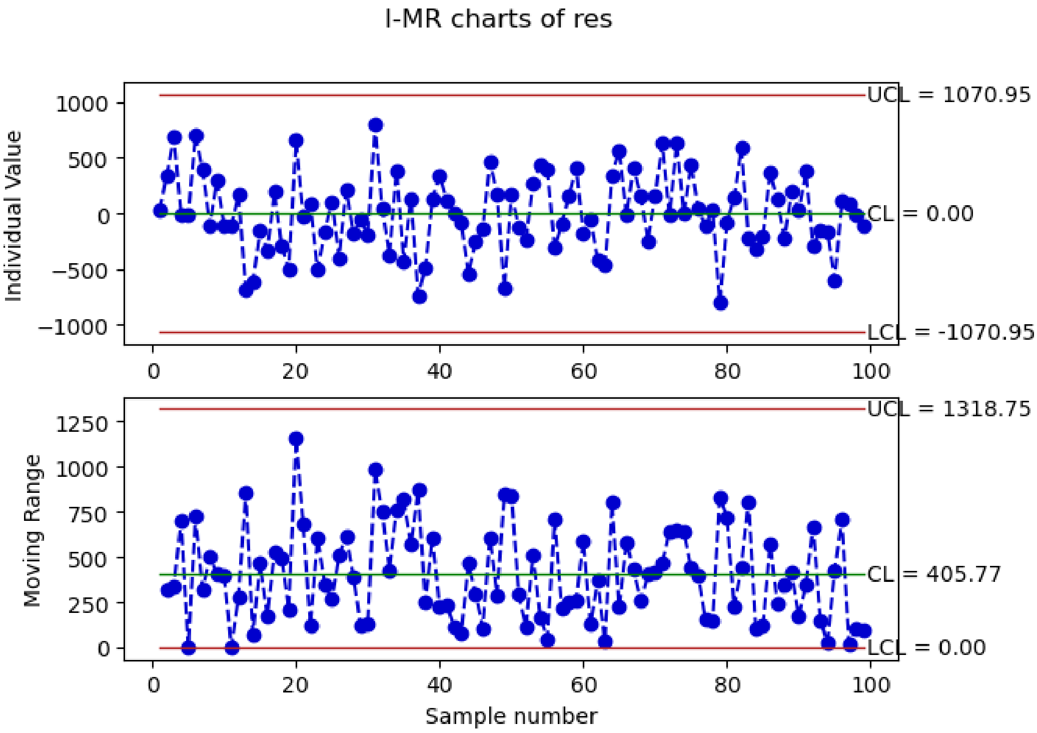
The normality assumption is met (Shapiro-Wilk's test p-value = 0.729).

The randomness assumption is met (Runs test p-value = 0.742).



The model is adequate.

Now design the I-MR cc using the new residuals.



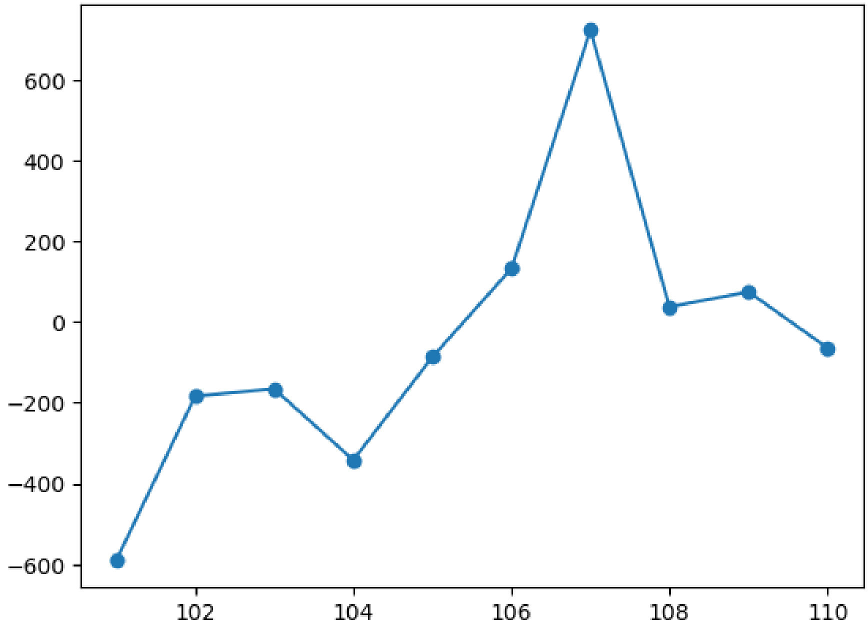
No out-of-control points are detected. Phase 1 is completed.

d)

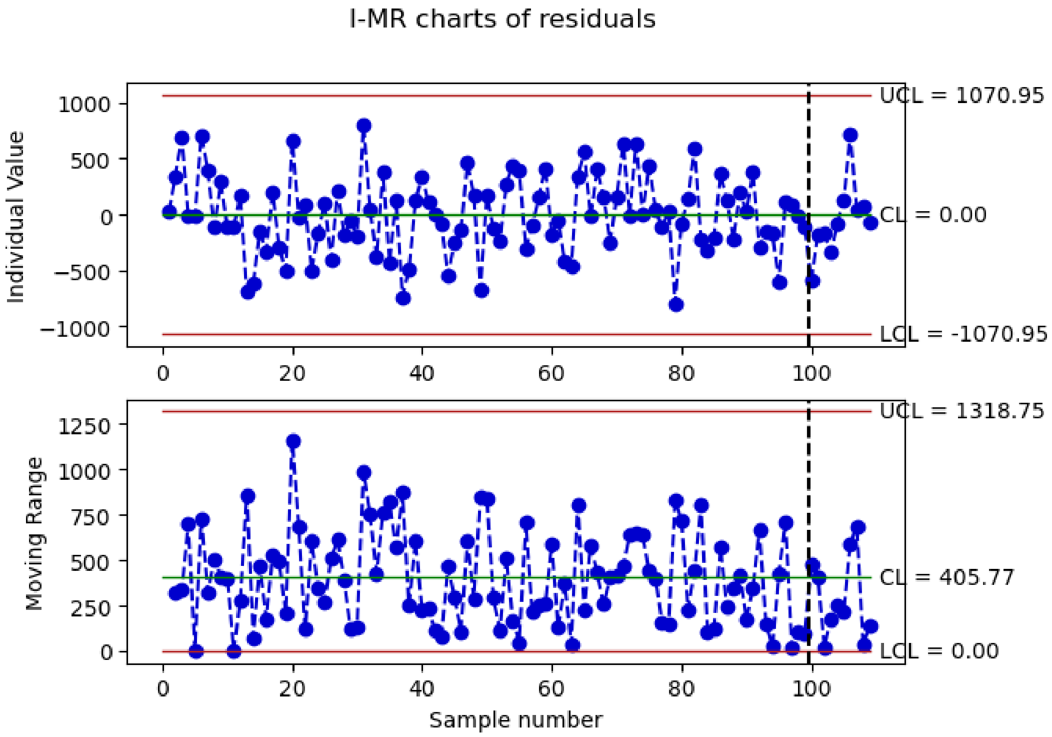
To determine whether the new data are in-control or not, the last model fitted in point 2 shall be used.

```
val =  + 0.201 const -0.001 layer + 0.443 lag1 -0.127 dummy
```

The resulting residuals for the layers from 101 to 110 are:



By plotting the new residuals on the previously design special cause control chart, we get:



The new observations are in control.

Exercise 2 solution

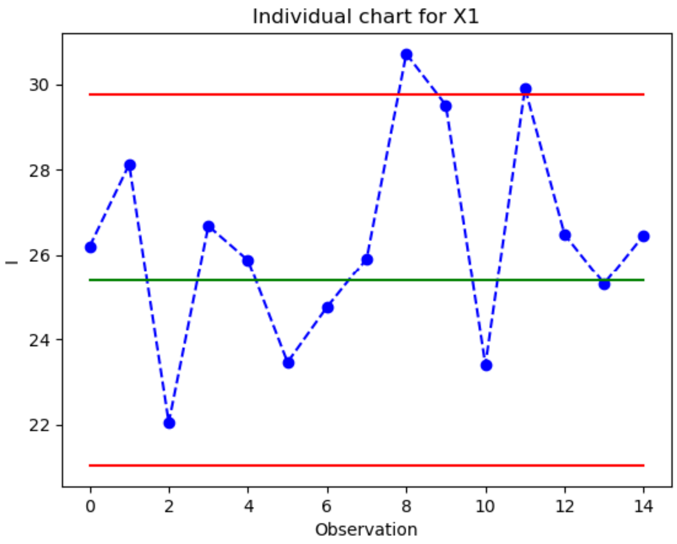
a)

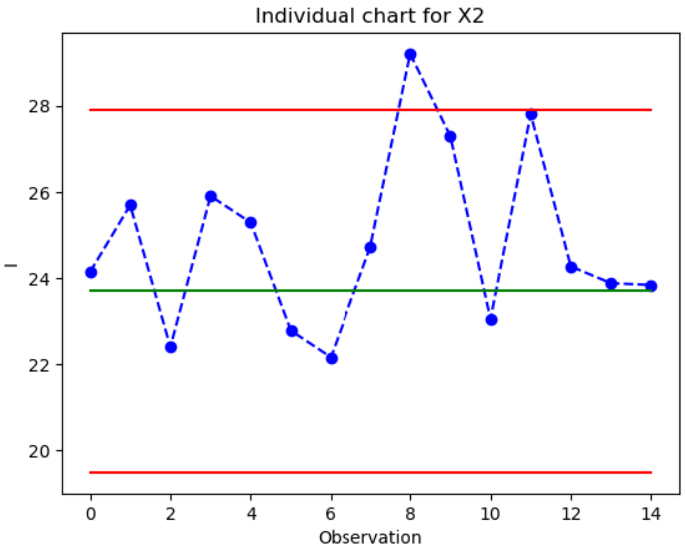
Using the Bonferroni's correction, each individual control chart is designed with a Type I error $\alpha' = \alpha/3$, where $\alpha = 0.01$, and hence K = 2.935.

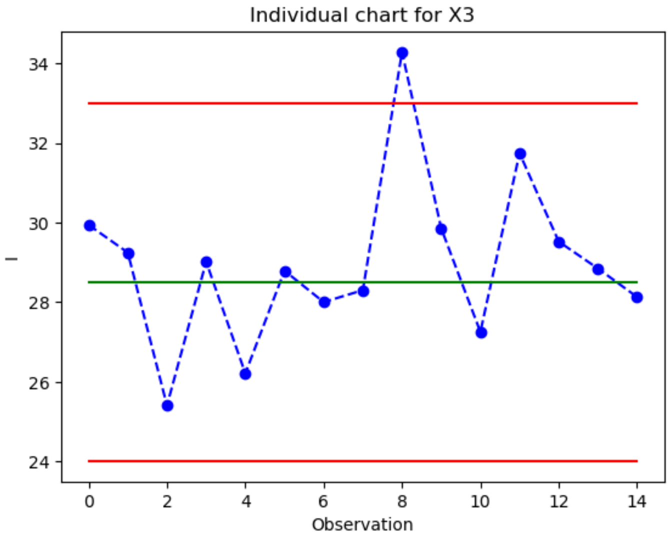
Being known the process parameters, the control limits of the three individual control charts are:

x_1	$UCL = \mu_1 + K\sigma_1 = 29.75$ $LCL = \mu_1 - K\sigma_1 = 21.05$
x_2	$UCL = \mu_2 + K\sigma_2 = 27.91$ $LCL = \mu_2 - K\sigma_2 = 19.48$
x_3	$UCL = \mu_3 + K\sigma_3 = 32.99$ $LCL = \mu_3 - K\sigma_3 = 24.01$

Phase 2 control charts are the following:







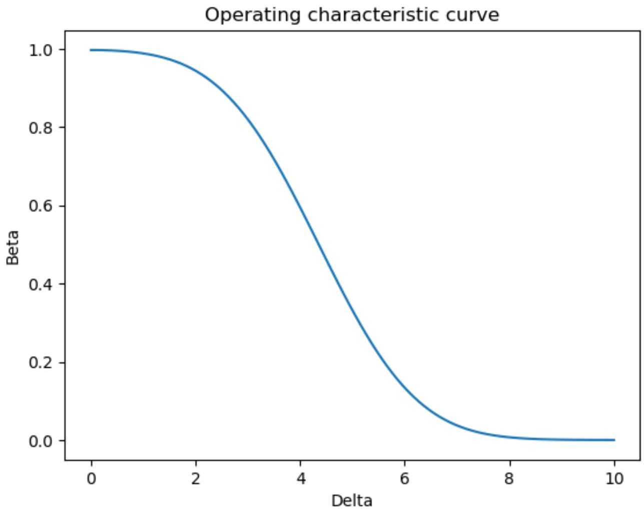
Observation n. 8 is out of control in all the three control charts. The control chart for x1 also signals an out of control for observation n. 11.

b)

The Type II error β_1 for the control chart con x_1 is:

$$\beta_1 = \Phi\left(K - \frac{\Delta\mu_1}{\sigma_1}\right) - \Phi\left(-K - \frac{\Delta\mu_1}{\sigma_1}\right)$$

The operating characteristic curve is:



The Type II error for $\Delta\mu_1 = 3$ is $\beta = 0.814$.

c)

Under in-control conditions, the scores of the first principal component are:

$$\mathbf{z}_1 = [z_{11,}z_{21,} \dots, z_{n1}]^T = (\mathbf{X} - \boldsymbol{\mu})\mathbf{u}_1, \text{ such that } \mathbf{z}_1 \sim N(\mu_{PC1}, \sigma_{PC1}^2)$$

Where:

$$\mu_{PC1} = 0$$

$$\sigma_{PC1}^2 = \lambda_1 = 5.402$$

With $\alpha = 0.01$ we have K = 2.576

Thus, the control chart to monitor the first principal component is:

$$UCL = \mu_{PC1} + K\sigma_{PC1} = 5.987$$

$$CL = \mu_{PC1} = 0$$

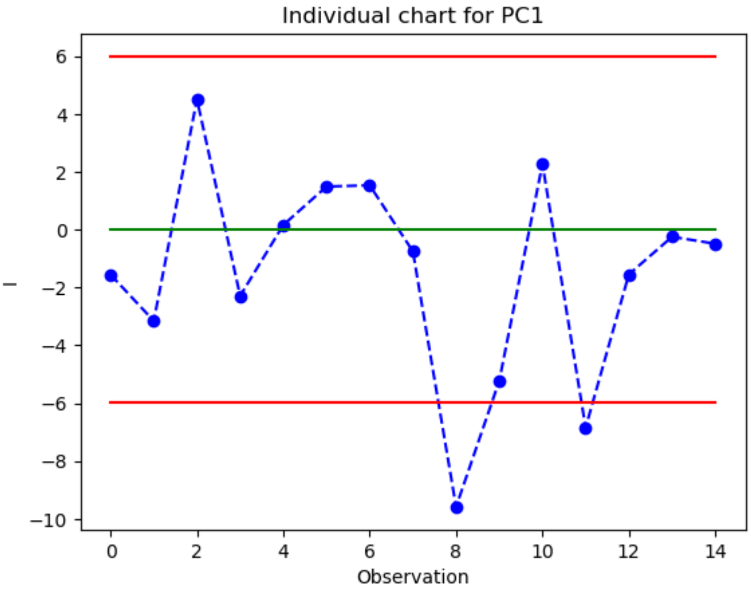
$$LCL = \mu_{PC1} - K\sigma_{PC1} = -5.987$$

To apply this control chart to the new data, they shall be projected along the direction spanned by the first principal component.

Being $\boldsymbol{u}_1 = [-0.589, -0.561, -0.582]^T$, the new data projections are:

$$\boldsymbol{z}_1 = -0.589(\boldsymbol{x}_1 - \mu_1) - 0.561(\boldsymbol{x}_2 - \mu_2) - 0.582(\boldsymbol{x}_3 - \mu_3)$$

The individual control chart on the first principal component is:



The control chart confirms the results obtained in point a), i.e., observations n. 8 and 11 are signaled as out of control.

d)

In out-of-control conditions we have $\boldsymbol{\mu}^* = [\mu_1 + \Delta\mu_1, \mu_2, \mu_3]^T$, and hence the out-of-control mean of the principal component is:

$$\mu^*{}_{PC1} = \mathrm{E}[u_{11}(X - \mu_1) + u_{12}(X - \mu_2) + u_{13}(X - \mu_3)] = u_{11}\Delta\mu_1$$

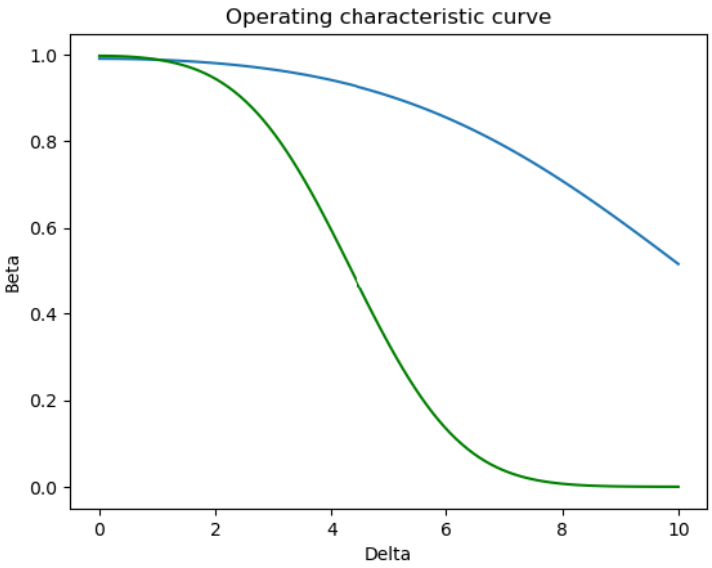
The Type II error can be computed as follows:

$$\beta_1 = \Phi\left(\frac{UCL - u_{11}\Delta\mu_1}{\sqrt{\lambda_1}}\right) - \Phi\left(\frac{LCL - u_{11}\Delta\mu_1}{\sqrt{\lambda_1}}\right) =$$

$$= \Phi\left(\frac{K\sqrt{\lambda} - u_{11}\Delta\mu_1}{\sqrt{\lambda_1}}\right) - \Phi\left(\frac{-K\sqrt{\lambda} - u_{11}\Delta\mu_1}{\sqrt{\lambda_1}}\right) =$$

$$= \Phi\left(K - \frac{u_{11}\Delta\mu_1}{\sqrt{\lambda_1}}\right) - \Phi\left(-K - \frac{u_{11}\Delta\mu_1}{\sqrt{\lambda_1}}\right)$$

The resulting operating characteristic curve is the one in blue in the next figure, where the green curve is the operating characteristic curve computed in point b):



The Type II error for $\Delta\mu_1 = 3$ is $\beta_1 = 0.964$

In case the shift affects the mean of one single variable, the control chart on the first PC is less effective than applying univariate control charts on the original variables. The reason is that the first PC is a basically a mean of the three variables (remind: $\boldsymbol{u}_1 = [-0.589, -0.561, -0.582]^T$), and hence the deviation on one signal variable is mitigated by the weight in the linear combination.

Exercise 3 solution

1)

Answer: d

Explanation: In the case where Alpha-to-Enter > Alpha-to-Remove there is a possibility the algorithm to fall in an infinite loop. For example, assume that we fix Alpha-to-Enter = 0.05 and Alpha-to-Remove = 0.01 and at some iteration of the stepwise approach in the forward step we have the smallest p-value = 0.03 for a variable X_k. Then as p-value(X_k)=0.03 < 0.05=Alpha-to-Enter, the algorithm will include X_k in the model. At the next step (backward elimination) though, we will have p-value(X_k)=0.03 > 0.01=Alpha-to-Remove and thus the variable will be moved out of the model and so we will get into an infinite loop.

2)

i): Answer: b

Explanation: At the first step of the backward elimination, we will remove X_2, as it has the highest p-value(X_2)=0.241, which exceeds the threshold Alpha-to-Remove=0.05.

ii): Answer: d

After the first backward elimination, we need to rerun the regression model with the remaining four predictors {X_1, X_3, X_4 and X_5} and all the p-values will be updated, and it is not possible to know in advance which will be the larger and whether this will still exceed the Alpha-to-Remove=0.05 threshold. Thus, we cannot tell.