Statistics for Data Science Project

Micluța-Câmpeanu Marius – 407

1. Time series

Dataset used

The dataset for this subject is the Air Quality Index (AQI) from Bucharest, recorded at University Square. Official data can be obtained from http://calitateaer.ro/public/monitoring-page/reports-reports-page/, but the process itself is tedious and time-consuming. Instead, I used another source which is more straightforward to use. The data can be viewed here: https://aerlive.ro/ica/universitate/ and it can be easily downloaded using the browser's developer tools and selecting the corresponding json response. Using cURL, it can be accessed using the following command in order to obtain more datapoints:

```
curl -k 'https://apps.roiot.ro/aerlive/api/cluster.php?q=history' \
--data-raw 'key=d09668ea-def5-44ea-8c77-ae32e9fa5572&s=2019-01-25&e=2020-06-22&cluster=8'\
> data/data_univ.json
```

Inspecting the two request parameters (`s` and `e`), we can determine that we sould have data spanning from 2019-01-25 until 2020-06-22. However, analyzing the data, we have 291 observations spanning from 2019-08-13 until 2020-06-03. The `key` parameter refers to the station (Universitate in this case).

The data has the following structure:

json	list [2]	List of length 2
status	integer [1]	100
data	list [6]	List of length 6
ica_data	list [3]	List of length 3
target	character [1]	'ica'
columns	list [291]	List of length 291
series	list [291]	List of length 291
[[1]]	list [2]	List of length 2
y	double [1]	20.74091
color	character [1]	'#008000'

The `columns` key represents the date of the measurement. The data is comprised of 6 time series:

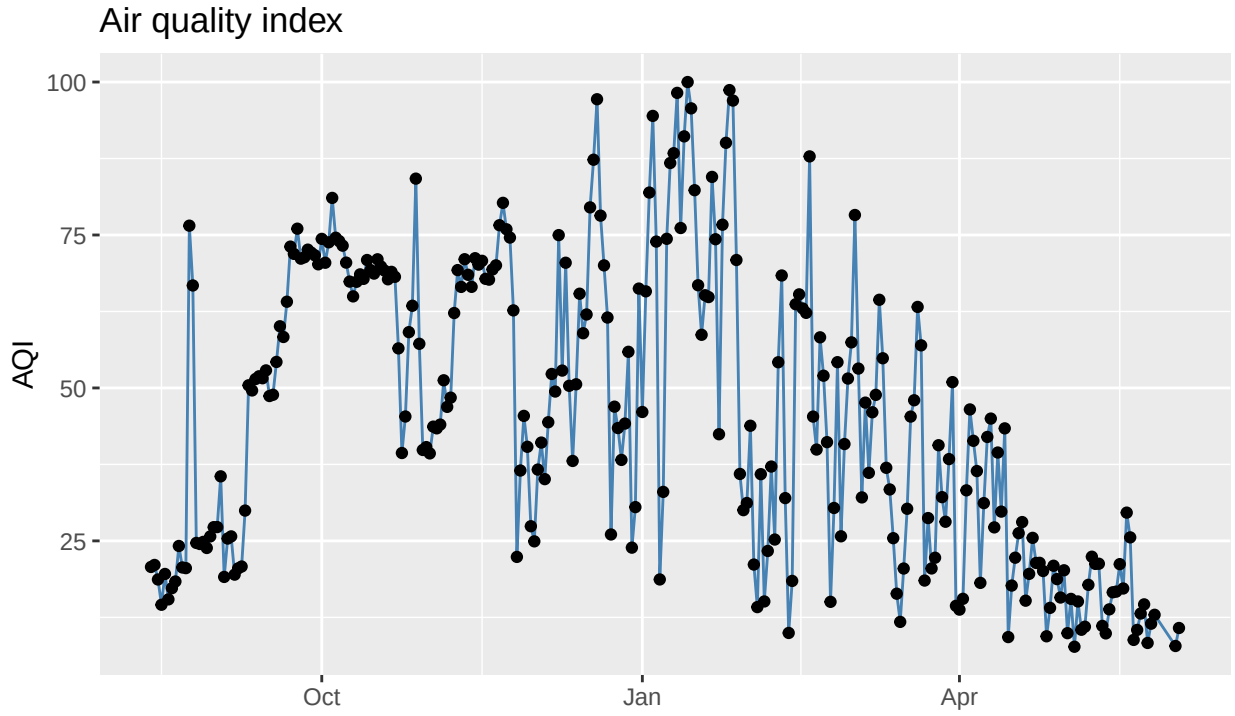
- ica_data
- pm10_data
- pm25_data
- co_data
- no2_data
- so2_data

I will limit this analysis to ICA (indicele calității aerului - AQI). From Wikipedia (https://en.wikipedia.org/wiki/Air_quality_index), the air quality index is used by governments to inform the general public about how polluted is the air or how polluted it will become.

According to the data source website (https://aerlive.ro/cum-masuram-poluarea/), since the project is new, only PM10 and PM2.5 are used to compute the AQI (ICA). The measurements for air pollutants CO, NO_2 and SO_2 are in testing phase ("1-3 months"). It is unspecified if these have been included by now, because the data spans almost ten months.

PM10 (Particulate matters) are particles with a diameter 2.5 and 10 micrometers; PM2.5 is for particles that have diameter 2.5um or less (https://en.wikipedia.org/wiki/Particulates).

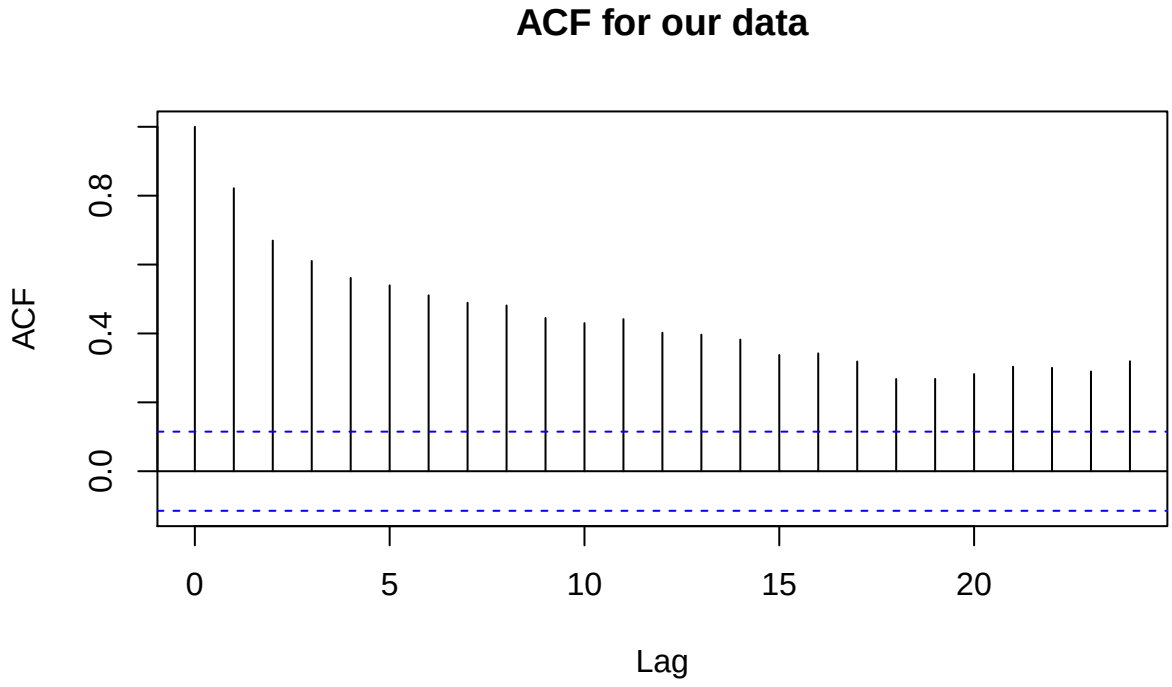
The data points are plotted below:

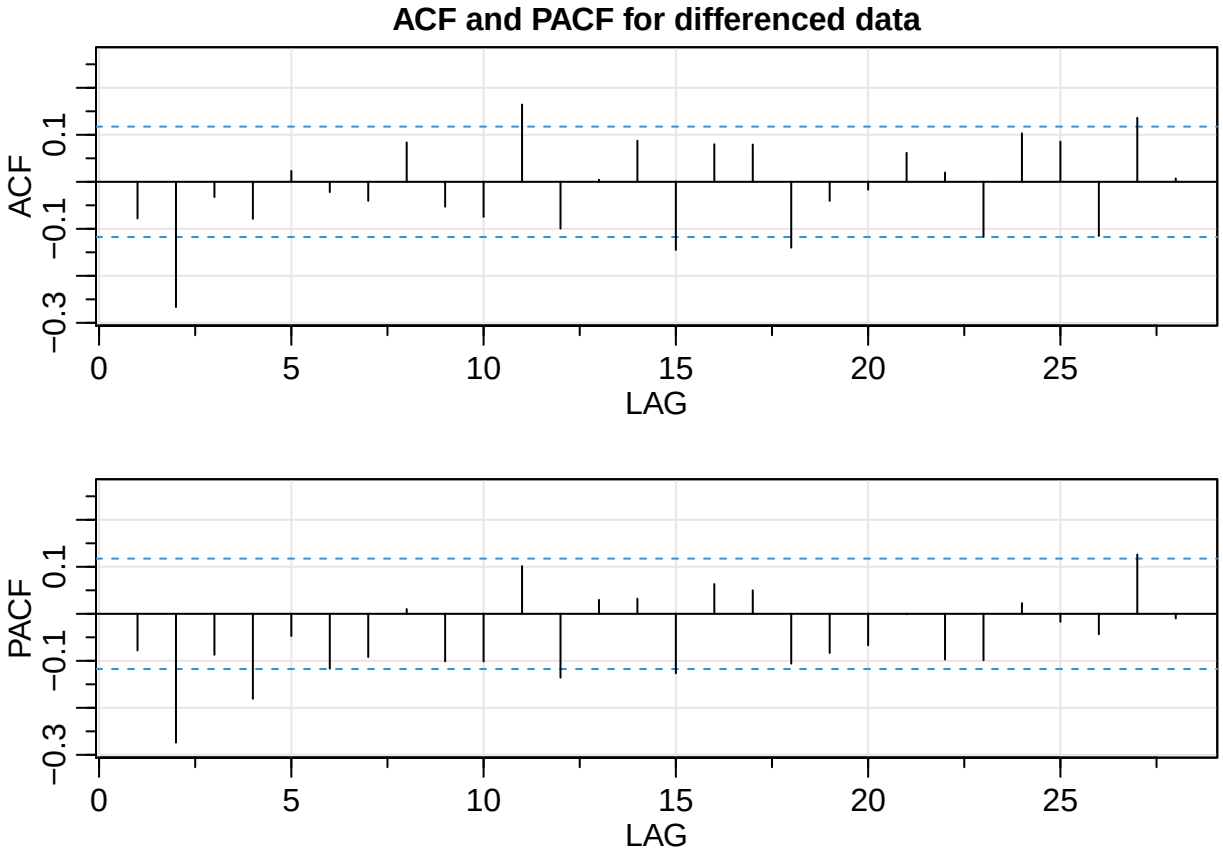


Building the model

From Lecture 6-1, a slowly decaying ACF (auto-correlation function) tells us if differencing is likely necessary. In order to verify this, first we plot the ACF.

Next, we plot the auto-correlation and partial auto-correlation functions for the differenced data.





The ACF plot suggests (from my understanding) a possible MA(3) or MA(5) model. The PACF plot suggests a possible AR(3). The plots might also suggest larger models, but the risk of overfitting is greater.

Unfortunately, there is not enough data to observe the seasonality, and the observed trend might be a bit misleading.

Next, we plot the residuals in order to see if the model fits our data well. Below, I only show the results for the best model that I found (based on the indications from the lectures/laboratory), because plots take up a lot of space. They are included in the notebook code, but they are simply not rendered in the output document. I have experimented with MA(3), MA(5) and AR(3).

AR(3) fails the Ljung-Box test.

MA(5) has slightly higher AIC and BIC scores, so I only included results for MA(3).

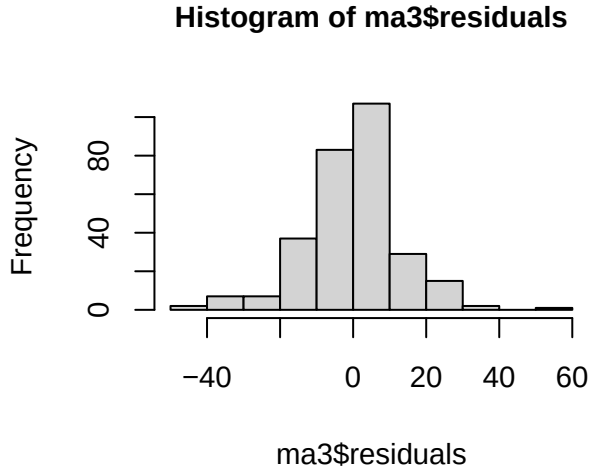
Analyzing the model, diagnostics

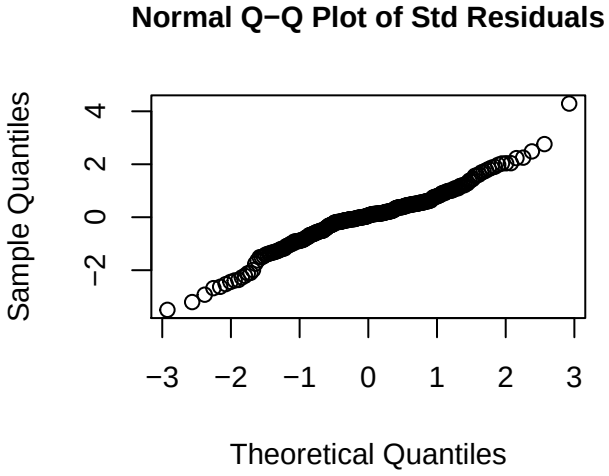
The histogram of residuals and the Q-Q plot are shown below. The Q-Q plot seems approximately linear. However, I have used the Shapiro-Wilk normality test and it strongly rejects the null hypothesis, which means that residuals do not come from a normal distribution (from the histogram, it is indeed skewed).

```
## [1] "AIC: "            "2322.22127623599"
```

```
## [1] "BIC: "            "2340.57068085089"
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ma3$residuals
## W = 0.96304, p-value = 9.413e-07
```

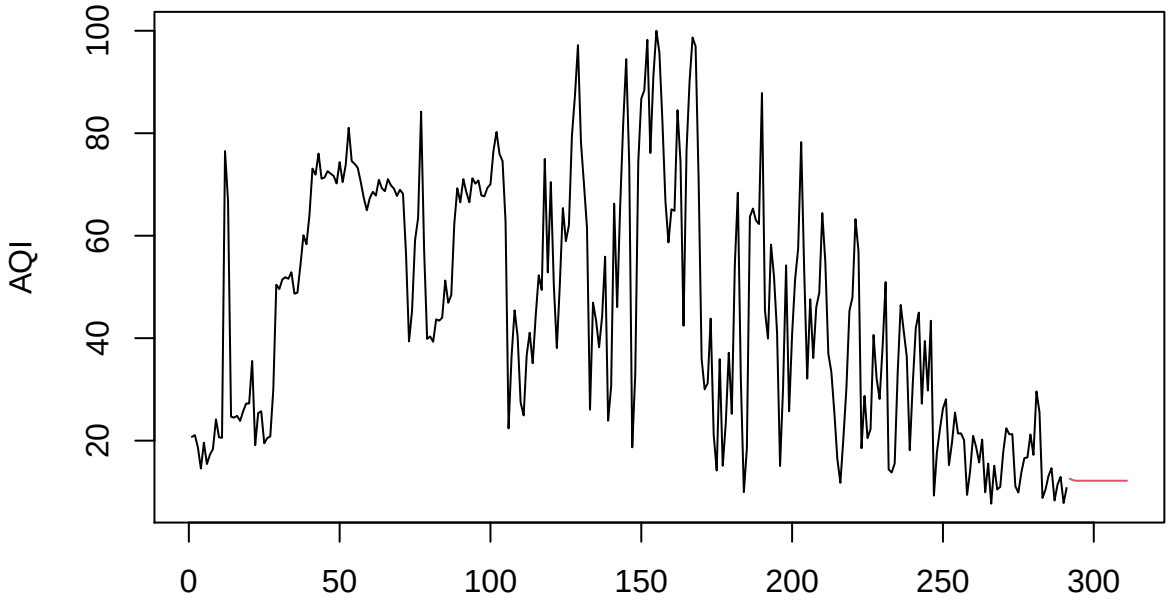




Predicting new values

We differenced once, so `d=1` and the model would be ARMA(0, 1, 3). The next 20 predicted values are plotted below. Due to the lack of better data (based of what I have read besides the lectures - see references), it is not very surprising to obtain a flat prediction. From my understanding, this happens to be the case when (at least for this model) the data does not provide a better trend. We might also obtain better results with

a seasonal model, since that is usually the case with weather-related phenomena. It might also be the case that further initial transformations of the data are required.



2. Two-way ANOVA

Datasets used

The first dataset is taken from http://users.stat.ufl.edu/~winner/data/birchpollen.dat and it originally appeared in the article "Morphological Differentiation of Betula (birch) Pollen in Northwest North America and its Palaeological Application" [Clegg et al.]. It consists of grain diameters of pollen from 5 species, "with varying numbers of locations within species (13, 6, 5, 19, 12), and 30 measurements per location".

The second dataset is from "Cooking Quality of Oregon-Grown Russet Potatoes" [Mackey and Stockman] (http://users.stat.ufl.edu/~winner/data/potato.dat). This experiment is the following: we have several factors regarding growing, storing and cooking potatoes. We are interested in finding the (combination of) factors that influence texture, flavor and moistness (according to some judges).

The factors have the following levels:

- growing area: 1 = Southern Oregon, 2 = Central Oregon
- two week holding temperature: 1 = 75°F (23.89°C), 2 = 40°F (4.44°C)
- size: 1 = large, 2 = medium
- storage period: 1 = 0 months, 2 = 2 months, 3 = 4 months, 4 = 6 months
- cooking method: 1 = boil, 2 = steam, 3 = mash, 4 = bake@350, 5 = bake@450

I used a text editor to remove the space from the beginning of each line and to replace the spaces with tab characters for easier processing in R.

After pre-processing, we test for the homogeneity of variances.

First experiment

```
##
##  Bartlett test of homogeneity of variances
##
## data:  groups
## Bartlett's K-squared = 159.02, df = 54, p-value = 2.772e-12
```

After looking closer at the data, we eliminate groups 1, 4 and 5, because ANOVA requires the assumption of homogeneity of variances (Lecture 10) for it to be meaningful and the previous Bartlett test disproved that assumption. Other combinations of the groups based on species do not seem to pass this statistical test. I did not attempt to remove levels based on location.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  groups
## Bartlett's K-squared = 14.953, df = 10, p-value = 0.1338
```

I am not sure if the methodology above is correct, because it looks to me as a bit of p-hacking. This is one of the reasons I used a second dataset.

We obtain that species and location have both a significant influence on the diameter of birch pollen (keeping in mind that we have eliminated three species from analysis). We also get a p-value of 0.00527 for "species" alone.

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## species       1   17.7   17.69   7.892 0.00527 **
## location      9  966.8  107.42  47.923 < 2e-16 ***
## Residuals   319  715.0    2.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Second experiment

Testing homogeneity for `holding temperature` and `size` for all three effects: `texture`, `flavor` and `moistness`.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  groups
## Bartlett's K-squared = 1.3046, df = 3, p-value = 0.728
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  groups
## Bartlett's K-squared = 3.0492, df = 3, p-value = 0.3841
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  groups
## Bartlett's K-squared = 3.8062, df = 3, p-value = 0.2832
```

The groups for holding temperature and size seem to have the same variance across all effects.

The holding temperature is significant for all effects, but size does not seem to influence flavor. Here we have `flavor~holding_temperature*size`:

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## tw_hold_temp         1  1.089  1.0890  14.627 0.000189 ***
## size                 1  0.000  0.0002   0.003 0.953865
## tw_hold_temp:size    1  0.016  0.0160   0.215 0.643597
## Residuals          156 11.615  0.0745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we have `moistness~holding_temperature*size`:

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## tw_hold_temp         1  1.332  1.3323   8.232 0.004686 **
## size                 1  2.025  2.0250  12.513 0.000532 ***
## tw_hold_temp:size    1  0.025  0.0250   0.154 0.694824
## Residuals          156 25.246  0.1618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we have `texture~holding_temperature*size`. Even though temperature and size are important on their own, the combining factor is irrelevant for moistness and texture:

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## tw_hold_temp         1  5.968   5.968  30.130 1.60e-07 ***
## size                 1  3.938   3.938  19.880 1.57e-05 ***
## tw_hold_temp:size    1  0.127   0.127   0.639    0.425
## Residuals          156 30.898   0.198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The cooking method does not seem to have a significant influence over texture, but combined with growing area it does:

```
##
##  Bartlett test of homogeneity of variances
##
## data:  groups
## Bartlett's K-squared = 2.658, df = 9, p-value = 0.9763
```

```
##                              Df Sum Sq Mean Sq F value   Pr(>F)
## cooking_method                4   1.43  0.3567   1.634 0.168502
## growing_area                  1   1.54  1.5406   7.060 0.008736 **
```

```
## cooking_method:growing_area   4   5.23  1.3076   5.992 0.000168 ***
## Residuals                   150  32.73  0.2182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we have `flavor~storage_period*cooking_method`, where both factors determine the flavor.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  groups
## Bartlett's K-squared = 20.119, df = 19, p-value = 0.3874
```

```
##                                Df Sum Sq Mean Sq F value   Pr(>F)
## storage_period                  3  2.024  0.6747  11.950 5.16e-07 ***
## cooking_method                  4  1.344  0.3359   5.948 0.000189 ***
## storage_period:cooking_method  12  1.447  0.1206   2.136 0.018192 *
## Residuals                     140  7.905  0.0565
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The other cases are similar to those above or do not pass Bartlett's test.

3. Linear regression

The dataset used is about the impact of container capacity for recycling on yield of materials [Baird et al.] from a recycling program in Scotland (http://users.stat.ufl.edu/~winner/data/scottish_recycle.dat).

```
##
## Call:
## lm(formula = x ~ y1 + y2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83659 -0.31913  0.03302  0.35019  1.02483
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.721744   0.577014   4.717 6.01e-05 ***
## y1           0.012142   0.003113   3.901 0.000548 ***
## y2          -0.006432   0.002002  -3.213 0.003298 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5107 on 28 degrees of freedom
## Multiple R-squared:  0.7092, Adjusted R-squared:  0.6884
## F-statistic: 34.14 on 2 and 28 DF,  p-value: 3.094e-08
```

We build the model based on recycling capacity and residual capacity (measured in liters/week).

For y1, we have p-value of 0.000548, so H:$\{\beta_1 = 0\}$ is rejected.

For y2, we have p-value of 0.003298, so H:$\{\beta_2 = 0\}$ is also rejected.

We accept both hypotheses of regression of x in y_1 and y_2.

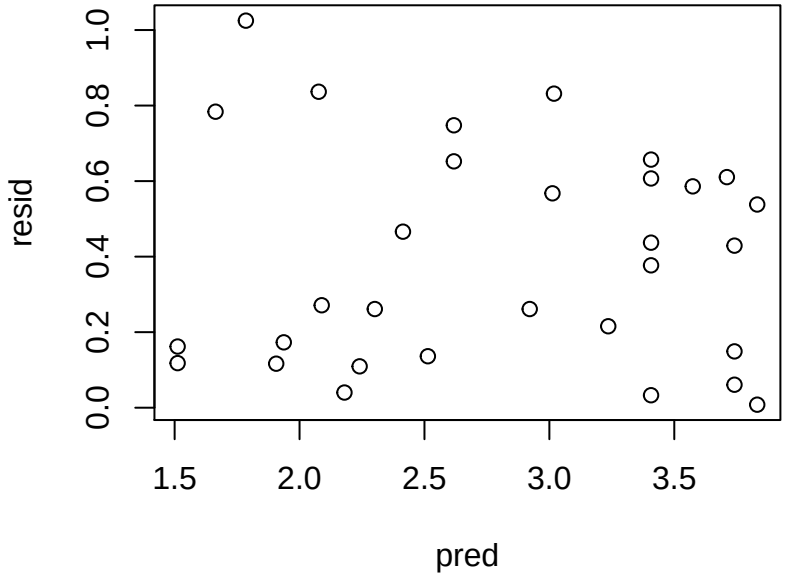
Since the R-squared value is not very close to 1, it might be the case that some "causes" are missing from the model.

```
## Analysis of Variance Table
##
## Response: x
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## y1         1 15.1165 15.1165  57.962 2.712e-08 ***
## y2         1  2.6918  2.6918  10.321  0.003298 **
## Residuals 28  7.3025  0.2608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the Shapiro-Wilk test, the residuals are normal, so the model proposed respects this assumption.

```
##
##  Shapiro-Wilk normality test
##
## data:  lmodel$residuals
## W = 0.9761, p-value = 0.698
```

From the goodness-of-fit analysis, we observe a bit of concentration to the right.



We repeat the analysis after adding the number of collected materials to the model.

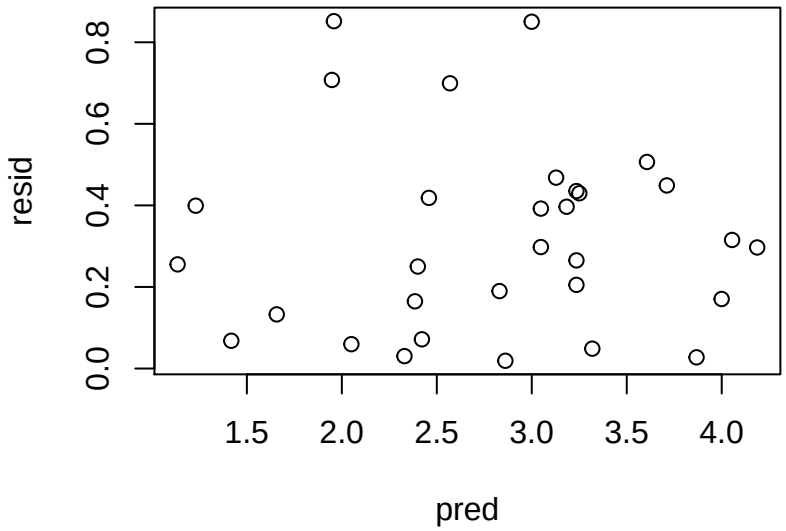
```
##
## Call:
```

```
## lm(formula = x ~ y1 + y2 + y3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70762 -0.28102 -0.02743  0.28266  0.85171
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.176991   0.496792   4.382 0.000160 ***
## y1           0.007360   0.002860   2.573 0.015887 *
## y2          -0.006347   0.001650  -3.848 0.000661 ***
## y3           0.187319   0.049617   3.775 0.000799 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4207 on 27 degrees of freedom
## Multiple R-squared:  0.8097, Adjusted R-squared:  0.7885
## F-statistic: 38.29 on 3 and 27 DF,  p-value: 7.253e-10
```

The residuals are still normal, which is a good sign.

```
##
##  Shapiro-Wilk normality test
##
## data:  lmodel2$residuals
## W = 0.97519, p-value = 0.6705
```

From the goodness-of-fit analysis, we observe a bit of concentration to the right. The values seem a bit more scattered now, but are still concentrated a little right next to the middle.



References

- the code is inspired from laboratories; I also tried to include below most of the resources I used
- https://www.r-graph-gallery.com/279-plotting-time-series-with-ggplot2
- http://www.cookbook-r.com/Graphs/Titles_(ggplot2)/
- https://www.earthdatascience.org/courses/earth-analytics/time-series-data/date-class-in-r/
- https://stackoverflow.com/questions/58907202/
- https://stats.stackexchange.com/questions/111010/
- https://stats.stackexchange.com/questions/143073/non-normal-residuals-in-arima
- https://stats.stackexchange.com/questions/79400/
- https://stackoverflow.com/questions/37115276/
- https://stackoverflow.com/questions/4241798/how-to-increase-font-size-in-a-plot-in-r
- https://stats.stackexchange.com/questions/286900/arima-forecast-straight-line
- https://stats.stackexchange.com/questions/124955/is-it-unusual-for-the-mean-to-outperform-arima/125016#125016
- https://rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf
- https://stackoverflow.com/questions/28032846/
- B.F. Clegg, W. Tinner, D.G, Gavin, F.S. Hu (2005). "Morphological Differentiation of Betula (birch) Pollen in Northwest North America and its Palaeological Application" The Holocene, Vol. 15, #2, pp. 229-237. (from http://users.stat.ufl.edu/~winner/data/birchpollen.txt)
- http://www.sthda.com/english/wiki/two-way-anova-test-in-r
- Source: A. Mackey and J. Stockman (1958). "Cooking Quality of Oregon-Grown Russet Potatoes", American Potato Journal, Vol.35, pp. 395-407 (http://users.stat.ufl.edu/~winner/data/potato.txt)
- https://en.wikipedia.org/wiki/Two-way_analysis_of_variance
- https://stackoverflow.com/questions/2933253/homoscedascity-test-for-two-way-anova
- https://stats.stackexchange.com/questions/60410/normality-of-dependent-variable-normality-of-residuals
- https://stats.stackexchange.com/questions/35132/assessing-normality-of-distribution
- https://www.sheffield.ac.uk/polopoly_fs/1.536444!/file/MASH_2way_ANOVA_in_R.pdf
- https://www.gormanalysis.com/blog/r-introduction-to-factors-tutorial/
- https://explorable.com/two-way-anova
- J. Baird, R. Curry, and T. Reid (2013). "Development and Application of a Multiple Linear Regression Model to Consider the Impact of Weekly Waste Container Capacity on the Yield from Kerbside Recycling Programs in Scotland," Waste Management & Research, Vol. 31, pp. 306-314 (http://users.stat.ufl.edu/~winner/data/scottish_recycle.txt)
- http://www.calvin.edu/~rpruim/courses/m343/F12/RStudio/LatexExamples.html
- https://en.wikipedia.org/wiki/Data_dredging