**Exercise 1**

1. The resulting matrix shows a suggested random assignment for the different experiment units, for example:

elements of the matrix: (1,2)=1,(2,2)=2,(3,2)=7 ==> the suggestion would be then to assign the unit 7 to temperature level 1 and humidity level 2.

1. Graphs – TODO: Add some lines about the graphs (eg. Interaction – non-parallel lines)
2. Conclusion:

Considering the decomposition of the population mean as:

u(i,j) = u + a(i) + b(i) + y(i,j)

As per the results from ANOVA:

The p-value testing H0: a(i) for all i is equal to 0 is 2.461e-10 (reject H0)

The p-value testing H0: b(i) for all i is equal to 0 is 4.316e-06 (reject H0)

The p-value testing H0: y(i,j) for all (i,j) is equal to 0 is 3.705e-07 (reject H0)

Thus, it’s possible to say that both factors have a main effect given that the null hypothesis for both of them was rejected (so they are a factor <> 0) in the equation mentioned above.

The fact that H0 is also rejected for y(i,j) means that there might be interaction between both factors.

As could be seen in the interaction plots from the item 1.2 there is an indication of interaction by the existence of non-parallel lines.

1. As can be seen on the estimations for the factors (environment1: 149.333 // environment2: -64.667), the environment has the biggest numerical influence over the time decay.

This can be a good question if it is important to know which factor could cause bigger changes in the outcome/result that is been measured.

However, is also important to notice that, as shown on the results above, that there is an interaction factor between the two factors. So it may happen that the combination of them could lead to more significant changes in the result.

1. *Checking the normality of the population.*

The residuals seems to follow a normal distribution, there are some extreme values outliers) though.

An extra check is also the Shapiro test. In this case, we also fail to reject the null hypothesis (that the sample) follows a normal distribution (p-value = 0.1911).

*Checking the assumption of equal population variances.*

As can be seen, there are two data-points that are extreme.

However, in general the residuals remains constant for different fitted values which is in line with the assumption of "of equal population variances".

**Exercise 2**

**Exercise 3**

1. The p-values produced with anova and summary commands are not simultaneous. The p-values in the lines starter2 to starter5 are for the hypothesis , and so on for the main effect of starter. This is the same for batch and position. According to the summary, starter4 has the least p-value equals to 6.10e-05. Therefore, we can conclude that the null hypothesis is strongly rejected and starter4 plays a significant role in the main model of the acidity. Similarly, bath2 and batch4 null hypothesis can be rejected as well with p-values equal to 0.0137 and 0.0127 respectively. The abstract model is: . To sum up, equals to for i = 2,3,5 (except ). For j = 2 and 4, is similar to (excluding . Finally, all in range (2,5) is the same as .
2. Here the table is based on simultaneous p-values for the null hypothesis. As it is clear from the table, starter4 has the most difference with others and significantly affects the acidity. P-values for cases, , , and, are less than 0.001 hence null hypothesis are rejected.
3. In simultaneous comparisons, the more inferences are made, the more likely erroneous inferences are to occur. Consequently, the p-value in part 2 is less trusted than in part 1 because of the error.
4. The intervals for , ,and, don’t contain 0. As we concluded in part 2, the starter4 lead to significantly different acidity. So, the confidence intervals shouldn’t include 0.

**Exercise 4**

**Exercise 5**

1. Script
2. The outcome of xtabs(~medicin+nausea) presents the data contained in the data frame in a aggregated view in which the different factors (in this case medicines) are presented in different rows and their respective outcomes (aggregated) are presented in the columns, in this case one column with the number of patients with or without nausea.

It’s possible to see from xtabs that Chlorpromazine had the biggest number of samples (patients) compared to the other medicines. It also had the best performance (when considering the number of patients without nausea after the treatment).

1. Given that the value for p-value = 0.042 we reject the null hypothesis that the populations are the same, in other words, there might be a difference between the different treatments (bear in mind that the p-value can change for each computation which can lead to a different conclusion about rejecting or not the null hypothesis).
2. The p-value calculated from chisquare test is different from the value calculated by the permutation test. The p-value would not be the same as the it is obtained from a random distribution on the permutation test.

**Exercise 6**

1. According to scatterplots, wind and temperature have the most influence on oxidant. Their plots approximately show linear correlation. Afterward, insolation and humidity could affect oxidant linearly.
2. Among all explanatory variables, wind has the least p-values. (8.20e-07). Therefore, it should be selected as the basis to start to find linear regression model. In step 2, temperature is chosen to be added to the model with p-value 1.17e-06. Adding insolation (third smallest p-value) to the model in step 3, its p-value demonstrates it doesn’t lead to significant affect. Finally, the appropriate model consists of two explanatory variables which are wind and temperature.
3. In the first and second and third step, day, insolation and humidity are removed from the model respectively due to their p-values. In this part, similarly we reach to the same model as in previous part. This model is: oxidant ~ wind + temperature
4. Wind and temperature equal to -0.427and 0.52 respectively.
5. To investigate the normality of residuals, we consider the QQ-plot. We can assume normality based on the graph. Moreover, the samples are scattered approximately in whole area according to the fitted-residuals plot.

**Exercise 7**

Using the function pairs it’s possible to see a potential correlation between expend and bad, lawyers, employ and pop.

Thus a first attempt to calculate the linear model will be made considering these factors.

After assessing the first model (with the qqnorm and the graph between fitted and residuas) it is possible to see that the variances for the different fitted values is concentrated in a region around smaller value of fitted expenses.

INCLUDE INFORMATION ABOUT SHAPIRO AFTER EACH ITERATION

2nd Iteration:

Considering that in the previous iteration lawyers and employ reject the null hypothesis the 2nd iteration will consider only these factors (bad will also be considered since it has the biggest estimated coeficient).

Additionally, now we will calculate the regression considering interaction between the variables.

Assessing the new regression parameters:

The qqnorm presents a curved shape with some points far from the line, the qqnorm shows a concentration around certain fitted values (<1000).

Given that bad is the variable with highest coeficient we will try to have better results by elevating bad to the power of 2.

After a third iteration the qqnorm graph presents a better slope and distance between the points.

Also the residuals x ffited graph shows that the residuals are more spread.

The model is:

expenses = 136.3 -8.186\*bad -0.1297\*lawyers + 0.08236\*employ - 0.1440\*bad^2