Diego Garcia-Olano
SMDE Fall 2013

DELIVERABLE FOR 12/17

FIRST PART

We want to define a complete set of experiments to analyze what is the best alternative for a specific modification on the system. We cannot modify all the factors that can affect the answer, so we must detect those more important and then define a DOE to conduct the experiments.

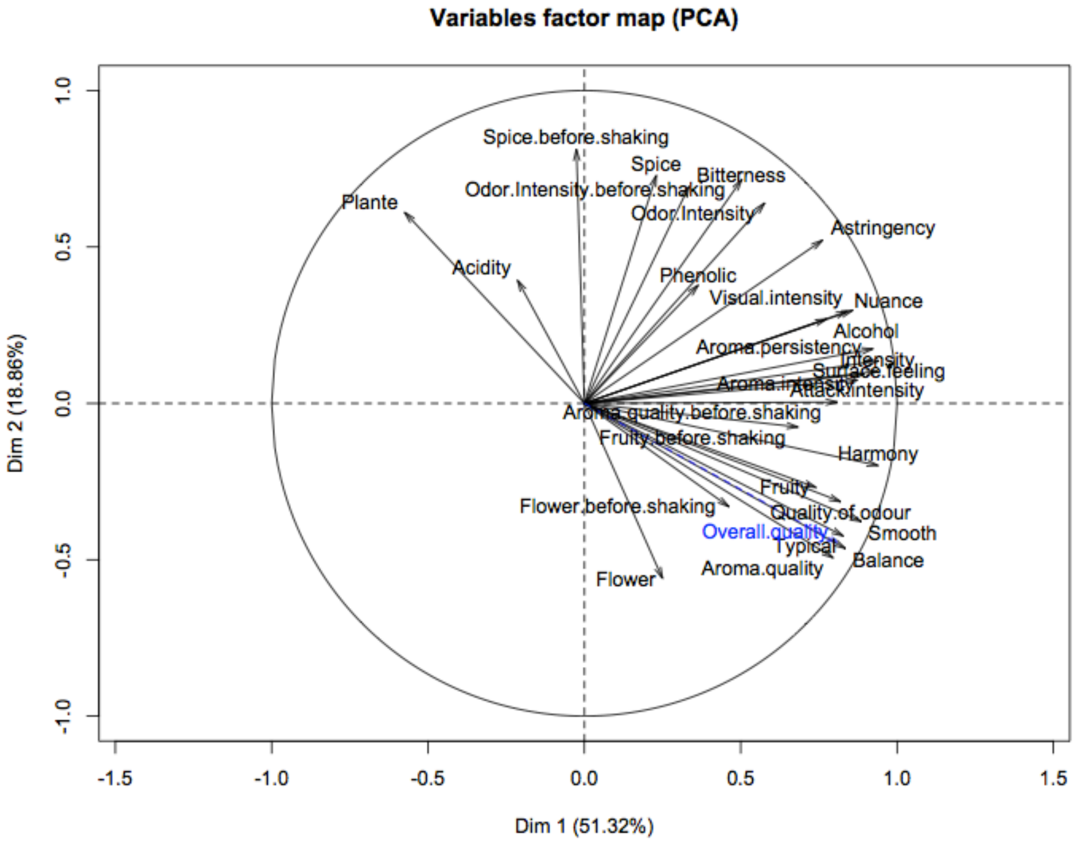
We use as data **decathlon** and **wine** of FactoMiner package. First we explore the data and depict the underlying relations between the different factors (use PCA).

Once this is done, for each dataset we set the objectives, defining the answer variable therein, select the factors, define an experimental design, build a model, and then test the model.

We first explore the **wine** data; a data frame with 21 rows (the number of wines) and 31 columns that refer to 21 wines of Val de Loire.

1. I would like to see the relation to Overall.quality of wine for other factors in the system. Another way of saying this is what factors are best predictors of Overall.quality of wine. Overall.quality of wine is thus our answer variable, and we will use PCA to determine what are the most relevant other factors which explain its variance in order to see what would be the best alternatives for improving that variable.

2. Based on the PCA output for variables, factors most similar to Overall.quality include Typical, Balance, Smooth, Aroma.quality, Quality.of.odour, Fruity, Fruity before shaking, Harmony, and Flower. These variables are thus positively correlated in some degree to Overall.quality and improvements in them be the best alternative for modifications on the system. Similarly, factors which are negatively related to Overall.quality include Acidity and Plante.



3. For the next step we need to define an experimental design.

combo	X100 (a)	X100m.hurdle (b)	X400 (c)	Long.jump (d)	Pole.vault (e)	Description
(1)	-	-	-	-	-	unchanged state
a	+	-	-	-	-	X100
b	-	+	-	-	-	X100.hurdle
ab	+	+	-	-	-	
c	-	-	+	-	-	X400
ac	+	-	+	-	-	
bc	-	+	+	-	-	
abc	+	+	+	-	-	
d	-	-	-	+	-	Long.jump
ad	+	-	-	+	-	
bd	-	+	-	+	-	
cd	-	-	+	+	-	
abd	+	+	-	+	-	
acd	+	-	+	+	-	
bcd	-	+	+	+	-	
abcd	+	+	+	+	-	
e	-	-	-	-	+	Pole.vault(e)
ae	+	-	-	-	+	
be	-	+	-	-	+	
ce	-	-	+	-	+	
de	-	-	-	+	+	
abe	+	+	-	-	+	
ace	+	-	+	-	+	
ade	+	-	-	+	+	
bce	-	+	-	-	+	
bde	-	+	-	+	+	
cde	-	-	+	+	+	
abce	+	+	+	-	+	
abde	+	+	-	+	+	
acde	+	-	+	+	+	
bcde	-	+	+	+	+	
abcde	+	+	+	+	+	

There are k = 11 factors at play so we would have 2^k (2048) rows in our table

(one row for each possible combination of factors) if we wrote it out completely.

We would not be able to run these many experiments in a realistic situation based on time and cost.

4. Build a model to see relation of factors to Overall.quality. This tests to see how the factors listed predict Overall quality

```
>RegModel.4 <-
lm(Overall.quality~Acidity+Aroma.quality+Balance+Flower+Fruity+Fruity.before.shaking+Harmony+Pl
ante+Quality.of.odour+Smooth+Typical,  data=wine)
```

```
>summary(RegModel.4)
```

```
Call: lm(formula = Overall.quality ~ Acidity + Aroma.quality + Balance +   Flower + Fruity +
Fruity.before.shaking +  Harmony + Plante +  Quality.of.odour + Smooth + Typical, data = wine)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.15877	-0.08013	0.00788	0.05970	0.18439

```
Coefficients:
                      Estimate      Std. Error  t value   Pr(>|t|)
(Intercept)           1.00798       1.59893      0.630   0.5441
Acidity               0.14734       0.17170     0.858   0.4131
Aroma.quality         -0.45458      0.37568     -1.210  0.2571
Balance               0.20209       0.49652     0.407   0.6935
Flower                  0.72199     0.34593     2.087   0.0665 .
Fruity                0.11676       0.55821     0.209   0.8390
Fruity.before.shaking  0.07459      0.47312     0.158   0.8782
Harmony               -0.09210      0.25966     -0.355  0.7310
Plante                -0.97036      0.49982     -1.941  0.0841 .
Quality.of.odour       0.49182      0.37624     1.307   0.2236
Smooth                 0.97358      0.57064     1.706   0.1222
Typical               -0.42653      0.34765     -1.227  0.2510
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1422 on 9 degrees of freedom
Multiple R-squared:  0.9512,     Adjusted R-squared:  0.8916
F-statistic: 15.96 on 11 and 9 DF,  p-value: 0.0001379
```

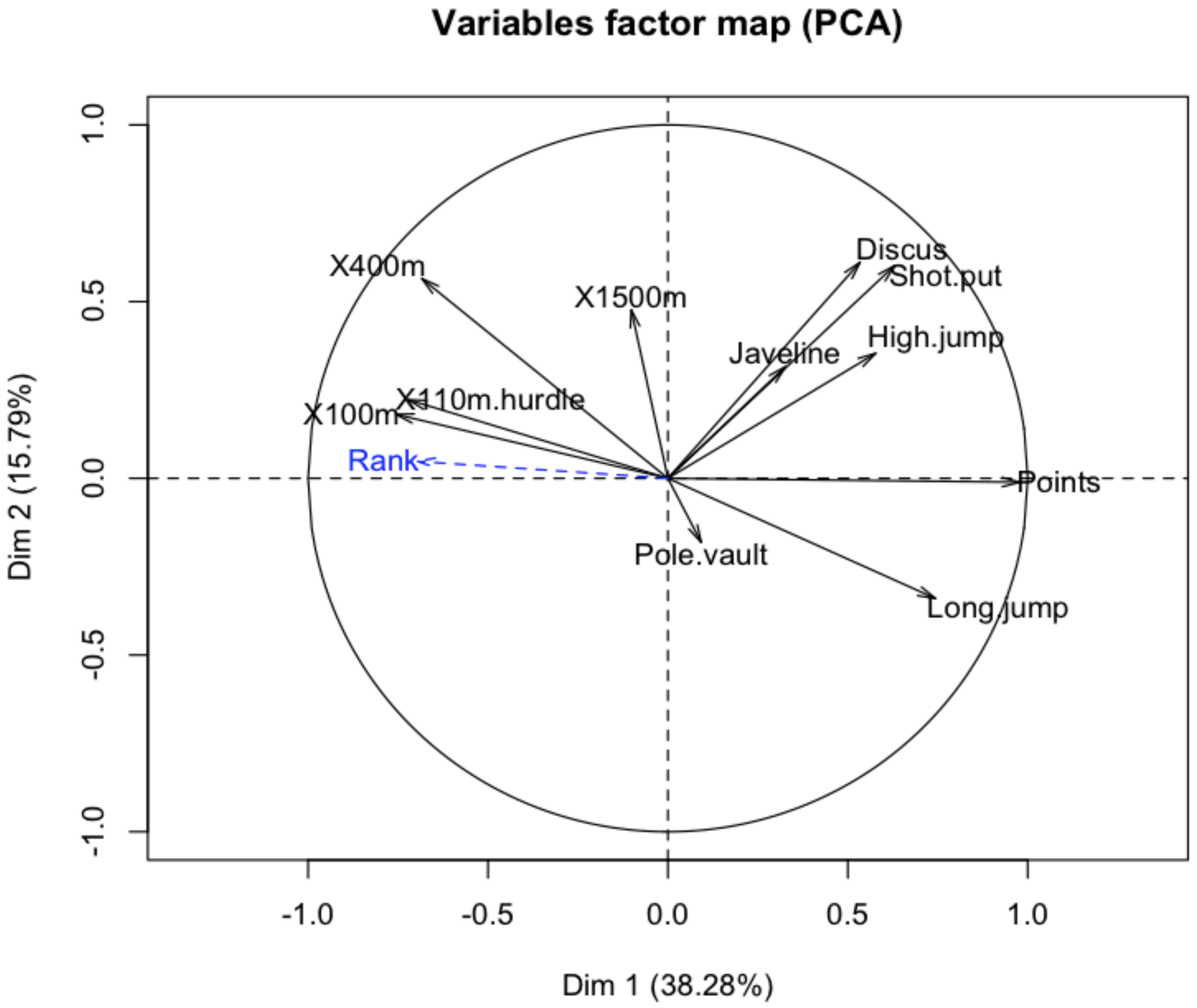
This output shows that these factors combined explain 95 of the amount of variability for our prediction explained by our model. It shows that Flower and Plant have the least probability of not being relevant as predictors of Overall.quality. Smooth, Quality.of.odour, Typical, and Aroma.quality respectively are the next best predictors.

5. Test the model.
We are not able to realistically test this model as is because we do not know how changes in each factor would change the overall score of Overall.quality as we are unable to replicate the experiment.

With the **decathlon** data, which is data frame with 41 rows and 13 columns, corresponds to the performance of the athletes for the 10 events of a decathlon, and their rank and the points obtained during two different sporting events.

1. I would like to see the relation to Rank for other "events" factors in the system. How do the events themselves predict rank? Rank is thus our answer variable and our objective is to see what factors are the best predictors of rank, and would be candidates to explore in an experiment

2. Based on the PCA output for variables, factors most correlated to Rank include X100, X100.hurdle and X400 and will be explored farther in our experiment. I discarded Points because it will be perfectly correlated to Rank and thus uninteresting. Similarly, factors that are negatively related to Rank include LongJump and Pole.vault and will be included as factors that effectively predict the inverse of rank.



3. For the next step we need to define an experimental design.

combo	X100 (a)	X100m.hurdle (b)	X400 (c)	Long.jump (d)	Pole.vault (e)	Description
(1)	-	-	-	-	-	unchanged state
a	+	-	-	-	-	X100
b	-	+	-	-	-	X100.hurdle
ab	+	+	-	-	-	
c	-	-	+	-	-	X400
ac	+	-	+	-	-	
bc	-	+	+	-	-	
abc	+	+	+	-	-	
d	-	-	-	+	-	Long.jump
ad	+	-	-	+	-	
bd	-	+	-	+	-	
cd	-	-	+	+	-	
abd	+	+	-	+	-	
acd	+	-	+	+	-	
bcd	-	+	+	+	-	
abcd	+	+	+	+	-	
e	-	-	-	-	+	Pole.vault(e)
ae	+	-	-	-	+	
be	-	+	-	-	+	

ce	-	-	+	-	+	
de	-	-	-	+	+	
abe	+	+	-	-	+	
ace	+	-	+	-	+	
ade	+	-	-	+	+	
bce	-	+	-	-	+	
bde	-	+	-	+	+	
cde	-	-	+	+	+	
abce	+	+	+	-	+	
abde	+	+	-	+	+	
acde	+	-	+	+	+	
bcde	-	+	+	+	+	
abcde	+	+	+	+	+	

4. Build a model to see relation of factors to Rank. This tests to see how the factors listed predict Rank.

```
> RegModel.5 <- lm(Rank~Long.jump+Pole.vault+X100m+X110m.hurdle+X400m, data=decathlon)
> summary(RegModel.5)
Call: lm(formula = Rank ~ Long.jump + Pole.vault + X100m + X110m.hurdle +   X400m, data =
decathlon)

Residuals:
    Min       1Q             Median       3Q   Max
-10.6970  -3.6137  -0.1756   2.6048  13.1210

Coefficients:
                        Estimate       Std. Error      t value          Pr(>|t|)
(Intercept)             46.621         86.001          0.542            0.5912
Long.jump                -10.447        4.171          -2.505           0.0171 *
Pole.vault               -6.572         3.447          -1.907           0.0648 .
X100m                   -7.600         4.865           -1.562           0.1272
X110m.hurdle            3.397          2.619           1.297            0.2031
X400m                   2.148          1.090           1.971            0.0566 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.89 on 35 degrees of freedom
Multiple R-squared:  0.5159,    Adjusted R-squared:  0.4468
F-statistic: 7.461 on 5 and 35 DF,  p-value: 7.465e-05
```

This output shows that these factors combined explain 51 of the amount of variability for our prediction explained by our model.
It shows that Long.jump, Pole.vault and X400 have the least probability of not being relevant as predictors of Rank amongst the factors choosen.

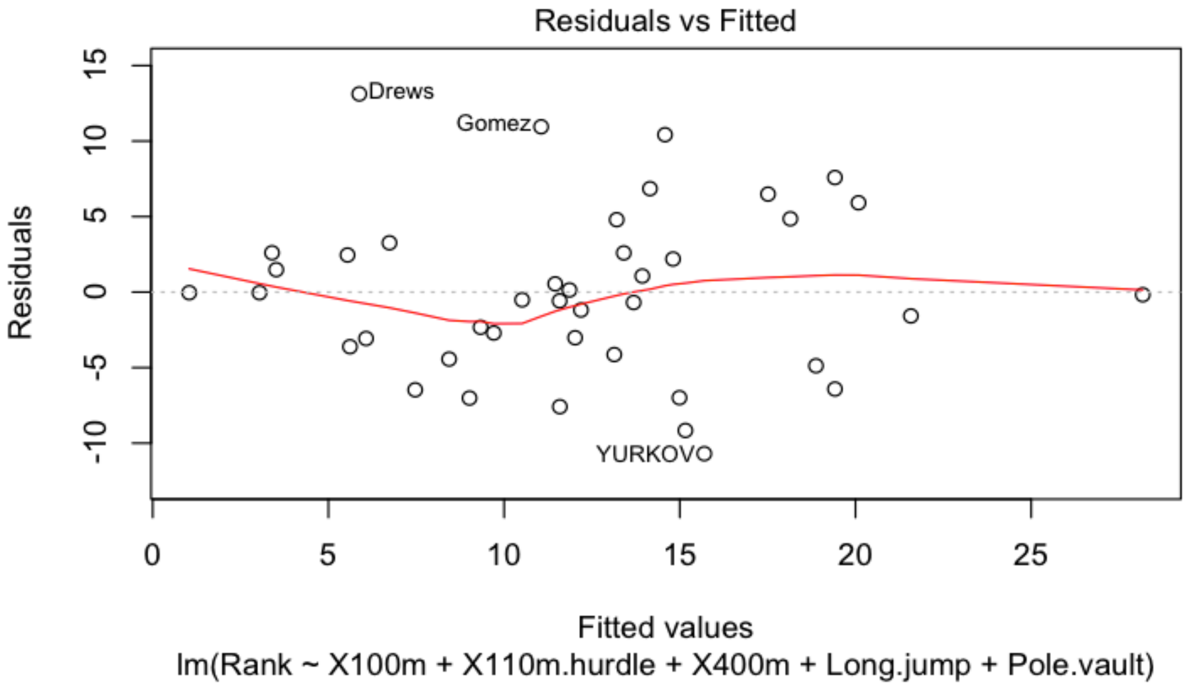
Similarly running anova on the model

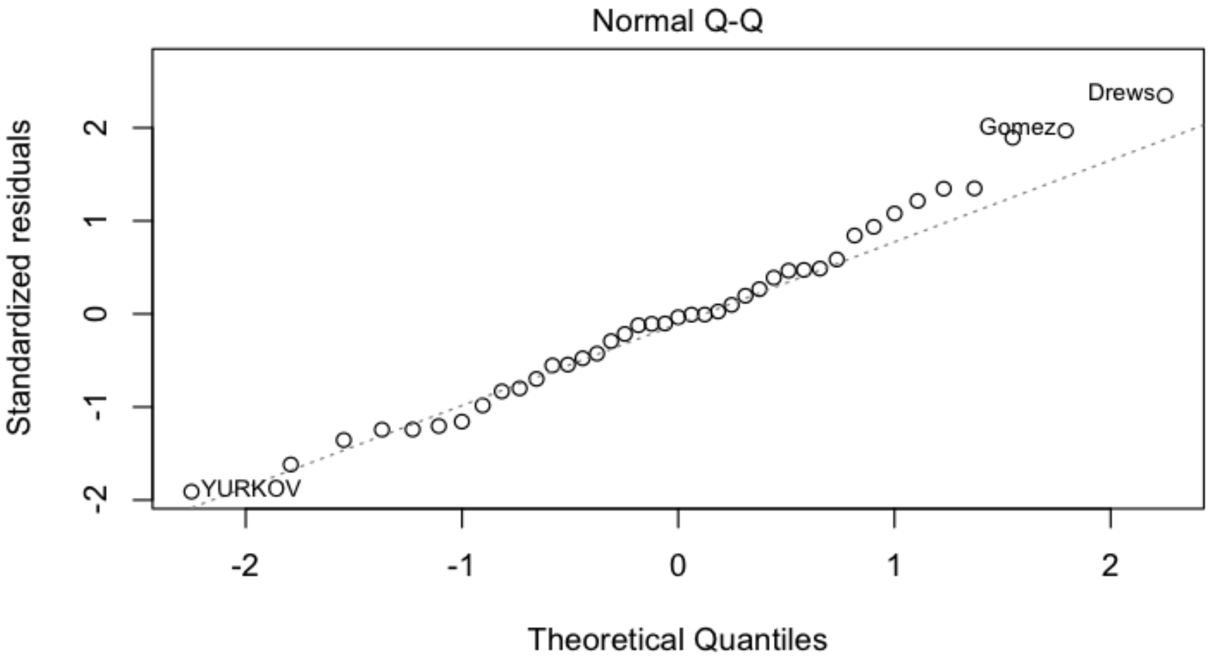
```
> anova(RegModel.5)
Analysis of Variance Table
```

```
Response: Rank
                    Df  Sum Sq Mean Sq     F value    Pr(>F)
X100m               1  220.82        220.82     6.3652     0.016335 *
X110m.hurdle        1  269.51        269.51     7.7687     0.008532 **
X400m               1  374.62        374.62     10.7984 0.002315 **
Long.jump           1  303.11        303.11     8.7372     0.005549 **
Pole.vault          1  126.11        126.11     3.6351     0.064810 .
Residuals           35 1214.22   34.69
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

and we can plot

```
> plot(RegModel.5)
```





5. Test the model.

Similar to the wine data section, we are not able to realistically test this model as is because we do not know how changes in each factor would change the overall Rank score as we are unable to replicate the experiment.

SECOND

Define the dataset to be used on the next deliverables. ‘

For this section, I created a dataset (attached: test-data.xlsx) composed of the top 30 countries according to their GDP's for 2012, and then added their Happiness Index Score, which is largely a function of Life Expectancy / Pollution, and then added how much a Big Mac costs in each country in US dollars, and finished it with some data from the UN 2013 Development report for those countries

(including 2011 GDP, GDP per capita, Health, Education and Military Spending).

There are 3 NA fields (ie, Iran does not have McDonalds, and China doesn't give stats on their education spending), but for class purposes I hope it will be okay to fill these in with reasonable data assumptions.

Below are the links from which the data was derived.

http://unstats.un.org/unsd/snaama/dnltransfer.asp?fID=2

http://www.happyplanetindex.org/data/
http://www.economist.com/content/big-mac-index
http://hdr.undp.org/opendata/