Diego Garcia-Olano
SMDE FALL 2013

Assignment 2: Exercises on MANOVA, GLM and DOE

First Exercise:

Explore the relation between the "soil" and the different wine features

Does Soil affect wine features?

```
> unique(wine$Soil)
[1] Env1      Env2      Reference Env4
Levels: Reference Env1 Env2 Env4

env1 = wine[which(wine$Soil == "Env1"),]
env2 = wine[which(wine$Soil == "Env2"),]
env4 = wine[which(wine$Soil == "Env4"),]
ref = wine[which(wine$Soil == "Reference"),]

lapply(list(env1,env2,env4,ref),nrow)
```

% rows for each = 7, 5, 2, 7 = **21 total**

See which factors have the greatest variability overall and test how they fare in relation to splitting data based on soil types.

```
rang = c();
for(i in 1:28){  rang[i] <- paste(sd(wine[,i+2]),i+2,sep=" - ")}
vars = sort(rang,decreasing=TRUE)[1:6]
vars
[1] "0.549463885896755 - 8"  "0.520739135862715 - 9"  "0.442038966177245 - 29" "0.432013111573513 - 30"
    "0.413986703260129 - 26" "0.372565091788063 - 28"

par( mfrow = c(2,3) )
fs = c()
for(i in 1:6){
        cu = strsplit(vars[i]," - ")[[1]][2]
        fs[i] = labels(wine)[[2]][as.numeric(cu)]
}
```

[1] "Visual.intensity" "Nuance" "Harmony" "Overall.quality" "Smooth" "Intensity"

Explore and analyze the data (at least 6 selected features)
Explain the obtained table and justify your answers

For the following tests, H0: means of factors tested over different soil groups are equal.

1. Visual.intensity based on soil

```
plot(wine$Visual.intensity~wine$Soil, main="Visual intensity vs soil", col=heat.colors(2))
tapply(wine$Visual.intensity,wine$Soil,summary)
tapply(wine$Visual.intensity,wine$Soil,sd)
tapply(wine$Visual.intensity,wine$Soil,length)

fit1 = lm(wine$Visual.intensity~wine$Soil)
anova(fit1)
```

%Pr(>F) = 0.2119 so can't reject null hypothesis

2. Nuance based on soil

```
plot(wine$Nuance~wine$Soil, main="Nuance vs soil", col=heat.colors(2))
tapply(wine$Nuance,wine$Soil,summary)
fit2 = lm(wine$Nuance~wine$Soil)
anova(fit2)
```

%Pr(>F) = 0.241 so can't reject null hypothesis

3. Harmony based on soil

```
plot(wine$Harmony~wine$Soil, main="Harmony vs soil", col=heat.colors(2))
tapply(wine$Harmony,wine$Soil,summary)
fit3 = lm(wine$Harmony~wine$Soil)
anova(fit3)
```

%Pr(>F) = 0.0416 so reject null hypothesis

4. Overall.quality based on soil

```
plot(wine$Overall.quality~wine$Soil, main="quality vs soil", col=heat.colors(2))
tapply(wine$Overall.quality,wine$Soil,summary)
fit4 = lm(wine$Overall.quality~wine$Soil)
anova(fit4)
```

%Pr(>F) =0.009113 ** <-- reject null hypothesis of means of groups being equal

5. Smooth based on soil

```
plot(wine$Smooth~wine$Soil, main="smooth vs soil", col=heat.colors(2))
tapply(wine$Smooth,wine$Soil,summary)
oneway.test(wine$Smooth~wine$Soil,data = wine)
```

% p-value = 0.01431 <-- reject null hypothesis of means of groups being equal

```
fit5 = lm(wine$Smooth~wine$Soil)
anova(fit5)
```

%Pr(>F) =0.02165 <-- reject null hypothesis

6. Intensity based on soil

```
plot(wine$Intensity~wine$Soil, main="intensity vs soil", col=heat.colors(2))
tapply(wine$Intensity,wine$Soil,summary)
oneway.test(wine$Intensity~wine$Soil,data = wine)    %  p-value = 0.01431
```

<-- reject null hypothesis of means of groups being equal

```
fit6 = lm(wine$Intensity~wine$Soil)
anova(fit6)
```

%Pr(>F) =0.03064 * <-- reject null hypothesis

%Find all significant factors.

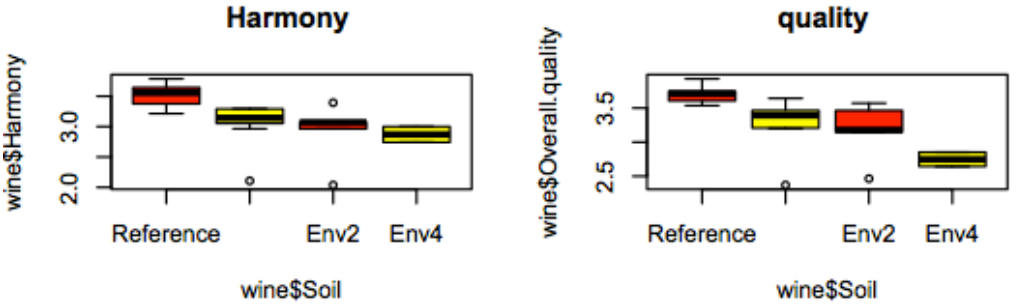
```
rang = c();
for(i in 1:28){  rang[i] <- paste(sd(wine[,i+2]),i+2,sep=" - ")}
vars = sort(rang,decreasing=TRUE)[1:28]
significant = c();
for( i in 1:28)
{
        ind =  strsplit(vars[i]," - ")[[1]][1]
        cu = as.numeric(strsplit(vars[i]," - ")[[1]][2])
        fs[i] = labels(wine)[[2]][cu]
        f = lm(wine[,cu]~wine$Soil)
        modviz = anova(f)
        a = modviz$"Pr(>F)"[1]
        if( a < .05 ){
                significant[i] = paste(fs[i], paste(" is effected significantly. pval = ",a))
        }
}
```

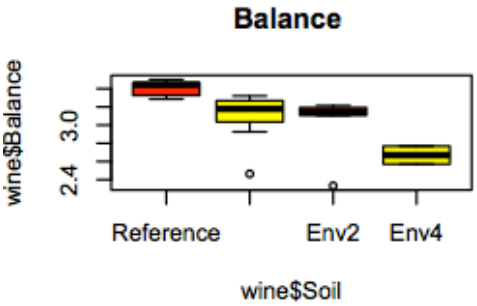
```
significant[which(!is.na(significant))]
 [1] "Harmony  is effected significantly. pval =  0.0415953008413679"
 [2] "Overall.quality  is effected significantly. pval =  0.00911307502339445"
 [3] "Smooth  is effected significantly. pval =  0.0216504889368781"
 [4] "Intensity  is effected significantly. pval =  0.0306419633190303"
 [5] "Balance  is effected significantly. pval =  0.01098007844513 05"
 [6] "Aroma.quality  is effected significantly. pval =  0.0028708710024 1839"
 [7] "Odor.Intensity.before.shaking  is effected significantly. pval =  0.0039685832598679"
 [8] "Aroma.persistency  is effected significantly. pval =  0.0341103467497511"
 [9] "Acidity  is effected significantly. pval =  0.047432946385 4842"
[10] "Spice.before.shaking  is effected significantly. pval =  0.00259948245103811"
[11] "Quality.of.odour  is effected significantly. pval =  0.0134196712205555"
[12] "Aroma.quality.before.shaking  is effected significantly. pval =  0.024751053537769"
[13] "Bitterness  is effected significantly. pval =  0.00176239597794868"
[14] "Plante  is effected significantly. pval =  0.0036336771382 5456"
```

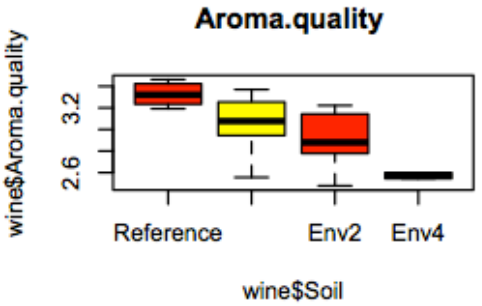
% plot all significant factors

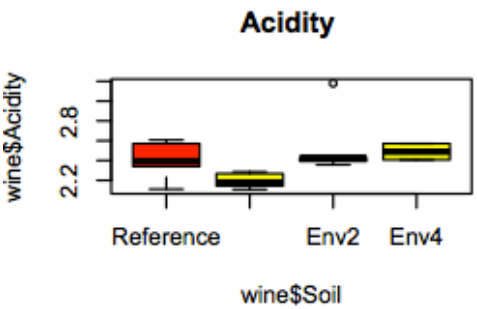
```
par( mfrow = c( 4 , 4 ) )
```

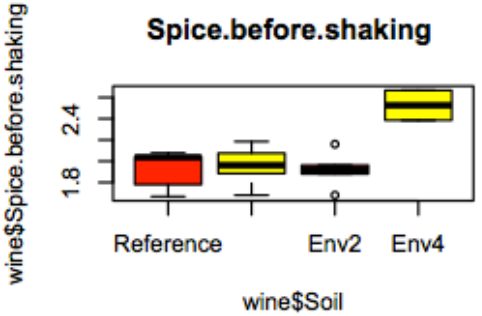
```
plot(wine$Harmony~wine$Soil, main="Harmony ", col=heat.colors(2))
plot(wine$Overall.quality~wine$Soil, main="quality ", col=heat.colors(2))
plot(wine$Smooth~wine$Soil, main="smooth ", col=heat.colors(2))
plot(wine$Intensity~wine$Soil, main="intensity ", col=heat.colors(2))
plot(wine$Balance~wine$Soil, main="Balance ", col=heat.colors(2))
plot(wine$Aroma.quality~wine$Soil, main="Aroma.quality ", col=heat.colors(2))
plot(wine$Odor.Intensity~wine$Soil, main="Odor intensity ", col=heat.colors(2))
plot(wine$Aroma.persistency~wine$Soil, main="Aroma.persistency ", col=heat.colors(2))
plot(wine$Acidity~wine$Soil, main="Acidity ", col=heat.colors(2))
plot(wine$Spice.before.shaking~wine$Soil, main="Spice.before.shaking ", col=heat.colors(2))
plot(wine$Quality.of.odour~wine$Soil, main="Quality.of.odour ", col=heat.colors(2))
plot(wine$Aroma.quality.before.shaking~wine$Soil, main="Aroma.quality.before.shaking ", col=heat.colors(2))
plot(wine$Bitterness~wine$Soil, main="Bitterness ", col=heat.colors(2))
plot(wine$Plante~wine$Soil, main="Plante ", col=heat.colors(2))
```

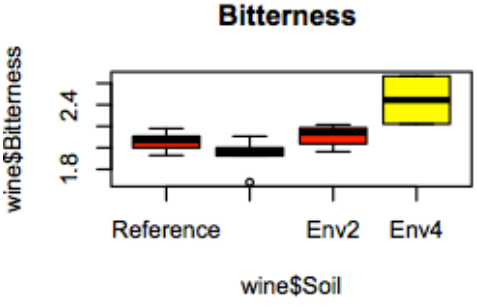


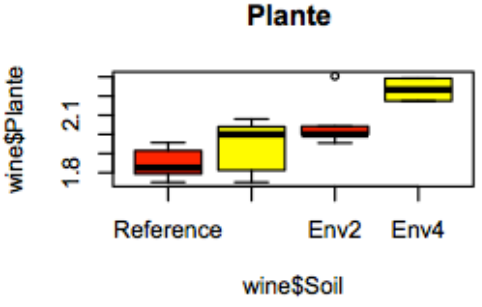


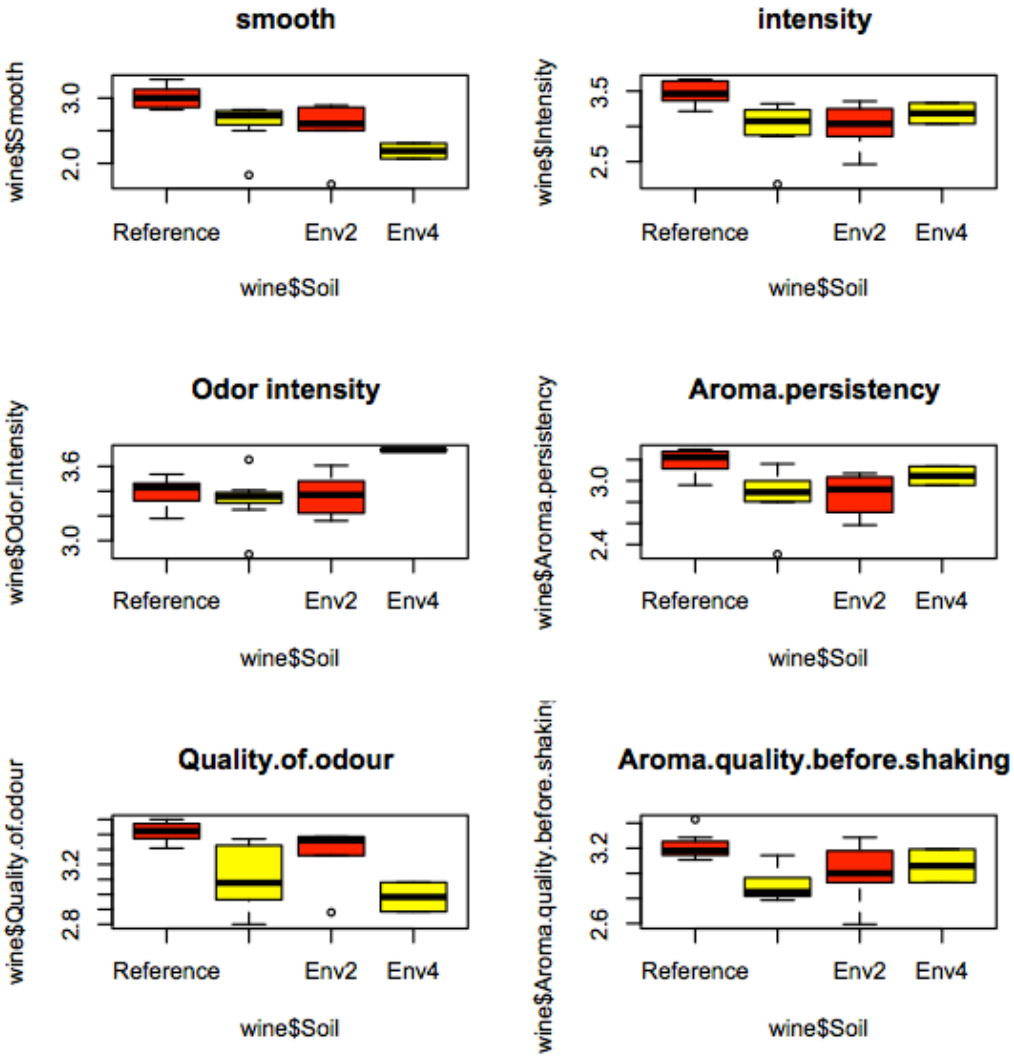












The preceding all have different group means, and illustrate that Soil does affect these wine factors.

Does Label and Soil affect "wine" features?

Check if they affect Overall.quality first:

```
> AnovaModel.2 <- (lm(Overall.quality ~ Label*Soil, data=wine))
> Anova(AnovaModel.2)
Anova Table (Type II tests)
Response: Overall.quality
              Sum Sq  Df  F value   Pr(>F)
Label         0.17182  2   0.5988  0.56509
Soil          1.79394  3   4.1680  0.03077 *
Label:Soil    0.03342  3   0.0777  0.97089
Residuals     1.72163 12
```

As before we observe that Soil has an affect on Overall.quality, but Label's main effect nor its interaction with Soil is significant.

```
> tapply(wine$Overall.quality, list(Label=wine$Label, Soil=wine$Soil), mean, na.rm=TRUE) # means
          Soil
Label                Reference     Env1              Env2          Env4
  Saumur         3.747000          3.333333          3.299333          2.7475
  Bourgueuil     3.702333     3.183000                    NA          NA
  Chinon         3.536000          3.200000          2.964000          NA

> tapply(wine$Overall.quality, list(Label=wine$Label, Soil=wine$Soil), sd, na.rm=TRUE) # std. deviations
                     Soil
Label          Reference      Env1          Env2          Env4
  Saumur       0.17907540  0.1033457   0.2357803   0.1477853
  Bourgueuil   0.05444569  0.7061083      NA             NA
  Chinon          NA             NA             0.7071068      NA

> tapply(wine$Overall.quality, list(Label=wine$Label, Soil=wine$Soil), function(x) sum(!is.na(x))) # counts
                     Soil
Label                Reference     Env1    Env2    Env4
  Saumur          3              3             3              2
  Bourgueuil      3              3             NA            NA
  Chinon          1              1             2             NA
```

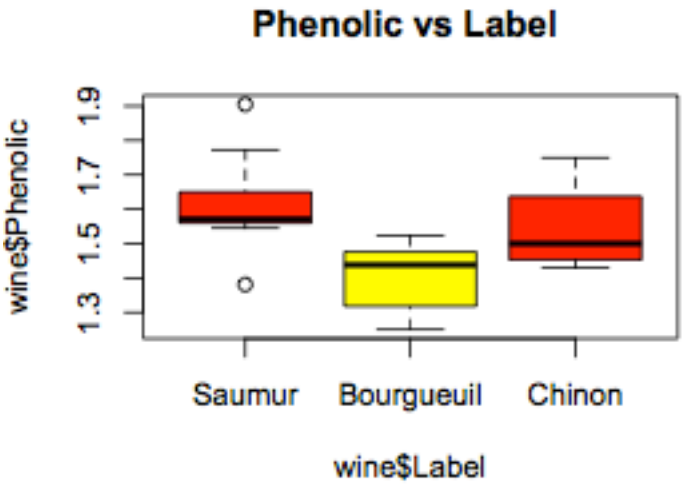
%Find all significant effects of Label or its interactions with Soil in predicting wine factors.

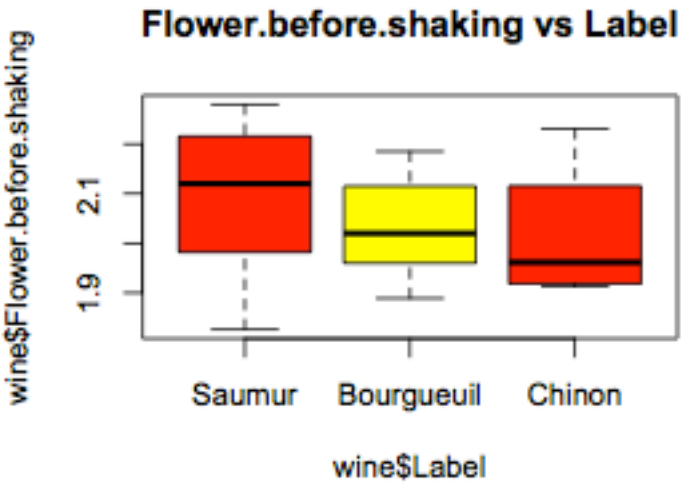
```
rang = c();
for(i in 1:28){  rang[i] <- paste(sd(wine[,i+2]),i+2,sep=" - ")}
vars = sort(rang,decreasing=TRUE)[1:28]
manova_significant = c();
for( i in 1:28)
{
        ind =  strsplit(vars[i]," - ")[[1]][1]
        cu = as.numeric(strsplit(vars[i]," - ")[[1]][2])
        fs[i] = labels(wine)[[2]][cu]
        AnovaModel <- (lm(wine[,cu] ~ wine$Label*wine$Soil, data=wine))
        sigs = Anova(AnovaModel)$"Pr(>F)"
        t = ""
        if(sigs[1] < .05){
                t = paste("Label main effect on", paste(fs[i], paste(" is effected significantly. pval = ",sigs[1])))
        }

        if(sigs[3] < .05){
                t = paste(t,paste("Label*Soil interaction effect on",
                                paste(fs[i], paste(" is effected significantly. pval = ",sigs[3]))))
        }
        manova_significant[i] = t
}
manova_significant[which(manova_significant != "")]
```

```
> manova_significant[which(manova_significant != "")]
[1] "Label main effect on Phenolic  is effected significantly. pval =  0.0366129957575019"
[2] "Label*Soil interaction effect on Flower  is effected significantly. pval =  0.0491181242450049"
[3] "Label main effect on Flower.before.shaking  is effected significantly. pval =  0.0331305749482079"
```

The only factor that the interaction between Label*Soil may have an effect on is on Flower, but its very close to the 5 % percent cutoff and is thus not completely satisfactory. Label has an effect on Phenolic and Flower.before.shaking and can be seen by the following:

```
par( mfrow = c( 1 , 2 ) )
plot(wine$Phenolic~wine$Label, main="Phenolic vs Label ", col=heat.colors(2))
plot(wine$Flower.before.shaking~wine$Label, main="Flower.before.shaking vs Label", col=heat.colors(2))
```

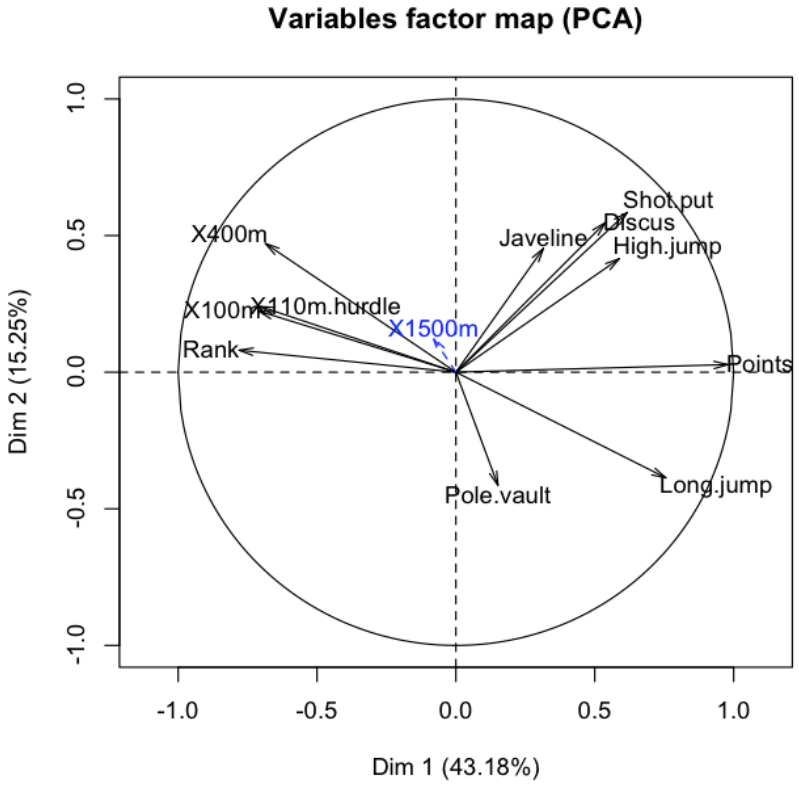




Thus generally speaking Label has little to no effect on wine features.

Second Exercise:
Predict the behavior of an athlete in the 1500 m

What is the expression that better predicts the behavior of an athlete for 1500m?
We first look at the PCA created with 1500m as a supplementary variable.



There are no clear factors which predict 1500m exactly, but X400 seems to do a good job, with X110m.hurdle, X100m, and Javaline being the next closest on the positive correlated side, whereas Pole.value seems to be a good indicator on the negative side. Collectively with all the factors included, they only explain 43+15 = 58 % percent of the variability of X1500m. Thus a possible expression to predict the behavior of on an athlete for 1500m would be:

```
> f = lm(X1500m ~ X400m + X110m.hurdle +  X100m + Javeline + Pole.vault )
> summary(f)
Call: lm(formula = X1500m ~ X400m + X110m.hurdle + X100m + Javeline +  Pole.vault)
Residuals:
    Min      1Q  Median      3Q     Max
-19.0683  -5.0345  -0.5956   3.6932  23.9385

Coefficients:
              Estimate     Std. Error   t value Pr(>|t|)
(Intercept)   134.8253     85.5776      1.575   0.124144
X400m         6.9497       1.6597       4.187   0.000181 ***
X110m.hurdle  -3.1806      4.2693       -0.745  0.461243
X100m         -15.8796     7.6383       -2.079  0.045015 *
Javeline      -0.5578      0.3230       -1.727  0.092986 .
Pole.vault    11.1315      5.5238       2.015   0.051616 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.632 on 35 degrees of freedom
Multiple R-squared:  0.4043,  Adjusted R-squared:  0.3192
F-statistic: 4.751 on 5 and 35 DF,  p-value: 0.00204
```

Thus it turns out that X400m and X100m as main effects are significant and Pole.vault could be good as well so we also check to see if any interactions produce interesting results. This equation however only represents 40.43% of the variability of X1500m so a predictor based on it will not be perfect.

```
> f = lm(X1500m ~ X400m + X110m.hurdle +  X100m + Javeline + Pole.vault + X400m*X100m +
                  X400m*X110m.hurdle + X400m*Pole.vault + X100m*X110m.hurdle)
> anova(f)
Analysis of Variance Table
Response: X1500m
                      Df  Sum Sq  Mean Sq  F value  Pr(>F)
X400m                 1   907.80  907.80   11.1655  0.002185 **
X110m.hurdle          1   269.78  269.78   3.3181   0.078183 .
X100m                 1   340.71  340.71   4.1905   0.049207 *
Javeline              1   308.64  308.64   3.7961   0.060475 .
Pole.vault            1   376.74  376.74   4.6337   0.039247 *
X400m:X100m           1   38.03   38.03    0.4677   0.499110
X400m:X110m.hurdle    1   659.97  659.97   8.1173   0.007722 **
X400m:Pole.vault      1   1.05    1.05     0.0129   0.910187
X110m.hurdle:X100m    1   27.44   27.44    0.3376   0.565444
Residuals             31  2520.43 81.30
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Explore different expressions describing the power and the features of each one of them. Justify your final decision.

Now we see that in addition to X400m, X100m, and Pole.vault, the interaction X400m:X110.hurdle is significant. This implies that we can use the following generalized linear model to predict the behavoir of X1500m:
lm(X1500m ~ X400m + X100m + Pole.vault + X400m*X110m.hurdle)

Predict the behavior of an athlete.

Now use the expression to predict the behavior for a specific athlete. Use the data contained in the table. Example, if you have a model that only uses X400m as a variable you can construct a new dataframe: *new <- data.frame(X400m=48)*
And then use it to predict: *predict(LinearModel.3, newdata=new, interval="prediction")*
or predict(LinearModel.3, newdata=new, interval="confidence")

Analyze and explain the results obtained. Is the model accurate? What do you expect?

From before our model is:
f = lm(X1500m ~ X400m + X100m + Pole.vault + X400m*X110m.hurdle)
Now I want to construct an example to see what it will predict for X1500m
I first construct a dataframe using the means of the values that I use in my value. Here i could construct the model using whatever values, but it seemed easier to do this for a first run as we would then expect the predicted value of X1500m to be near its mean value in the data.

```
> mean(Pole.vault)          ->  4.762439
> mean(X110m.hurdle)        -> 14.60585
> mean(X100m)               -> 10.99805
> mean(X400m)               -> 49.61634
> new <- data.frame(X400m=49.6,X100m=11,Pole.vault=4.76, X110m.hurdle=14.6)
> predict(f,newdata = new, interval="prediction")
        fit       lwr       upr
1   277.237  258.0475  296.4265
```

Thus the fitted predicted value of what X1500m will be is 277.237 with a lower bound of 258.0475 and a higher bound of 296.426 using a predication interval of 95%. As the mean for X1500m from the decatholon data is mean(X1500m) -> 279.0249 this seems reasonable.

To get a confidence interval we run:

```
>predict(f,newdata = new, interval="confidence")
        fit       lwr       upr
1   277.237  273.9637  280.5103
```

Our Confidence Interval is then 277.237 ± 3.2733 which gives a lowerbound of 273.96 and an upperbound of 280.51. This interval overall is closer around the mean than the prediction interval as we would expect, and represents with 95% certain where we expect the population parameter (X1500m) to fall.

To calculate a tolerance interval, ie, limits taken from the estimated interval within which a stated proportion of the population (here, X1500m) is expected to occur, we use the tolerance library as follows:

```
> library(tolerance)
> normtol.int(x=X1500m, alpha= 0.05, P=0.95, side=2)
     alpha    P      x.bar          2-sided.lower   2-sided.upper
1    0.05   0.95   279.0249       250.5338              307.516
```

This results in a 95% tolerance interval for 95% of data of this type, based on the 41 observations (rows) of the decathlon data set. It says that with 95% confidence, 95% of the data will exist within interval between 250.53 and 307.52 around the mean of 279.02.

Third Exercise:

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, we must detect those more important and then define a DOE to conduct the experiments.

our data:

UN 2013 Development report data
for 50 highest ranked HDI (human development index) countries

available at http://hdr.undp.org/opendata/

1) Preprocessing data first to make it clean:

```
p3d <-read.csv("designofexperiments/doe-assignment2-part3-data.csv",na.strings="..",stringsAsFactors=FALSE)
dim(p3d)
 % 50 rows by 52 columns with na's
```

% fix columns which are numeric but are read in as strings

-> just had to change from Number to General in Excel

%remove columns with most NA's so that we have more experiments than factors

```
> nas = c()
> for(i in 1:52){ nas[i] = length(which(is.na(p3d[,i])))}
> nas
[1] 0 0 0 0 0 0 1 1 2 2 4 3 11 0 0 0 0 0 0 0 13 13 13 3 3 13 13 13 13 22 22 6 6 4 0 2 2 2 3 3 2 2 2 2 3 2
2 2 16 2 6 5
```

```
>  nas[30]  ->   22
> nas[31]   ->   22
> nas[49]   ->   16
> nas[c(21:23,26:29)]  -> 13 13 13 13 13 13 13
```

%these factors correspond to mainly "inequality adjusted measures"

```
> p3d <- p3d[ , c(1:20,24,25,32:48,50:52)]
> dim(p3d)
[1] 50 42
```

%double check columns

```
> nnas = c()
> for(i in 1:42){ nnas[i] = length(which(is.na(p3d[,i])))}
> nnas
 [1] 0 0 0 0 0 0 1 1 2 2 4 3 11 0 0 0 0 0 0 0 3 3 6 6 4 0 2 2 2 3 3 2 2 2 2 3 2 2 2 2 6 5
```

%double check, that no rows have high na's

```
> rnas = c()
> for(i in 1:50){ rnas[i] = length(which(is.na(p3d[i,])))}
> rnas
[1] 0 0 0 0 0 0 0 0 0 0 0 0 14 0 0 0 0 0 0 0 0 0 0 24 0 1 0 0 0 4 0 0 17 0 1 2 0 3 0 0 0 2 0 0 0 9 1 1 5 4
```

% need to remove index 13, 24, 33, 46, which correspond to the countries which provided the least amount of data to the UN

```
> p3d[c(13,24,33,46),][1]
              Countries
13 Hong Kong, China (SAR)
24          Liechtenstein
33                Andorra
46             Seychelles

> p3d <- p3d[ c(1:12,14:23,25:32,34:45,47:50),]
> dim(p3d)
[1] 46 42
```

%now fill in nas with reasonable values for each column by using mean for that column.

```
cnas = c()
for(i in 1:42){ cnas[i] = length(which(is.na(p3d[,i])))}
cnas
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 4 0 0 0 1 2 0 3 0 0 0 2 0 0 0
> which(cnas > 0)
[1] 12 13 23 24 27 28 29 36 40 41 42

for(i in c(12,13,23,24,27,28,29,36,40,41,42))
{
        tmp_mean = mean(p3d[,i],na.rm=TRUE)
        rrs = c()
        for(j in 1:46){
                if(is.na(p3d[j,i])){
                        p3d[j,i] = tmp_mean
                }
        }
}
```

```
%final verify of no NAs
fnas = c()
for(i in 1:46){ fnas[i] = length(which(is.na(p3d[i,])))}
fnas
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

%save file to cleanupversion
write.csv(p3d,"designofexperiments/doe-assignment2-part3-cleanedup-data.csv")

> str(p3d)
'data.frame':   46 obs. of  42 variables:
 $ Countries       : chr  "Norway" "Australia" "United States" "Netherlands" ...
 $ HDI.rank        : int  1 2 3 4 5 6 7 7 9 10 ...
 $ IMMUNIZATION.COVERAGE..DTP...of.one.year.olds..2010         : num  99 97 99 99 97 95 98 99 ...
 $ IMMUNIZATION.COVERAGE.Measles...of.one.year.olds..2010     : num  93 94 92 96 96 91 90  ...
 $ HIV.PREVALENCE..YOUTH.Female....ages.15.24..2009     : num  0.1 0.1 0.2 0.1 0.1 0.1 0.1 0.1  ...
 $ HIV.PREVALENCE..YOUTH.Male...ages.15.24..2009         : num  0.1 0.1 0.3 0.1 0.1 0.1 0.1  ...
 $ MORTALITY.RATES.Infant.2010..deaths.per.1.000.live.births.       : int  3 4 7 4 3 5 3 2 4 2 ...
 $ MORTALITY.RATES.Under.five.2010.deaths.per.1.000.live.births.           : int  3 5 8 4 4 6 4 3 5 3 ...
 $ MORTALITY.RATES.Adult.Female.2009..per.1.000.adults.   : int  50 45 78 56 53 57 57 47 43 42 ...
 $ MORTALITY.RATES.ADULT.MALES.2009..per.1.000.adults.       : int  83 79 134 75 99 86 97 74 74 ...
 $ MORTALITY.RATES...cardiovascular.diseases.and.diabetes.2008.per.1.000.people.    : int  124 112  ...
 $ HEALTH.CARE.QUALITY.Physicians.2005.10..per.1.000.people.   : num  4.1 3 2.7 3.9 3.5 2.4 3.2 ...
 $ Satisfaction.with.health.care.quality.2007.9....satisfied.    : num  68 60 56 77 47 64 47 81 81 54 ...
 $ Human.Development.Index..HDI..2012    : num  0.955 0.938 0.937 0.921 0.92 0.919 0.916 0.916  ...
 $ Life.expectancy.at.birth..years.2012     : num  81.3 82 78.7 80.8 80.6 80.8 80.7 81.6 82.5 83.6 ...
 $ Mean.years.of.schooling..years.2010   : num  12.6 12 13.3 11.6 12.2 12.5 11.6 11.7 11 11.6 ...
 $ Expected.years.of.schooling.2011..years.       : num  17.5 19.6 16.8 16.9 16.4 19.7 18.3 16 15.7 15.3 ...
 $ Gross.national.income..GNI..per.capita..2012    : num  48688 34340 43480 37282 35431 ...
 $ GNI.per.capita.rank.minus.HDI.rank.2012      : int  4 15 6 8 10 26 19 6 2 11 ...
 $ Nonincome.HDI.2012       : num  0.977 0.978 0.958 0.945 0.948 0.978 0.96 0.94 0.926 0.942 ...
 $ Inequality.adjusted.life.expectancy.index.Value        : num  0.928 0.93 0.863 0.916 0.915  ...
 $ Inequality.adjusted.life.expectancy.index.Loss.....2012     : num  3.7 4.7 6.6 4.3 4 5.2 4.3 3.3 4.1 3.5 ...
 $ Gender.Inequality.Index.Rank.2012      : num  5 17 42 1 6 31 19 2 3 21 ...
 $ Gender.Inequality.Index.Value          : num  0.065 0.115 0.256 0.045 0.075 0.164 0.121  ...
 $ Maternal.mortality.ratio..deaths.per.100.000.live.births..2010     : int  7 7 21 6 7 15 6 4 8 5 ...
 $ Adolescent.fertility.rate..births.per.1.000.women.ages.15.19..2012     : num  7.4 12.5 27.4 4.3  ...
 $ Seats.in.national.parliament....female..2012      : num  39.6 29.2 17 37.8 32.4 32.2 19 44.7 26.8 13.4 ...
 $ Population.with.at.least.secondary.education....ages.25.and.older..Female.2006.10: num  95.6 92.2  ...
 $ Population.with.at.least.secondary.education....ages.25.and.older..Male.2006.10  : num  94.7 92.2 ...
 $ Labour.force.participation.rate....ages.15.and.older..female.2011     : num  61.7 58.8 57.5 ...
 $ Labour.force.participation.rate....ages.15.and.older..Male.2011     : num  70.1 72.3 70.1 71.7 ...
 $ GDP.2011      : num  233 782 13238 622 2814 ...
 $ GDP.per.capita.2011      : num  46982 34548 42486 37251 34437 ...
 $ Gross.fixed.capital.formation....of.GDP..2011     : num  20.2 27.1 14.7 18.6 18.2 18.9 11.5  ...
 $ Consumer.Price.Index..2005...100..2010       : int  112 116 112 108 108 115 107 108 104 100 ...
 $ General.government.final.consumption.expenditure.2000....of.GDP.   : num  19.3 17.6 14.3 22  ...
 $ General.government.final.consumption.expenditure.2011....of.GDP.    : num  21.5 18 17.5 28.1  ...
 $ Health...of.GDP.2000    : num  6.4 5.4 5.8 5 8.2 6 4.6 7 5.6 6.2 ...
 $ Health...of.GDP.2010    : num  8 5.9 9.5 9.4 9 8.4 6.4 7.8 6.8 7.8 ...
 $ Education...of.GDP.2005.10        : num  7.3 5.1 5.4 5.9 4.6 7.2 5.7 7.3 5.4 3.8 ...
 $ Military...of.GDP.2000     : num  1.7 1.8 3 1.5 1.5 1.2 0.7 2 1.1 1 ...
 $ Military....of.GDP.2010         : num  1.5 1.9 4.8 1.4 1.4 1.2 0.6 1.3 0.8 1 ...
```

Find a problem that could be solved using the DOE methodology and:

1. Set the objectives.
2. Select the process variables.
3. Define an experimental design.
4. Execute the design.
5. Check that the data are consistent with the experimental assumptions.
6. Analyze and interpret the results, detect effects of main factors and interactions.

1. I would like to study what are the factors which best predict 2011 GDP
I first run a Principal Components Analysis with 2011 GDP as my supplmentary variable, but because there are so many factors its difficult to look at the full graph and tell which factors are most important. In looking at the eigenvalues for the components I can see that first ten components explain the variance of the data.

```
> p3d.PCA
<-p3d[, c("HDI.rank", "IMMUNIZATION.COVERAGE..DTP...of.one.year.olds..2010",
 "IMMUNIZATION.COVERAGE.Measles...of.one.year.olds..2010", +   "HIV.PREVALENCE..YOUTH.Female....ages.15.24..2009",
"HIV.PREVALENCE..YOUTH.Male...ages.15.24..2009", +   "MORTALITY.RATES.Infant.2010..deaths.per.1.000.live.births.",
"MORTALITY.RATES.Under.five.2010.deaths.per.1.000.live.births.",
+   "MORTALITY.RATES.Adult.Female.2009..per.1.000.adults.", "MORTALITY.RATES.ADULT.MALES.2009..per.1.000.adults.",
+   "MORTALITY.RATES...cardiovascular.diseases.and.diabetes.2008.per.1.000.people.",
+ "HEALTH.CARE.QUALITY.Physicians.2005.10..per.1.000.people.",
+   "Satisfaction.with.health.care.quality.2007.9....satisfied.", "Human.Development.Index..HDI..2012",
+   "Life.expectancy.at.birth..years.2012",
+   "Mean.years.of.schooling..years.2010", "Expected.years.of.schooling.2011..years.",
+   "Gross.national.income..GNI..per.capita..2012",
+   "GNI.per.capita.rank.minus.HDI.rank.2012", "Nonincome.HDI.2012", "Inequality.adjusted.life.expectancy.index.Value",
+   "Inequality.adjusted.life.expectancy.index.Loss.....2012", "Gender.Inequality.Index.Rank.2012",
"Gender.Inequality.Index.Value",
+   "Maternal.mortality.ratio..deaths.per.100.000.live.births..2010",
"Adolescent.fertility.rate..births.per.1.000.women.ages.15.19..2012",
+   "Seats.in.national.parliament....female..2012",
"Population.with.at.least.secondary.education....ages.25.and.older..Female.2006.10",
+   "Population.with.at.least.secondary.education....ages.25.and.older..Male.2006.10",
"Labour.force.participation.rate....ages.15.and.older..female.2011",
+   "Labour.force.participation.rate....ages.15.and.older..Male.2011", "GDP.per.capita.2011",
"Gross.fixed.capital.formation....of.GDP..2011",
+   "Consumer.Price.Index..2005...100..2010", "General.government.final.consumption.expenditure.2000....of.GDP.",
+   "General.government.final.consumption.expenditure.2011....of.GDP.", "Health...of.GDP.2000",
"Health...of.GDP.2010", "Education...of.GDP.2005.10",
+   "Military...of.GDP.2000", "Military....of.GDP.2010", "GDP.2011")]

> res<-PCA(p3d.PCA , scale.unit=TRUE, ncp=5, quanti.sup=c(41: 41), graph = FALSE)
> plot.PCA(res, axes=c(1, 2), choix="var", new.plot=TRUE, col.var="black", col.quanti.sup="blue", label=c("var",
"quanti.sup"), lim.cos2.var=0)

> res$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1  1.334989e+01       3.337471e+01                          33.37471
comp 2  6.061375e+00       1.515344e+01                          48.52815
comp 3  4.105395e+00       1.026349e+01                          58.79164
comp 4  2.586817e+00       6.467044e+00                          65.25868
comp 5  2.464693e+00       6.161733e+00                          71.42042
comp 6  1.779869e+00       4.449672e+00                          75.87009
comp 7  1.667025e+00       4.167563e+00                          80.03765
comp 8  1.411577e+00       3.528942e+00                          83.56659
```

comp 9 1.006419e+00 2.516047e+00 86.08264
comp 10 9.276533e-01 2.319133e+00 88.40177
….
comp 39 8.166555e-05 2.041639e-04 99.99998
comp 40 6.587630e-06 1.646908e-05 100.00000

I run the PCA with less factors to visually see which ones are more positively/negatively correlated with GDP.2011. I also discard columns which are rankings as they are not in reality numeric, but instead ordinal, and come up with the following model.

```
> sssm = lm( p3d$"GDP.2011" ~
+ p3d$"MORTALITY.RATES.ADULT.MALES.2009..per.1.000.adults."+
+ p3d$"MORTALITY.RATES...cardiovascular.diseases.and.diabetes.2008.per.1.000.people."+
+ p3d$"IMMUNIZATION.COVERAGE..DTP...of.one.year.olds..2010"+
+ p3d$"Mean.years.of.schooling..years.2010"+
+ p3d$"Expected.years.of.schooling.2011..years."+
+ p3d$"Adolescent.fertility.rate..births.per.1.000.women.ages.15.19..2012"+
+ p3d$"Population.with.at.least.secondary.education....ages.25.and.older..Female.2006.10" +
+ p3d$"Population.with.at.least.secondary.education....ages.25.and.older..Male.2006.10"+
+ p3d$"Labour.force.participation.rate....ages.15.and.older..Male.2011"+
+ p3d$"General.government.final.consumption.expenditure.2011....of.GDP."+
+ p3d$"Military....of.GDP.2010")
> summary(sssm)

Residuals:
   Min     1Q  Median     3Q    Max
-3891.4  -846.8   -96.5   633.6  6900.1

Coefficients:
                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                     -45358.621  17087.7  -2.654   0.0120 *
p3d$MORTALITY.RATES.ADULT.MALES.2009..per.1.000.adults.             16.030     12.068   1.328   0.1929
p3d$MORTALITY.RATES...cardiovascular.diseases.and.diabetes.2008..  -15.374      6.857   -2.242   0.0316 *
p3d$IMMUNIZATION.COVERAGE..DTP...of.one.year.olds..2010            365.788    152.484   2.399   0.0221 *
p3d$Mean.years.of.schooling..years.2010                          977.453     352.761   2.771   0.0090 **
p3d$Expected.years.of.schooling.2011..years.                    -257.223     248.151  -1.037   0.3073
p3d$Adolescent.fertility.rate..births.per.1.000.women.ages.15.19..2012  51.748   30.408   1.702   0.0979 .
p3d$Population.with.at.least.secondary.education....ages.25.and.older..Female.  -19.375   69.803   -0.278   0.7830
p3d$Population.with.at.least.secondary.education....ages.25.and.older..Male.    9.047   75.123   0.120   0.9049
p3d$Labour.force.participation.rate....ages.15.and.older..Male.2011   48.744   58.425   0.834   0.4099
p3d$General.government.final.consumption.expenditure.2011....of.GDP.  39.906   70.583   0.565   0.5755
p3d$Military....of.GDP.2010                                     521.635     228.189   2.286   0.0286 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1744 on 34 degrees of freedom
Multiple R-squared:  0.4476,        Adjusted R-squared:  0.2688
F-statistic: 2.504 on 11 and 34 DF,  p-value: 0.01986
```

I then select the most significant of these factors,
"MORTALITY.RATES...cardiovascular.diseases.and.diabetes.2008.per.1.000.people."
"IMMUNIZATION.COVERAGE..DTP...of.one.year.olds..2010"
"Mean.years.of.schooling..2010"
"Military% of.GDP.2010"
and design an experiment to see which of these factors or interactions of these would lead to increases in GDP.2011. This is of course an unrealistic experiment, as I can't in actuality change what the GDP for a given year is, nor what the social factors measured were, but as a hypothetical situation it is of interest. For the purpose of the experiment we define:

a = "MORTALITY.RATES...cardiovascular.diseases.and.diabetes.2008.per.1.000.people."
b = "IMMUNIZATION.COVERAGE..DTP...of.one.year.olds..2010"
c = "Mean.years.of.schooling..2010"
d = "Military% of.GDP.2010"

We then get the low (-) and high (+) values for each factor as follows:

```
> min(p3d$MORTALITY.RATES.ADULT.MALES.2009..per.1.000.adults.)    -> 65
> max(p3d$MORTALITY.RATES.ADULT.MALES.2009..per.1.000.adults.)  -> 324

> min(p3d$IMMUNIZATION.COVERAGE..DTP...of.one.year.olds..2010 ) ->  92
> max(p3d$IMMUNIZATION.COVERAGE..DTP...of.one.year.olds..2010 ) -> 99

> min(p3d$Mean.years.of.schooling..years.2010)  -> 7.3
> max(p3d$Mean.years.of.schooling..years.2010)  -> 13.3

> min(p3d$Military....of.GDP.2010) -> 0.1
> max(p3d$Military....of.GDP.2010) ->  6.9
```

Additionally as we can't actually measure what these combinations would actually produce we need to select response values for GDP.2011 from the existing values for it.

```
> summary(p3d$GDP.2011)
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
   4.791    58.350   229.700   806.100   629.400 13240.000
```

Plotting a histogram of it shows that the data is not normally distributed

```
>histogram(p3d$GDP.2011, breaks=20)
```

We calculate and find in fact the data is distributed according to a lognormal function

```
> library(MASS)
> paramslognrom = fitdistr(p3d$GDP.2011, "lognormal", lower = 0.001)
> paramslognrom
    meanlog       sdlog
  5.2861813   1.7369850
> ks.test(p3d$GDP.2011, "plnorm", meanlog=paramslognrom$estimate[1],  sdlog=paramslognrom$estimate[2])

        One-sample Kolmogorov-Smirnov test
        data:  p3d$GDP.2011
        D = 0.0828, p-value = 0.8851
        alternative hypothesis: two-sided
```

Thus we select 16 random elements from the distribution and place them in our table:

```
> gdpdistrib = rlnorm(1000,meanlog=5.286,  sdlog=1.7369850)
> round(sample(gdpdistrib, 16, replace=F),2)
 [1]   227.84   825.57   959.26   167.16   142.22  2238.51    27.59   136.98   161.99    35.63  1630.48
33.73   163.88   101.66 11562.35     9.48
```

exp	a	b	c	d	Response
-	-	-	-	-	227.84
a	+	-	-	-	825.57
b	-	+	-	-	959.26
ab	+	+	-	-	167.16

c	-	-	+	-	142.22
ac	+	-	+	-	2238.51
bc	-	+	+	-	27.59
abc	+	+	+	-	136.98
d	-	-	-	+	161.99
ad	+	-	-	+	35.63
bd	-	+	-	+	1630.48
abd	+	+	-	+	33.73
cd	-	-	+	+	163.88
acd	+	-	+	+	101.66
bcd	-	+	+	+	11562.35
abcd	+	+	+	+	9.48

Then according to Yates algorithm, we add four columns to the right for each replication, one to hold the divider necessary and then a final column to measure the main effects and interactions

obs	rep1	rep2	rep3	rep4	div	Response	
227.84	1053.4	2179.8	4725.1	18424	16	1151.52	mean
825.57	1126.4	2545.3	13699	-11327	8	-1415.86	a
959.26	2380.7	1861.8	2011.3	10630	8	1328.72	b
167.16	164.57	11837	-13338	-16338	8	-2042.22	ab
142.22	197.62	-194.4	-2143	10341	8	1292.63	c
2238.5	1664.2	2205.7	12773	-7492	8	-936.49	ac
27.59	265.54	-1723	-3377	7550.5	8	943.82	bc
136.98	11572	-11615	-12961	-10617	8	-1327.17	abc
161.99	597.73	73.01	365.47	8974.1	8	1121.76	d
35.63	-792.1	-2216	9975.5	-15350	8	-1918.69	ad
1630.5	2096.3	1466.6	2400.1	14916	8	1864.50	bd
33.73	109.39	11306	-9892	-9584	8	-1198.04	abd
163.88	-126.4	-1390	-2289	9610.1	8	1201.26	cd
101.66	-1597	-1987	9839.7	-12292	8	-1536.50	acd
11562	-62.22	-1470	-597.1	12129	8	1516.11	bcd
9.48	-11553	-11491	-10020	-9423	8	-1177.90	abcd

According to our data, then factors/interactions with the greatest effects are: b:d with 1864.50, b:c:d with 1516.11, b with 1328.72, and c with 1292.63.

Thus in this particular case the interactions between b "IMMUNIZATION.COVERAGE..DTP…of.one.year.olds..2010" **and c** "Mean.years.of.schooling..2010" **and d** "Military% of.GDP.2010" **have the greatest effect followed by the main effects of b and c.**