SMDE Second assignment

DIEGO CAMPOS MILIAN

Define your RNG

1. **Implement an RNG.**

Congruential RNG

```
congRNG <- function(mult, inc, seed, mod, num){

  aux = 0.0
  output = 0.0
  result = numeric()

  for ( i in 1:num){
    aux = (mult*seed+inc) %% mod
    output = aux/mod
    result[i] = output
    seed = aux
  }

  return(result)
}
```

Simulate your data.

1. **Define, for each factor (from 1 to 5) a distribution (the RVGs that you prefer, uniform, normal, exponential, etc.). For the factors 6 to 10 define a function that uses the previous variables, as an example F6=F1+2F3.**
2. **Define an answer variable that will be composed by a function that combines a subset of the previous factors plus a normal distribution you know (to add some random noise**).

Factor	Distribution	Upper value	Lower value	Factor	Combination
F1	Uniform	250	80	F6	2*F1
F2	Exponential	100	0	F7	F1 + 3*F2
F3	Power 10	50	10	F8	F4 + F5
F4	Logarithmic	180	20	F9	F3 - F5
F5	Logistic	80	5	F10	F4 - F2
				Answer	F6 +F7 + F4 +F8 +F9

SMDE Second assignment

DIEGO CAMPOS MILIAN

OBTAIN AN EXPRESSION TO GENERATE NEW DATA.

1. **Explore the possible relations of all the factors and the answer variable, you can use any technique developed during the course.**
2. **Describe what you find on this analysis and, explain if it is coherent with the knowledge you have from the data.**
3. **Propose an expression (using a LRM) to generate new data. This is the method that you are going to use to generate new values using a subset of the factors, for a more complex dataset one can use other approaches like Simulation.**

```
Loadings:
       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11
F1      0.327  0.145 -0.208  0.416 -0.135  0.304  0.204          0.709
F2     -0.324  0.246         -0.136 -0.256  0.310 -0.292  0.734         -0.102  -0.107
F3      0.330 -0.192          0.275  0.270         -0.837
F4      0.304  0.309 -0.261 -0.114 -0.468 -0.586 -0.178                  -0.107  -0.348
F5      0.335                -0.422  0.190  0.278         -0.106         -0.756
F6      0.327  0.145 -0.208  0.416 -0.135  0.304  0.204  0.114 -0.700
F7     -0.314  0.308 -0.113         -0.318  0.410 -0.299 -0.640
F8      0.329  0.191         -0.298                       0.152          0.248   0.809
F9     -0.334 -0.114          0.459 -0.185 -0.289                        -0.579   0.450
F10            0.489  0.835  0.221
Answer -0.204  0.610 -0.343  0.114  0.652 -0.176
```

For determining the relations between factors and the answer variable I have used a PCA. In the resulting loading table we can see in the second component, where Answer is the main variable, seems that is mainly correlated to F4, F7 and F10, and a little correlated with F2. Considering that we known that the main factors are F1, F2 and F4, seems coherent to have F4, F7 and F10 as the main factors, knowing that F7 and F10 are mixups of F1, F2 and F4. So I can conclude that the result of the analysis is more or less correct.

For the linear regression model that I used to generate the data I decided to go with the expression with the highest R-squared value. The final result has shown that almost any combination that have F1, F2 and F4 will have always and R-squared over 0.95, what is almost perfect. So, the final expression choosed is the follow one that also incorporates F3.

```
lm(formula = Answer ~ F1 + F2 + F4 + F3, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-92.206 -23.442   0.438  20.458  98.719

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.3123   644.0704   0.176    0.860
F1            2.8381     0.4163   6.817 1.61e-11 ***
F2            2.9211     0.3696   7.903 7.21e-15 ***
F4            1.1260     0.2191   5.140 3.31e-07 ***
F3           -0.9079    14.1803  -0.064    0.949
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.02 on 995 degrees of freedom
Multiple R-squared:  0.9726, Adjusted R-squared:  0.9725
F-statistic:  8834 on 4 and 995 DF,  p-value: < 2.2e-16
```

SMDE Second assignment

DIEGO CAMPOS MILIAN

DOE

1. **Define a DOE to explore with what parametrization of the 10 factors the answer obtains the best value (define what means best, i.e. maximize or minimize the value).**
2. **Detect and analyze the interactions.**

Objective: So the goal will be to determine the main factors that maximize the Answer variable.

Process variable: So for the experiment I have reduced the variables to process to only the ones that form the LRM explained before, F1, F2, F3 and F4 .

Considering that we are going to use an LRM we have to check for homoscedasticity, normality and independence. As the results of the test shown, all p-values are above 0.05, does not seem to be heteroscedasticity, nonnormality or dependence between the variables.

```
> lmtest::dwtest(LM, alternative = "two.sided")

 Durbin-Watson test

data:  LM
DW = 2, p-value = 0.9996
alternative hypothesis: true autocorrelation is not 0

> shapiro.test(residuals(LM))

 Shapiro-Wilk normality test

data:  residuals(LM)
W = 0.99796, p-value = 0.2657

> lmtest::bptest(LM)

 studentized Breusch-Pagan test

data:  LM
BP = 3.499, df = 3, p-value = 0.3209
```

Experimental design: So what I done is separate the data in four different chunks, and for each of them use the minimum and maximum value of each factor to do a 2^k factorial design. Then use the means of the results with the Yates algorithm to determine the effect of each variable and its interactions.

Results (A=F1 ,B=F2, C=F4, D=F3). So in the final results we can see that F2 is the variable with most effect, followed by F1 and somewhat by F4. While F3 seems to have an inverse effect, but is almost negligible. Must to say that seems that any interactions between the variable do not affect the answer by any means. Therefore if we want to maximize the answer we will maximize F2.

	A	B	C	D	answer	x1	x2	x3	x4	effect
1	0	0	0	0	336.6461	1155.1107	6506.755	13733.2582	27317.650	1707.35315
2	1	0	0	0	818.4646	5351.6446	7226.503	13584.3922	3854.548	481.81850
3	0	1	0	0	2434.9130	1514.9845	6432.322	1927.2740	16786.136	2098.26695
4	1	1	0	0	2916.7316	5711.5184	7152.070	1927.2740	0.000	0.00000
5	0	0	1	0	516.5830	1117.8942	963.637	8393.0678	1439.495	179.93690
6	1	0	1	0	998.4015	5314.4281	963.637	8393.0678	0.000	0.00000
7	0	1	1	0	2614.8500	1477.7680	963.637	0.0000	0.000	0.00000
8	1	1	1	0	3096.6685	5674.3019	963.637	0.0000	0.000	0.00000
9	0	0	0	1	318.0378	481.8185	4196.534	719.7476	-148.866	-18.60825
10	1	0	0	1	799.8563	481.8185	4196.534	719.7476	0.000	0.00000
11	0	1	0	1	2416.3048	481.8185	4196.534	0.0000	0.000	0.00000
12	1	1	0	1	2898.1233	481.8185	4196.534	0.0000	0.000	0.00000
13	0	0	1	1	497.9747	481.8185	0.000	0.0000	0.000	0.00000
14	1	0	1	1	979.7933	481.8185	0.000	0.0000	0.000	0.00000
15	0	1	1	1	2596.2417	481.8185	0.000	0.0000	0.000	0.00000
16	1	1	1	1	3078.0602	481.8185	0.000	0.0000	0.000	0.00000

SMDE Second assignment

DIEGO CAMPOS MILIAN