

Random dataset of correlated variables with assigned binary variable (dependent event) .

The inverse-logit function can be used to construct random sample of data. Data set generated this way can be useful for simulations.

Below I am going to demonstrate the way how to construct correlated variables and assign to them dependent event. In the final step will be measured discriminant power of variable x_1 , x_2 , and x_1+x_2 with usage of modified KS.test implemented in R library.

1. Assumptions

Let's assume that event depends on two correlated variables (case A) and such event is not fully deterministic.

Case A: Level2 = 0.5

corr(x_1, x_2) = 70%

$x_1 \sim N(0,1)$; $x_2 \sim N(0,1)$

Let's construct dependent variable in the following way:

$$P(x_1, x_2) = \frac{1}{1 + \exp(-(w_1 x_1 + w_2 x_2 + w_3 N(0,1)))}$$

w_1, w_2 – weight of variables

w_3 – noise level

R code

```
createdataset<-function (size, level1, level2, w1, w2){  
  
  # Case A corr(x1,x2 )=70%  
  #level1<-0.5  
  # Case B corr(x1,x2 )=0  
  #level1<-0  
  
  #level2 - noise level/unexplained variance  
  
  # 1'st variable  
  x1<-rnorm(size)  
  # 2'nd variable  
  x2<-((1-level1)*x1+level1*rnorm(size))
```

```

#noise / unknown factor
noise<-rmnorm(size)

#dependent variable (Probability)
z1<-1/(1+exp((w1*x1+w2*x2+level2*noise)))

#binary dependent variable with 50% occurrences of 1
zz1<-ifelse(z1 > median(z1),1,0)

}

createdataset(size=10000, level1=0, level2=1,w1=1,w2=1)

```

2. Example of Analysis

Now we can generate sample of data (size =10000) and calculate KS.statistics for variables.

```

source('createdataset.R')

createdataset(size=10000,level1=0.5,level2=1,w1=1,w2=1)

#Let's check correlations between created variables x1 and x2:

cor(x1,x2)

```

[1] 0.7036143

```

source('dkstest.R')

dKS.test(10000,x1,zz1)

```

Two-sample Kolmogorov-Smirnov test

data: variable0 and variable1
D = 0.5928, p-value < 2.2e-16
alternative hypothesis: two-sided

```
dKS.test(10000,x2,zz1)
```

Two-sample Kolmogorov-Smirnov test

data: variable0 and variable1
D = 0.5524, p-value < 2.2e-16
alternative hypothesis: two-sided

```
dKS.test(10000,x1+x2,zz1)
```

Two-sample Kolmogorov-Smirnov test

```
data: variable0 and variable1  
D = 0.6444, p-value < 2.2e-16  
alternative hypothesis: two-sided
```

As we can see above discriminant power measured by D statistic is the highest for x_1+x_2 .

email: datascientist2020@yahoo.com