DATA SIMULATION PROJECT FOR HS-616

Vaishali Chaudhuri

May 23, 2015

Title : A simulated study that shows the association of Intelligence Quotience and Maternal and Infant Factors

Reference Links :http://jama.jamanetwork.com/article.aspx?articleid=194901

Study design , Setting and Participants described in paper

This population study is a prospective longitudinal sub-sample derived from the main Copenhagen Perinatal Cohort comprising of 9125 individuals born at the Copenhagen University Hospital between October 1959 and December 1961.This sub -cohort consists of a sample of 973 men and women When the cohort was established, demographic, socioeconomic, prenatal, and postnatal medical data were recorded prospectively during pregnancy, at delivery, and at a 1-year examination. Information on duration of breastfeeding was collected by a physician who interviewed the mothers at the 1-year examination.

Introduction: This is an Data Simulation Project that references the above link and builds the story based on the paper. Breastfeeding has clear short-term benefits for child survival through reduction of morbidity and mortality from infectious diseases.The paper concludes that certain other parameters (parental and infant) determine Intelligence during adult stage of life determined by WAIS scores Analytics on this simulated dataset is aimed to first generate the data set and then find the answer which mystery parameter has long term benefits on IQ and what's the relation between IQ and the mystery parameter.

Participant's data

- Sex : 976 singletons (490 males and 486 females)
- Age : Mean assessment age of 27.2 years (SD = 4.4; range, 20-34 years)

Main Outcome Measure Intelligence was assessed using the Wechsler Adult Intelligence Scale (WAIS) at a mean age of 27.2 years in the mixed-sex sample

Factors that affect the outcome There is a main factor (not revealed but left for the analyst to come up with) on which the out come depended however thirteen potential confounders were included as covariates: It is upto the analyst to predict which is the primary variable on which the WAISscore is dependant .

```
N=1000
generateTable <- function(N){

  ## Statistical Data  for the Parents ##

  MA <- runif(N, min=(29.3-6.6), max=(29.3+6.6))   # Maternal Age at time of pregnancy
  PSS <-runif(N, min=(4.6-1.9), max=(4.6+1.9)) # Social_Status
  BE <- runif(N, min=(2.6-0.8), max=(2.6+0.8)) # Breadwinners_Education
  MH <- runif(N, min=(163.3-5.4), max=(163.3+5.4)) # Mother's Height (cm)
  MW <- runif(N, min=(4.2-2.5), max=(4.2+2.5)) # Mother's weight gain during pregnancy (kg)
  SM <- sample(c("SMOKER", "NON_SMOKER"), N, replace=TRUE, prob=c(.4, .6))#smokers & nonsmokers
  CC <- ifelse(SM=="SMOKER", runif(N *(0.4),min=(3.7-1.2), max=(3.7+1.2)),0)
  NP <- runif(N, min=(2.0-1.2), max=(2.0+1.2)) # No. of pregnancies
  PC <- runif(N, min=(70.6-37.6), max=(70.6+37.6)) # Pregnancy Complications
  DC <- runif(N, min=(71.6-40.5), max=(71.6+40.5)) # Delivery Complications

  ##### Infant Characteristics
  #Intelligence scores were also affected by 3 factors defined as infant characteristics   at the    ti
  GA <-runif(N, min=(39.2-2.0), max=(39.2+2.0)) # Estimated gestational age(GA) (wk)
  BW <-runif(N, min=(3251-562), max=(3251+562)) # Birth weight(BW) (g)
  BL <-runif(N, min=(51.1-2.6), max=(51.1+2.6)) # Birth height(BL) (cm)

  #Generating data frame based on parental and  infant characteristics
  dataframe1<- data.frame(MA,PSS,BE,MH,MW,CC,NP,PC,DC,GA,BW,BL)
  }
  P_dataset<-generateTable(10e3)
head(P_dataset)
```

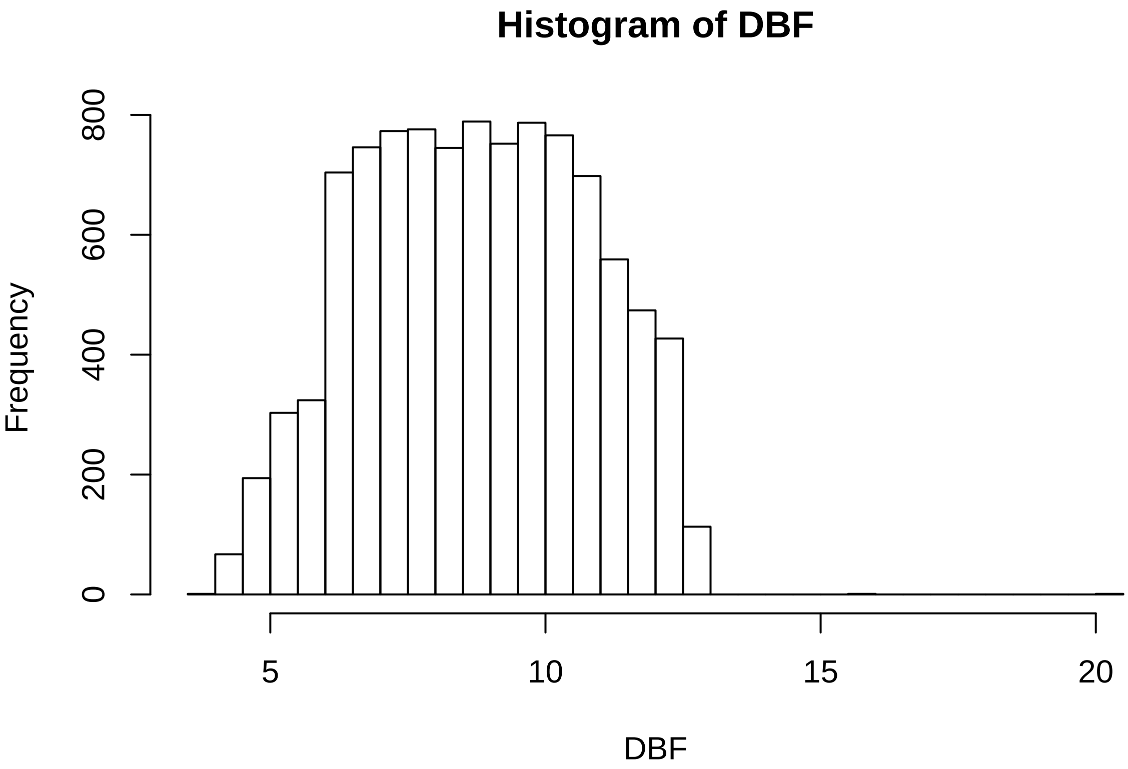
```
##         MA      PSS       BE       MH       MW       CC        NP
## 1 24.84585 5.462692 2.817610 166.7111 3.082430 0.000000 1.5626635
## 2 32.51772 4.189294 3.079230 158.5673 5.796400 4.047844 2.2810410
## 3 26.04591 6.066875 3.048667 164.1036 5.694737 3.415259 3.0410999
## 4 29.09487 4.800967 2.848665 160.6888 3.212658 0.000000 2.7550966
## 5 29.30397 4.312437 2.078316 164.9809 3.350370 3.067798 2.9481748
## 6 27.11469 3.897421 2.223900 161.1644 6.185680 0.000000 0.8079217
##          PC       DC       GA       BW       BL
## 1  76.10619 60.31765 40.20017 3431.962 53.55797
## 2  43.19080 86.75394 39.27888 3651.594 49.51617
## 3  59.64457 65.02808 39.85835 3323.125 52.96368
## 4  37.56743 40.76393 41.15889 2801.826 49.50128
## 5 103.56741 55.53199 37.48487 3525.405 53.59523
## 6  89.13911 96.66655 40.38937 3374.633 52.83531
```

```
# Adding a few outliers to the simulated data as is the case in actual world

  P_dataset$MA[1] <- 43
  P_dataset$PSS[1] <- 2.0
  P_dataset$BE[1] <- 3.2
  P_dataset$BL[1] <- 52
  P_dataset$CC[1] <- 0


  P_dataset$MA[6] <- 52
  P_dataset$PSS[6] <- 2.0
  P_dataset$BE[6] <- 1.2
  P_dataset$BL[6] <- 42
  P_dataset$CC[6] <- 0
 head(P_dataset)
```

```
##         MA      PSS       BE       MH       MW       CC        NP
## 1 43.00000 2.000000 3.200000 166.7111 3.082430 0.000000 1.5626635
## 2 32.51772 4.189294 3.079230 158.5673 5.796400 4.047844 2.2810410
## 3 26.04591 6.066875 3.048667 164.1036 5.694737 3.415259 3.0410999
## 4 29.09487 4.800967 2.848665 160.6888 3.212658 0.000000 2.7550966
## 5 29.30397 4.312437 2.078316 164.9809 3.350370 3.067798 2.9481748
## 6 52.00000 2.000000 1.200000 161.1644 6.185680 0.000000 0.8079217
##          PC       DC       GA       BW       BL
## 1  76.10619 60.31765 40.20017 3431.962 52.00000
## 2  43.19080 86.75394 39.27888 3651.594 49.51617
## 3  59.64457 65.02808 39.85835 3323.125 52.96368
## 4  37.56743 40.76393 41.15889 2801.826 49.50128
## 5 103.56741 55.53199 37.48487 3525.405 53.59523
## 6  89.13911 96.66655 40.38937 3374.633 42.00000
```

Duration of breastfeeding was positively associated with mother's age, social status, education, birth weight and negatively associated with cigarette consumption. So we form an equation giving weightage to these parameters according to their association

```
 DBF<- DBF <- with (P_dataset ,
              10^(-0.3) * (MA) +
              10^(-1) * (PSS) +
              10^(-1.2) * (BE) +
             -10^(-0.4) * (CC) -6)
             hist(DBF, breaks=50)
```



```
#a=-0.3,b=-1,c=-1.2,d=-2.4,e=-0.4

P_dataset$DBF<- DBF
head(P_dataset)
```

```
##          MA      PSS       BE       MH       MW       CC        NP
## 1 43.00000 2.000000 3.200000 166.7111 3.082430 0.000000 1.5626635
## 2 32.51772 4.189294 3.079230 158.5673 5.796400 4.047844 2.2810410
## 3 26.04591 6.066875 3.048667 164.1036 5.694737 3.415259 3.0410999
## 4 29.09487 4.800967 2.848665 160.6888 3.212658 0.000000 2.7550966
## 5 29.30397 4.312437 2.078316 164.9809 3.350370 3.067798 2.9481748
## 6 52.00000 2.000000 1.200000 161.1644 6.185680 0.000000 0.8079217
##          PC       DC       GA       BW       BL       DBF
## 1  76.10619 60.31765 40.20017 3431.962 52.00000 15.952957
## 2  43.19080 86.75394 39.27888 3651.594 49.51617  9.299206
## 3  59.64457 65.02808 39.85835 3323.125 52.96368  6.493283
## 4  37.56743 40.76393 41.15889 2801.826 49.50128  9.241811
## 5 103.56741 55.53199 37.48487 3525.405 53.59523  8.027842
## 6  89.13911 96.66655 40.38937 3374.633 42.00000 20.337451
```

To choose the correct coefficient values to the different parameters responsible for the "Duration of Breast feeding" equation . The manipulate function is used to set the parameters

```
logistic <- function(t) 1 / (1 + exp(-t))
library(manipulate)
manipulate({
  DBF <- with (P_dataset ,
               10^a * (MA) +
               10^b * (PSS) +
               10^c * (BE) +
              -10^e * (CC) -6)
              hist(DBF, breaks=50)
},
a=slider(-9, 9, step=0.1, initial = -1),
b=slider(-9, 9, step=0.1, initial = -1),
c=slider(-9, 9, step=0.1, initial = -1),
d=slider(-9, 9, step=0.1, initial = -1),
e=slider(-9, 9, step=0.1, initial = -1))

#a=-0.3,b=-1,c=-1.2,e=-0.4
```

Finally the output is in the form of IQ score of the participants which is WAIS score of the participants

```
  WAISscore <- 20*DBF - DBF^2 + rnorm(N, sd=2)  # add some noise

P_dataset$WAISscore<- WAISscore
colnames(P_dataset) <- c("MaternaL_Age","Social_Status","Parent_Educn","Mothers_ht","Mothers_wt_gain","
IQ_cohort <-P_dataset
names(IQ_cohort) <- tolower(names(IQ_cohort)
head(IQ_cohort)
write.csv(IQ_cohort,file="/Users/vchaudhuri/Desktop/HS-616/IQ_data.csv")
```

Analytics

Exploratory Analysis (TO make sense of this data table and predict which variable should be considered in prediction of the outcome) I used the correlogram package to generate a correlation matrix of the different vraiables. It is very useful to highlight the most correlated variables in a data table. In this plot, correlation coefficients is colored according to the value. Correlation matrix can be also reordered according to the degree of association between variables. The R corrplot package is used here.

```
library("corrplot")
M<-cor(IQ_cohort)
head(round(M,2))
corrplot(M, method="circle")


## the corrplot package narRows down the Independant varaiable to Duration of Breast Feeding(DBF). Howe
```

Analyzed each parameter vs iq_level

```
library("ggplot2")
ggplot(IQ_cohort)+geom_point(aes(x=maternal_age,y=iq_level,color='red'))
ggplot(IQ_cohort)+geom_point(aes(x=social_status,y=iq_level))
ggplot(IQ_cohort)+geom_point(aes(x=parent_educn,y=iq_level))
ggplot(IQ_cohort)+geom_point(aes(x=mothers_ht,y=iq_level))
ggplot(IQ_cohort)+geom_point(aes(x=mothers_wt_gain,y=iq_level))
ggplot(IQ_cohort)+geom_point(aes(x=cig_cons,y=iq_level))
ggplot(IQ_cohort)+geom_point(aes(x=no._of_pregn,y=iq_level))
ggplot(IQ_cohort)+geom_point(aes(x=preg_compl,y=iq_level))
ggplot(IQ_cohort)+geom_point(aes(x=delivery_compl,y=iq_level))
ggplot(IQ_cohort)+geom_point(aes(x=gestational_age,y=iq_level))
ggplot(IQ_cohort)+geom_point(aes(x=birth_wt,y=iq_level))
ggplot(IQ_cohort)+geom_point(aes(x=birth_len,y=iq_level))
ggplot(IQ_cohort)+geom_point(aes(x=DBF,y=iq_level))
```

The two variables that seem to affect IQ level seems to be maternal age and Durtaion of breast feeding

```
## We first try to fit linear modelling on this data with a small sample set:

Population_dataset<-generateTable(30)
with (IQ_cohort, plot(DBF, iq_level))
fit1 <- lm(iq_level ~ DBF, IQ_cohort)
abline(fit1, col="red")
summary(fit1)


Population_dataset<-generateTable(30)
with (IQ_cohort, plot(maternal_age, iq_level))
fit2 <- lm(iq_level ~ maternal_age, IQ_cohort)
abline(fit2, col="blue")
summary(fit2)

Population_dataset<-generateTable(100)
with (IQ_cohort, plot(DBF, iq_level))
fit1_quad <- lm(iq_level ~ DBF + I(DBF^2), data=IQ_cohort)
abline(fit1_quad, col="red")
summary(fit1_quad)





#R-squared goes from essentially 0 to close to
#1 (it would be a perfect fit if we had not added any noise).
```