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 .

```
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
  MA[1] <- 45
  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)

DBF<- DBF <- (
             10^(-0.3) * (MA) +
             10^(-1) * (PSS) +
             10^(-1.2) * (BE) +
            -10^(-0.4) * (CC) -6)
# Finally the output is in the form of IQ score of the participants which is WAIS score of the partic

WAISscore <- 20*DBF - DBF^2 + rnorm(N, sd=2)

#Generating data frame based on parental and  infant characteristics
dataframe1<- data.frame(MA,PSS,BE,MH,MW,CC,NP,PC,DC,GA,BW,BL,DBF,WAISscore)

}

P_dataset<-generateTable(10e3)

head(P_dataset)
```

```
##         MA      PSS       BE       MH       MW       CC       NP        PC
## 1 45.00000 5.630244 2.848779 161.1664 2.797476 0.000000 2.736067  53.16498
## 2 35.22045 2.871786 2.892034 164.2781 4.889457 3.539332 2.437927  57.69148
## 3 23.23265 4.378425 2.051052 161.4906 4.260270 4.118337 1.713255  33.10616
## 4 29.25597 5.589476 2.689307 163.5695 3.233374 0.000000 1.740213  56.81725
## 5 28.74541 4.962382 2.562159 164.8129 6.160049 3.860006 2.492847 107.36497
## 6 25.25194 6.250096 2.177162 158.8509 4.866257 0.000000 1.352927  92.40625
##          DC       GA       BW       BL       DBF WAISscore
## 1  77.11274 39.42380 3413.101 51.66645 17.296196  44.81853
## 2 104.40796 39.80721 2858.689 52.05690 10.712658 102.09320
## 3  56.84296 39.41184 3024.029 52.90034  4.571621  70.53890
## 4 105.25749 38.12141 2993.569 50.18472  9.391351  97.52406
## 5  45.44066 38.42723 3618.827 50.74769  7.528036  93.77505
## 6  44.15999 40.12371 3501.898 50.78997  7.418331  94.42386
```

```
# 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$DBF[1] <- (
               10^(-0.3) * (P_dataset$MA[1]) +
```

```r
               10^(-1) * (P_dataset$PSS[1]) +
               10^(-1.2) * (P_dataset$BE[1]) +
              -10^(-0.4) * (P_dataset$CC[1]) -6)
   P_dataset$WAISscore[1] <- 20*(P_dataset$DBF[1]) - (P_dataset$DBF[1])^2


  P_dataset$MA[6] <- 47
  P_dataset$PSS[6] <- 2.0
  P_dataset$BE[6] <- 3.2
  P_dataset$BL[6] <- 52
  P_dataset$DBF[6] <- (
               10^(-0.3) * (P_dataset$MA[6]) +
               10^(-1) * (P_dataset$PSS[6]) +
               10^(-1.2) * (P_dataset$BE[6]) +
              -10^(-0.4) * (P_dataset$CC[6]) -6)
   P_dataset$WAISscore[1] <- 20*(P_dataset$DBF[6]) - (P_dataset$DBF[6])^2
 head(P_dataset)
```

```
##         MA      PSS       BE       MH       MW       CC       NP        PC
## 1 43.00000 2.000000 3.200000 161.1664 2.797476 0.000000 2.736067  53.16498
## 2 35.22045 2.871786 2.892034 164.2781 4.889457 3.539332 2.437927  57.69148
## 3 23.23265 4.378425 2.051052 161.4906 4.260270 4.118337 1.713255  33.10616
## 4 29.25597 5.589476 2.689307 163.5695 3.233374 0.000000 1.740213  56.81725
## 5 28.74541 4.962382 2.562159 164.8129 6.160049 3.860006 2.492847 107.36497
## 6 47.00000 2.000000 3.200000 158.8509 4.866257 0.000000 1.352927  92.40625
##          DC       GA       BW       BL       DBF WAISscore
## 1  77.11274 39.42380 3413.101 52.00000 15.952957  36.67491
## 2 104.40796 39.80721 2858.689 52.05690 10.712658 102.09320
## 3  56.84296 39.41184 3024.029 52.90034  4.571621  70.53890
## 4 105.25749 38.12141 2993.569 50.18472  9.391351  97.52406
## 5  45.44066 38.42723 3618.827 50.74769  7.528036  93.77505
## 6  44.15999 40.12371 3501.898 52.00000 17.957706  94.42386
```

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 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

```r
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
```

The next steps is to change the abbreviated column names to more descriptive names and create a simulated Data Table which describes the IQ level of an adult at 27.2 years of age that might could have been influenced by a bunch of factors in infancy or by the characteristics determined by the parents . Our job is to back run analytics and find out which one .

```
colnames(P_dataset) <- c("MaternaL_Age","Social_Status","Parent_Educn","Mothers_ht","Mothers_wt_gain","C
IQ_cohort <-P_dataset
names(IQ_cohort) <- tolower(names(IQ_cohort))
head(IQ_cohort)
```

```
##   maternal_age social_status parent_educn mothers_ht mothers_wt_gain
## 1     43.00000      2.000000     3.200000   161.1664        2.797476
## 2     35.22045      2.871786     2.892034   164.2781        4.889457
## 3     23.23265      4.378425     2.051052   161.4906        4.260270
## 4     29.25597      5.589476     2.689307   163.5695        3.233374
## 5     28.74541      4.962382     2.562159   164.8129        6.160049
## 6     47.00000      2.000000     3.200000   158.8509        4.866257
##   cig_cons no._of_pregn preg_compl delivery_compl gestational_age birth_wt
## 1 0.000000     2.736067   53.16498       77.11274        39.42380 3413.101
## 2 3.539332     2.437927   57.69148      104.40796        39.80721 2858.689
## 3 4.118337     1.713255   33.10616       56.84296        39.41184 3024.029
## 4 0.000000     1.740213   56.81725      105.25749        38.12141 2993.569
## 5 3.860006     2.492847  107.36497       45.44066        38.42723 3618.827
## 6 0.000000     1.352927   92.40625       44.15999        40.12371 3501.898
##   birth_len durn_breast_feed  iq_level
## 1  52.00000        15.952957  36.67491
## 2  52.05690        10.712658 102.09320
## 3  52.90034         4.571621  70.53890
## 4  50.18472         9.391351  97.52406
## 5  50.74769         7.528036  93.77505
## 6  52.00000        17.957706  94.42386
```

```
write.csv(IQ_cohort,file="/Users/vchaudhuri/Desktop/HS-616/IQ_data.csv")
```

Analytics

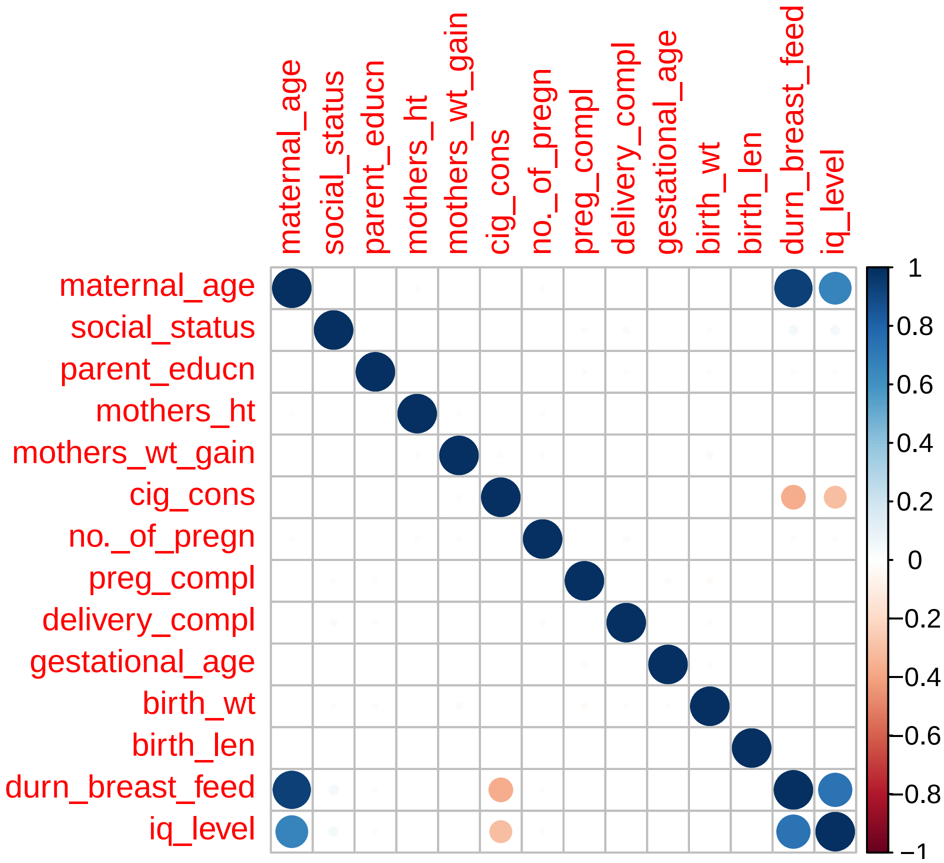
Exploratory Analysis (To make sense of this data table and predict which variable should be considered in prediction of the outcome) I used the correlogram package to generate a correlation matrix of the different vraiables. It is very useful to highlight the most correlated variables in a data table. 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))
```

```
##                 maternal_age social_status parent_educn mothers_ht
```

```
## maternal_age           1.00          0.00          0.00        0.01
## social_status          0.00          1.00          0.01        0.00
## parent_educn           0.00          0.01          1.00        0.00
## mothers_ht             0.01          0.00          0.00        1.00
## mothers_wt_gain        0.01         -0.01          0.01        0.01
## cig_cons              -0.01          0.00          0.00        0.00
##                 mothers_wt_gain cig_cons no._of_pregn preg_compl
## maternal_age               0.01    -0.01         0.01       0.01
## social_status             -0.01     0.00         0.01      -0.01
## parent_educn               0.01     0.00         0.00       0.01
## mothers_ht                 0.01     0.00         0.01       0.00
## mothers_wt_gain            1.00     0.01        -0.01      -0.01
## cig_cons                   0.01     1.00         0.00       0.00
##                 delivery_compl gestational_age birth_wt birth_len
## maternal_age             -0.01           -0.01     0.00      0.00
## social_status             0.02            0.01    -0.02     -0.01
## parent_educn              0.01            0.00     0.01      0.01
## mothers_ht                0.00           -0.01    -0.01     -0.01
## mothers_wt_gain           0.00            0.00     0.03     -0.01
## cig_cons                  0.00            0.01     0.01      0.01
##                 durn_breast_feed iq_level
## maternal_age                0.93     0.67
## social_status               0.05     0.05
## parent_educn                0.01     0.01
## mothers_ht                  0.01     0.01
## mothers_wt_gain             0.00     0.00
## cig_cons                   -0.36    -0.31
```

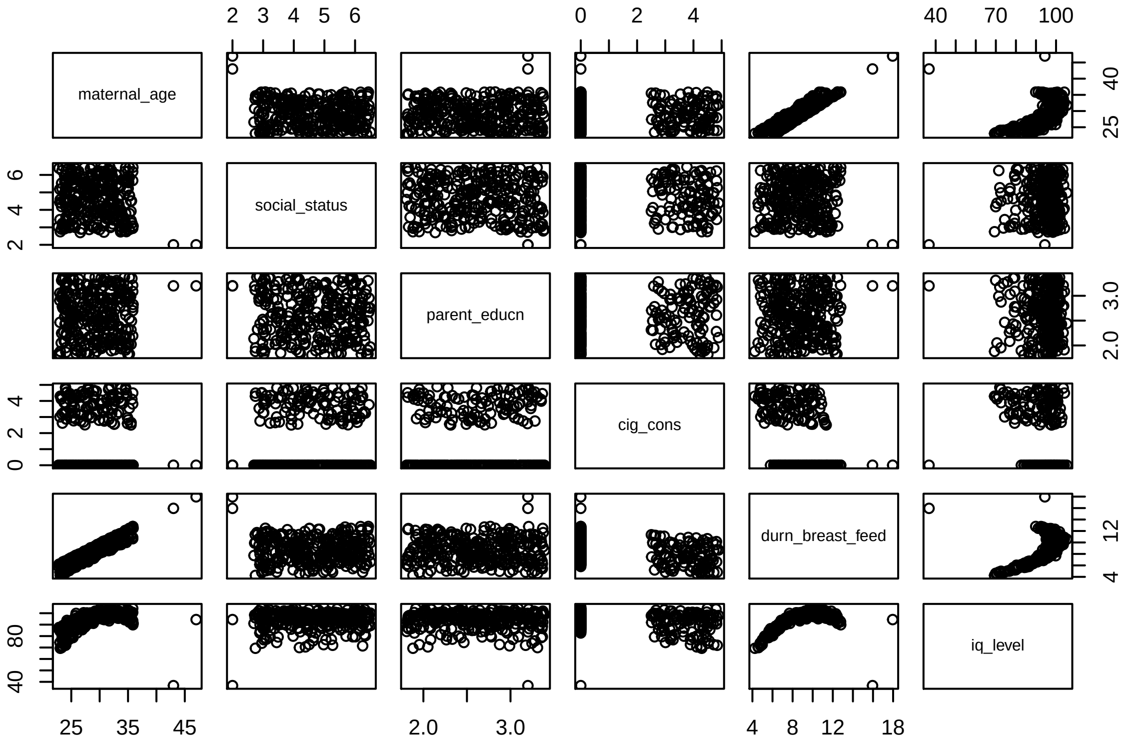
```
corrplot(M, method="circle")
```




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

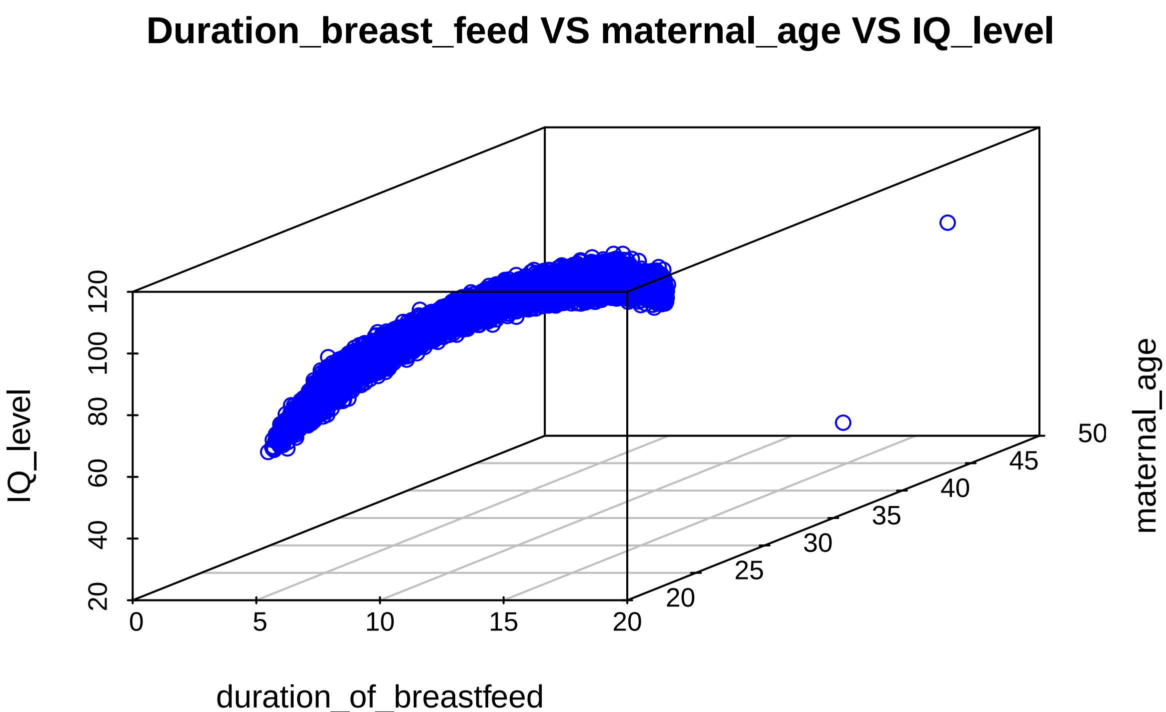
A scatterplot matrix compares each variable in a dataset against each of the other variables using scatter plots. To understand the relationship between each variable and the outcome this graph was plotted.

```
Population_dataset<-IQ_cohort[1:300,]
plot(Population_dataset[c("maternal_age", "social_status", "parent_educn", "cig_cons", "durn_breast_fee
```



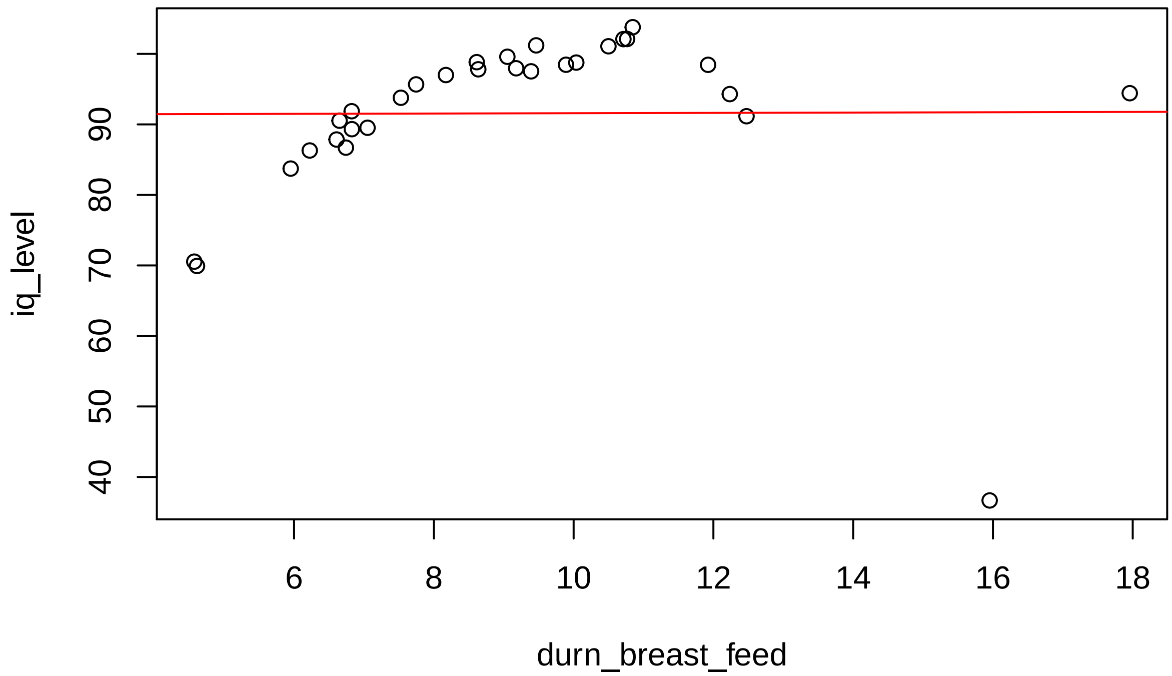
The two variables that seem to affect IQ level seems to be maternal age and Durtaion of breast feeding. So all three are plotted into a 3D scatter plot

```
library("scatterplot3d")

IQ3d <-scatterplot3d( IQ_cohort$durn_breast_feed,IQ_cohort$maternal_age,IQ_cohort$iq_level,
                     color = "blue",
              xlab= "duration_of_breastfeed",ylab= "maternal_age",zlab= "IQ_level",
              main="Duration_breast_feed VS maternal_age VS IQ_level")
```



Checking which parameter and which function gives the best fit

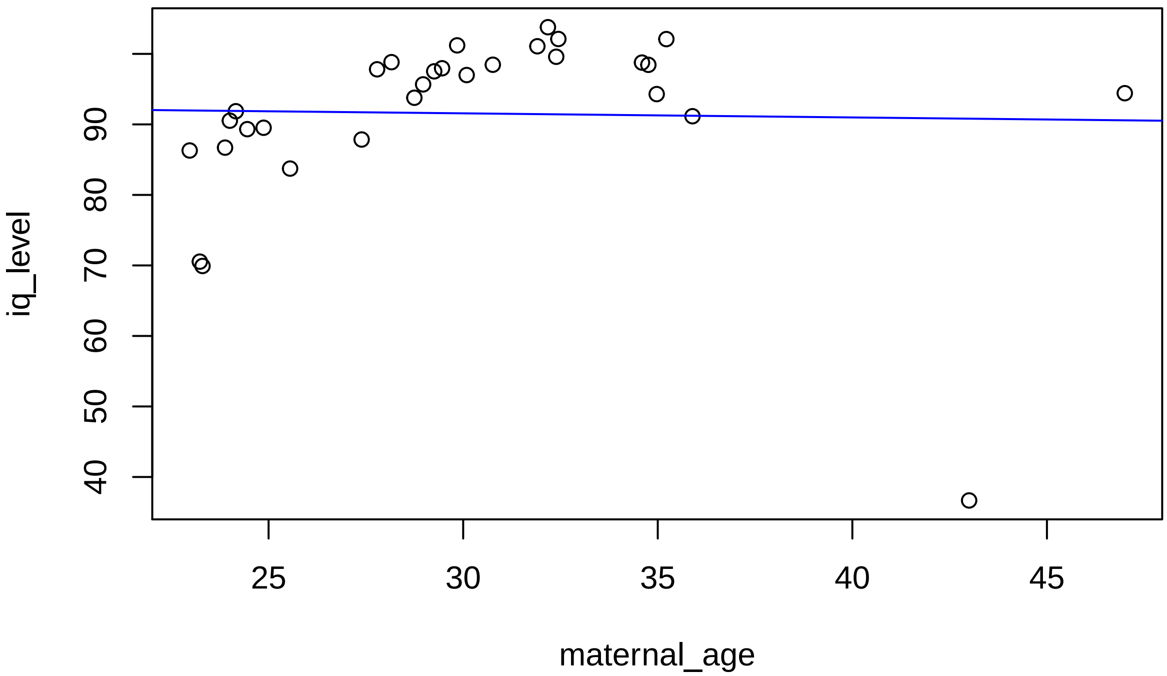
```
#Fitting IQ_level as a function of Duration of Breast feeding in a linear model with a small sample siz

Population_dataset<-IQ_cohort[1:30,]
with (Population_dataset, plot(durn_breast_feed, iq_level))
fit1 <- lm(iq_level ~ durn_breast_feed, Population_dataset)
abline(fit1, col="red")
```




```r
summary(fit1)
```

```
##
## Call:
## lm(formula = iq_level ~ durn_breast_feed, data = Population_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.045  -2.140   3.398   7.105  12.177
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      91.35771    7.93019  11.520 3.88e-12 ***
## durn_breast_feed  0.02272    0.82825   0.027    0.978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.44 on 28 degrees of freedom
## Multiple R-squared:  2.688e-05,  Adjusted R-squared:  -0.03569
## F-statistic: 0.0007527 on 1 and 28 DF,  p-value: 0.9783
```

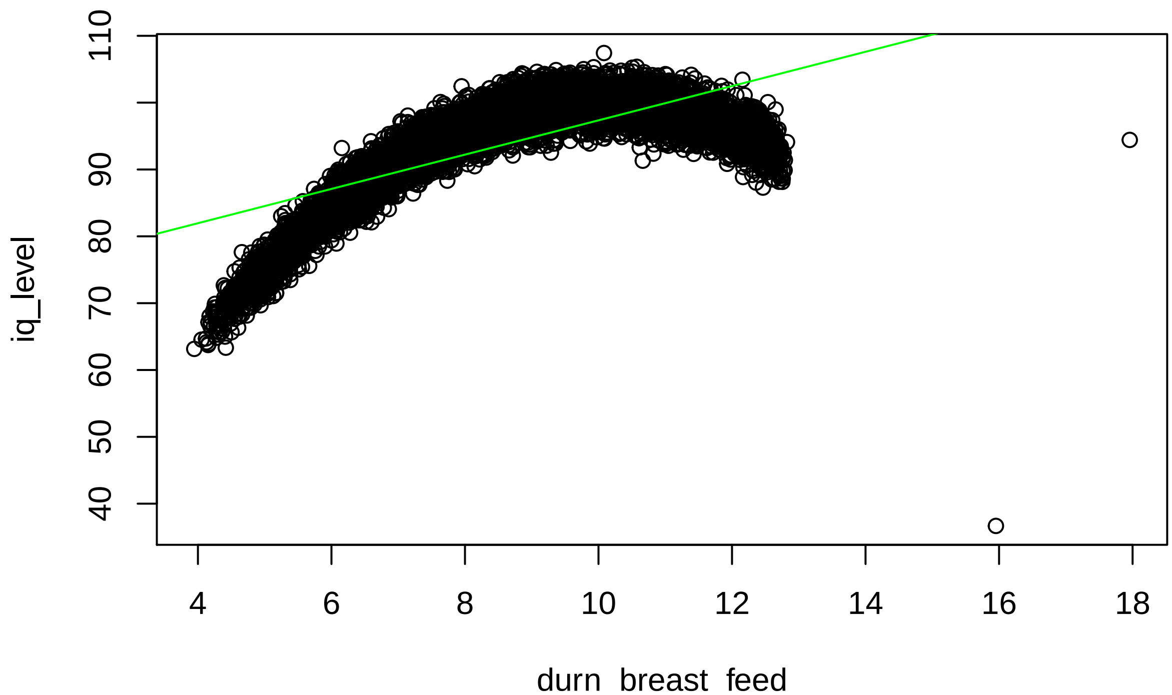
```r
#Fitting IQ_level as a function of Maternal Age in a linear model with a small sample size

Population_dataset<-IQ_cohort[1:30,]
with (Population_dataset, plot(maternal_age, iq_level))
fit2 <- lm(iq_level ~ maternal_age, Population_dataset)
abline(fit2, col="blue")
```




```
summary(fit2)
```

```
##
## Call:
## lm(formula = iq_level ~ maternal_age, data = Population_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.137  -2.509   3.944   7.154  12.341
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  93.30984   13.31400   7.008 1.27e-07 ***
## maternal_age -0.05809    0.43559  -0.133    0.895
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.44 on 28 degrees of freedom
## Multiple R-squared:  0.0006349,  Adjusted R-squared:  -0.03506
## F-statistic: 0.01779 on 1 and 28 DF,  p-value: 0.8949
```

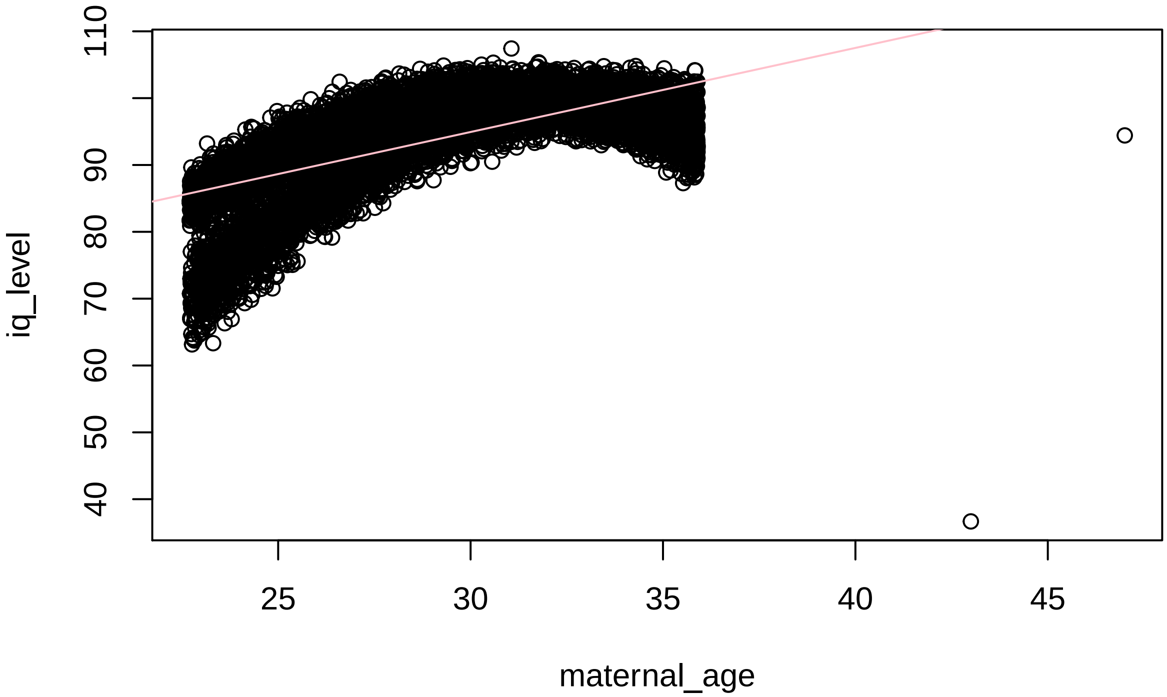
```
#Fitting IQ_level as a function of Duration of Breast feeding in a linear model with a the entire sampl

Population_dataset<-IQ_cohort
with (Population_dataset, plot(durn_breast_feed, iq_level))
fit3 <- lm(iq_level ~ durn_breast_feed, Population_dataset)
abline(fit3, col="green")
```




```r
summary(fit3)
```

```
##
## Call:
## lm(formula = iq_level ~ durn_breast_feed, data = Population_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.949  -2.552   1.111   3.455  10.352
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      71.70419    0.21200   338.2   <2e-16 ***
## durn_breast_feed  2.56503    0.02367   108.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.899 on 9998 degrees of freedom
## Multiple R-squared:  0.5401, Adjusted R-squared:  0.5401
## F-statistic: 1.174e+04 on 1 and 9998 DF,  p-value: < 2.2e-16
```

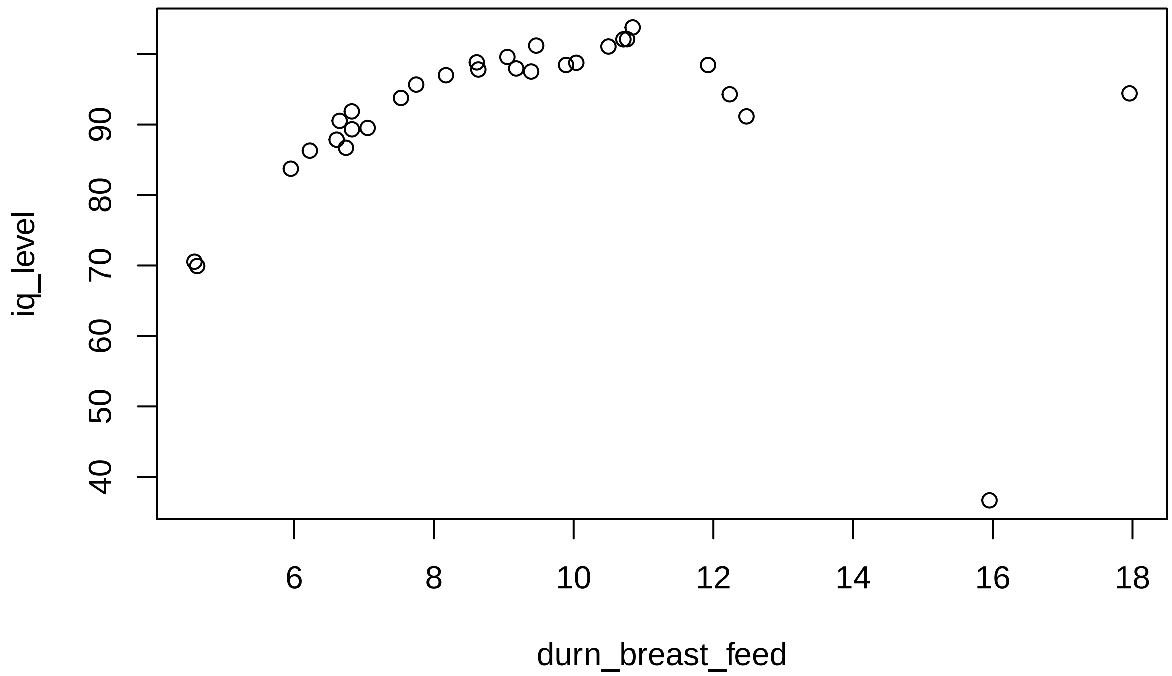
```r
#Fitting IQ_level as a function of Maternal_Age in a linear model with a the entire sample size IQ_coho

Population_dataset<-IQ_cohort
with (Population_dataset, plot(maternal_age, iq_level))
fit4 <- lm(iq_level ~ maternal_age, Population_dataset)
abline(fit4, col="pink")
```




```r
summary(fit4)
```

```
##
## Call:
## lm(formula = iq_level ~ maternal_age, data = Population_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.615  -2.814   0.919   3.781  11.791
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57.19566    0.41340  138.35   <2e-16 ***
## maternal_age  1.25801    0.01399   89.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.372 on 9998 degrees of freedom
## Multiple R-squared:  0.4472, Adjusted R-squared:  0.4471
## F-statistic:  8087 on 1 and 9998 DF,  p-value: < 2.2e-16
```

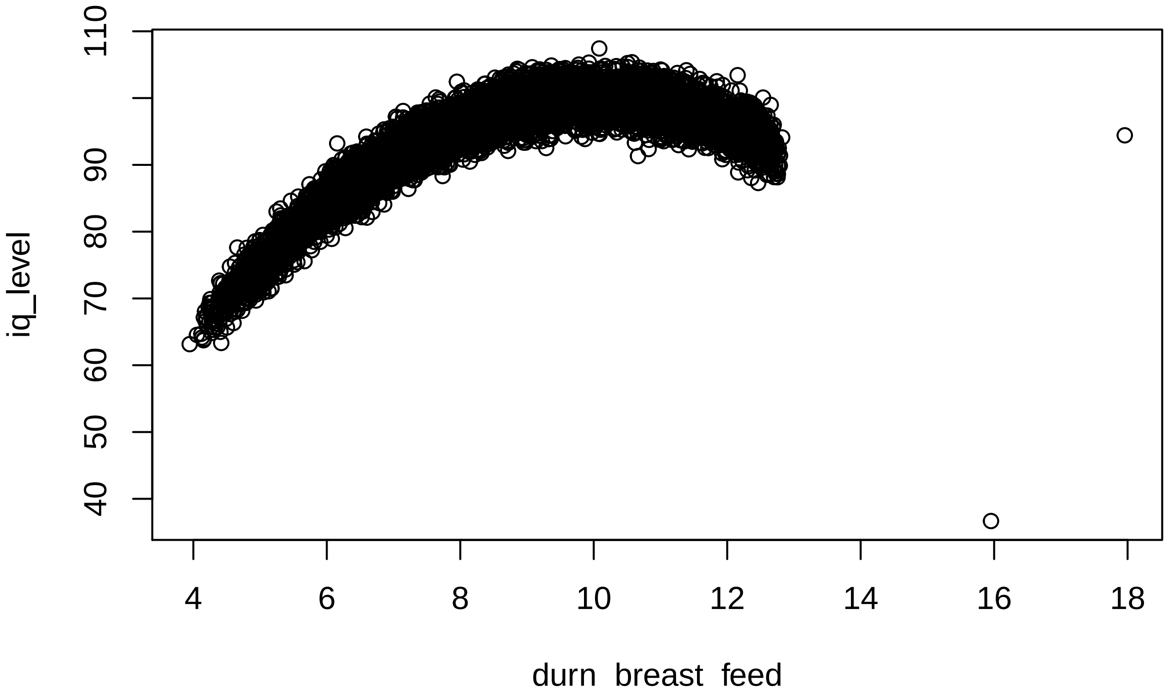
```r
#Fitting IQ_level as a function of duration of Breast feeding in a Quadratic model with a small sample 

Population_dataset<-IQ_cohort[1:30,]
with (Population_dataset, plot(durn_breast_feed, iq_level))
```




```
fit1_quad <- lm(iq_level ~ durn_breast_feed + I((durn_breast_feed)^2), data=Population_dataset)
summary(fit1_quad)
```

```
##
## Call:
## lm(formula = iq_level ~ durn_breast_feed + I((durn_breast_feed)^2),
##     data = Population_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.145  -0.815   0.792   2.482  29.574
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               27.0904    16.3698   1.655 0.109524
## durn_breast_feed          13.4369     3.2245   4.167 0.000284 ***
## I((durn_breast_feed)^2)  -0.6312     0.1486  -4.248 0.000229 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.6 on 27 degrees of freedom
## Multiple R-squared:  0.4006, Adjusted R-squared:  0.3562
## F-statistic: 9.023 on 2 and 27 DF,  p-value: 0.0009975
```

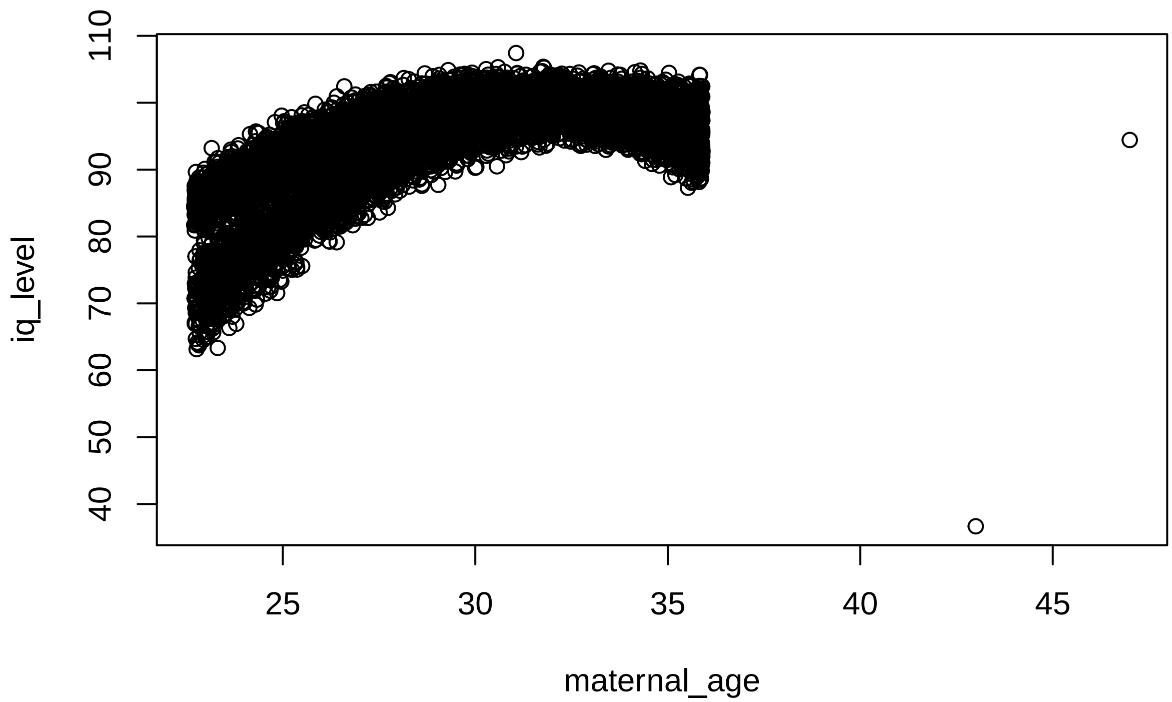
```
#Fitting IQ_level as a function of duration of Breast feeding in a Quadratic model with a large sample s

Population_dataset<-IQ_cohort
with (IQ_cohort, plot(durn_breast_feed, iq_level))
```




```
fit1_quad <- lm(iq_level ~ durn_breast_feed + I((durn_breast_feed)^2), data=Population_dataset)
summary(fit1_quad)
```

```
##
## Call:
## lm(formula = iq_level ~ durn_breast_feed + I((durn_breast_feed)^2),
##     data = Population_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.899  -1.368  -0.006   1.388  56.032
##
## Coefficients:
##                          Estimate Std. Error  t value Pr(>|t|)
## (Intercept)              1.551902   0.344006    4.511 6.52e-06 ***
## durn_breast_feed        19.623481   0.081325  241.296  < 2e-16 ***
## I((durn_breast_feed)^2) -0.978520   0.004629 -211.399  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.095 on 9997 degrees of freedom
## Multiple R-squared:  0.9159, Adjusted R-squared:  0.9159
## F-statistic: 5.446e+04 on 2 and 9997 DF,  p-value: < 2.2e-16
```

```
#Fitting IQ_level as a function of maternal_age in a Quadratic model with a large sample size

Population_dataset<-IQ_cohort
with (IQ_cohort, plot(maternal_age, iq_level))
```




```
fit2_quad <- lm(iq_level ~ maternal_age + I((maternal_age)^2), data=Population_dataset)
summary(fit1)
```

```
##
## Call:
## lm(formula = iq_level ~ durn_breast_feed, data = Population_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.045  -2.140   3.398   7.105  12.177
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      91.35771    7.93019  11.520 3.88e-12 ***
## durn_breast_feed  0.02272    0.82825   0.027    0.978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.44 on 28 degrees of freedom
## Multiple R-squared:  2.688e-05,  Adjusted R-squared:  -0.03569
## F-statistic: 0.0007527 on 1 and 28 DF,  p-value: 0.9783
```

CONCLUSION

R-squared goes from essentially 0 to close to 1 when Iqlevel is a quadratic function of Duration of Breast Feeding