Analytical skills of young gifted children

Khawla Bounouh 3137274 Mohamed Taha Tiyal 3137276

January 7, 2022

1 Introduction

Children who are gifted are defined as those who demonstrate an advanced ability or potential in one or more specific areas when compared to others of the same age or to their environment. While no two gifted children are the same, many share common gifted characteristics and traits, such as: →Advanced learning and reasoning above their age peers, Strong ability in observation and analysis and Excellent memory.
In many cases, in order to classify if a kid is gifted or not, tests known as IQ tests are used. It's an assessment that measures a range of cognitive abilities, like the ones mentioned above, and provides a score that is intended to serve as a measure of an individual's intellectual abilities and potential. Hence, any person who has a score above 130 is considered gifted.
An investigator was interested in understanding the relationship, if any, between the analytical skills of young gifted children and seven other different variables. Therefore, he gathered all the data he needed and evaluated the analytical skills of young gifted children using IQ tests. The Scores of these tests are our dependent variable.
From the seven variables he gathered, we will only focus on three of them as our independent variables:
MotherIQ: Which the IQ of the mother of the gifted child.
Read: Average number of hours per week the parents read to the child.
Edutv: Average number of hours per week the child watched an educational program on TV.
We will then research whether these three variables can predict the IQ score of gifted children. Additionally, we will focus on "Read" as an independent variable by investigating whether the mean IQ of gifted children significantly differs from 0 if this variable is: less than or equal to 2.1 hours or above 2.1 hours.

2 Methods

2.1 dataset

Data were collected from different schools located in a large city on a set of thirty-six children who were identified as gifted children soon after they reached the age of four. This data is considered to be a random sample and was released in a book in 1994.
This data contains clear and precise observations, in addition to very direct and coherent numbers. However, the lack of information and specifications on the location of these schools and city makes the data less reliable. In addition to the sample size, that is relatively small and lacks diversity, which makes it hard to come up with general and conclusive results.

2.2 data cleaning

Before the data set could be analyzed, columns dedicated to variables that we won't be considering in this study were filtered out. Four columns were then deleted which left us with three remaining variables, in addition to our dependent variable "IQ score".

2.3 Basic exploration of data

Before any tests were carried out, the data set was explored through various plots. First, the correlation between the dependent variable "score" and the independent variables: "motherIQ" and "Edutv" was explored using a 3D-scatterplot and two other distinct scatterplots. The variable "read" was not included in the plots because its influence on the dependent variable will be investigated through a 2-sample unpaired t-test later in the paper.

2.4 Multivariate linear regression

Multiple linear regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.
Multivariate linear regression was studied and depicted through a regression line. We chose ("score") as our main response variable whereas ("motherIQ"), ("read"), and the ("edutv") represent our explanatory variables. The main objective of this analysis is to study to which degree the variance of the response variable was related, or in other words could be factually explained , to our explanatory variables. Afterwards, hypothesis testing was used to find out whether the intercept and slopes of the regression equation were significantly different than 0. As a result all five assumptions for the usage of multivariate regression were checked. First, the normal distribution of residuals was checked using a density plot and a Q-Q plot of the residuals. Second, the homoscedasticity of the residuals was checked through a scatter plot of the residuals vs predicted values. At last, the independence of the variables and the linearity were checked with visual representations.

2.5 2-sample Unpaired t-test

Before the t-test was carried out, the independent variable "read: Average number of hours per week the parents read to the child " was divided into two groups:
Group (1): Average number of hours per week less then or equal to 2.1
Group (2): Average number of hours per week above 2.1
Hence, we ended up with two distinct subtables: A subtable containing Group (1) and another subtable containing Group (2). Afterwards, a 2-sample t-test was performed, to evaluate the mean IQ scores of each group, if any, the difference between them.
Finally, we're gonna proceed to check the assumptions for the t-test: Normal distribution using a Shapiro-wilK test and an F-test to test the assumption that both samples have a common variance.

3 Results

3.1 Basic exploration of data

The 3D plot of the dependent variable "score" and the independent variables "motherIQ" and "Edutv" in Figure 1 shows that a higher IQ score can be associated with a higher mother IQ and a less number of hours watching educational programs on TV("edutv").

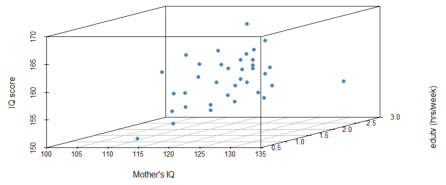
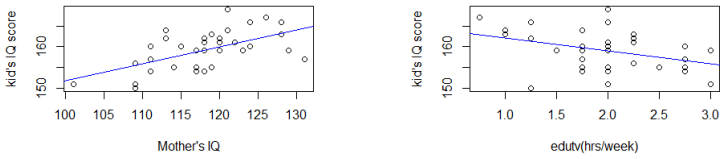


Figure 1: 3D plot of Gifted Children's IQ according to edutv and mother's IQ

Let's plot each independent variable with the dependent variable in order to get a clearer visualization of the data :

By displaying the relationship of each independent variable with the dependant variable visually, we get a clear image of the type of correlation they have. The plot on the left of figure 2 displays a



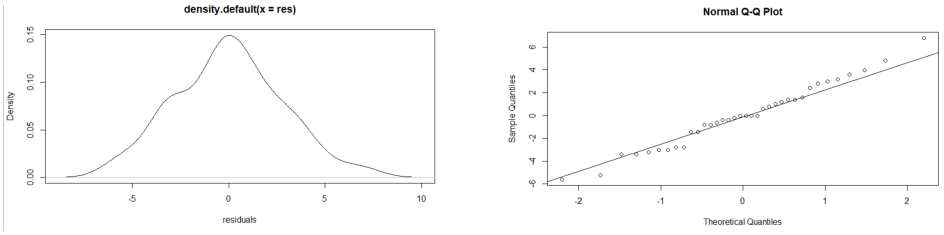
(a) positive relationship (b) negative relationship

Figure 2: The relationship between the dependent variable and the independent variables

positive relationship between the dependent variable and the mother's IQ, while the plot on the right displays a negative relationship between the dependent variable and the hours of edutv.
Other than observe the relationship between the variables, the visualisation of data, is also a straightforward method of identifying outliers. In my data however, I think that no significant outlier is worth removing from the observations, but I thought that it would be interesting to mention that the scatter plot of "score" against "Edutv" displays an atypical situation: There is a kid who watches an average of 2 hours/week of edutv and is still profoundly gifted with an IQ score of near 170. It is strange because the scatter plot clearly shows that the other kids his age who watch edutv for longer hours have less and less IQ. My guess is that he was subject to other specific factors in his environment that deeply stimulated his analytical skills, or maybe the educational program he was watching really helped him develop his cognitive skills. But this is beyond the scope of my study, and I think that it is pretty normal to find these kinds of exceptions.

3.2 Multivariate linear regression

Before starting the test, let's first test out the assumptions of the multivariate linear regression. The independence is checked since all the variables are independent from each other, and the linearity was also checked with the different data visualisation plots.(see the R script for all the plots) Moreover, based on Figure 3, we can conclude that the assumption that the residuals are normally distributed is satisfied. As the density plot in Figure 3/a approximates the shape of a normal distribution and as the residuals in the Q-Q plot of Figure 3/b almost all fall in a straight line.



(a) (b)

Figure 3: The relationship between the dependent variable and the independent variables

Furthermore, Figure 4 shows that the assumption of multivariate linear regression that the residuals are homoscedastic is also satisfied. Since there is no systematic change in the spread of residuals over the range of measured values.

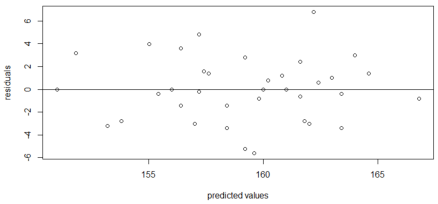


Figure 4: residuals vs predicted values

```
Call:
lm(formula = z ~ x + y + t, data = dataset_filtered)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6035 -1.7634 -0.0054  1.4456  6.7931

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 87.74186   12.18228   7.202 3.52e-08 ***
x            0.40044    0.08076   4.959 2.24e-05 ***
y           -0.79303    0.95214  -0.833    0.411
t           11.99889    2.44313   4.911 2.57e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.917 on 32 degrees of freedom
Multiple R-squared:  0.637,     Adjusted R-squared:  0.603
F-statistic: 18.72 on 3 and 32 DF,  p-value: 3.405e-07
```

Figure 5

We will now interpret the results of or multiple linear regression. The main variable is the children's IQ and the explanatory variables are the time spent watching edutv per week, the mother's IQ, the time spent being read to. First let's proceed and identify our sample equation :
Children's IQ = 87.74186 + 0.40044 * (mother's IQ) – 0.79303 * (Edutv) + 11.99889 * (read)
Now considering the results of our analysis we notice that both the p-values for the mother's IQ and the time spent being read to are in the range to make these variables statistically significant ($p < alpha = 0.05$) whereas the p-value for the explanatory variable ("Edutv") is not statistically significant since its p-value is greater than 0.05. We can also verify that these values are non zero by comparing their absolute value to their standard errors. The coefficient of the variable ("motherIQ") is the lowest and thus the one with the highest influence on the children's IQ. The value of R^2 (0.637) shows us that there is a particularly strong linear relationship between the response variable and the explanatory variables since the closer R^2 is to one the stronger the linear relationship, furthermore the value R^2 provides us with the clear cut explanation for this linearity since it basically means that 63.7 percent of the variance in the response variable can be explained by our model depicting these explanatory variables. In addition to that, the value of adjusted R^2 (0.603) is very close to R^2 which means that our data isn't very noisy and can be somewhat reliable, which is particularly important when working with small-sized samples. Since two of our p-values were within range and our R^2 was pretty high on a scale from 0 to 1 we can reject the null hypothesis that there isn't any kind of relationship between our explanatory variables and our main variable, although we can't accept the alternative hypothesis either since we still have one variable that is not statistically significant. There is however, one thing that has been made clear, and that is that the higher the mother's IQ is the more chances there are for the child to be gifted.(we can affirm that based on this variable's p-value and low coefficient).

3.3 2-Sample Unpaired T-Test

In this section, I am investigating the following research question: How does "the average hours per week the parents read to the child" influence their IQ score? The logic behind this question is that the longer a parent reads to the child, the more they will learn, memorize and analyse, and the better their cognitive skills will get. Thus, my expectation would be that the kids that get read to the most have a higher IQ score. In my study, the variable "the average hours per week the parents read to the child" consists of the two following categories:
A: Average of 2.1 hours per week or less
B: Above 2.1 hours per week

Before running the t-test, I first do some preliminary tests to check independent t-test assumptions. Let's run an F-test first to see if both samples have a common variance: The p-value is p = 0.7942. It's greater than the significance level alpha = 0.05. In conclusion, there is no significant difference between the variances of the two samples. Therefore, we can use the classic t-test which assumes equality of the two variances. Next let's see if both samples are normally distributed, we'll use the functions with() and shapiro.test() to compute Shapiro-Wilk test for each group of samples.
From the output, the two p-values are greater than the significance level 0.05 implying that the distribution of the data are not significantly different from the normal distribution. In other words, we can assume the normality.

Results of the 2-sample Unpaired t-test indicated that we have enough evidence to reject the null hypothesis that the mean IQ Score of gifted children in the two samples are equal as the p-value p = 0.01381 is smaller than 0.05.

```
> res <- t.test(data1$score, data2$score, var.equal = TRUE)
> res

        Two Sample t-test

data:  data1$score and data2$score
t = -2.5966, df = 34, p-value = 0.01381
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.6626465 -0.8123535
sample estimates:
mean of x mean of y
 157.0625  160.8000
```

4 Discussion

Based on the results of our multivariate linear regression analysis we can conclude that our model « explains » the higher IQ for children based on the independent variables that we chose ; there is a real statistical significance that can be established to link our main variable and our independent variables. Yet there are some limitations ; As we were writing this paper we were aware of the difficulties related to sampling, it is almost impossible to find the perfect dataset, sampling is one of the most important conditions of our statistical analysis since it will let us determine whether our study can be generalized to the whole population or if there any occurred biases that may mess with the objectivity of our results. One of the limitations of the study is the relatively small sample size which means that some assumptions of statistical inference were not respected ; indeed all of the participants of the survey live in the same city and are therefore subject to the same cultural and demographic conditions overall. Furthermore only children older than 4 years old were taken into account in our surveying, we do not know the parents' background and how often,if they do, push their kids towards intellectual activities ; all of these factors may have introduced a bias in our sample. Therefore it is safe to assume that our sampling lacks scrutiny and cannot be considered as a measurement for the entire population. As a result our study can legitimately be deemed observational rather than experimental, which means that this study could never identify the explanatory variables as direct causes for the high IQ score of the children but only prove correlation between them at best. We have found a somewhat decent R^2 in our regression analysis which affirms that this model is indeed a good fit for our hypothesis. However, one of our independent variables (time spent being read to) have no statistical significance whatsoever.
To conclude, this study has found evidence for the hypothesis that the mother's IQ and the time spent reading to kids can indeed predict a fairly large portion of the variance of the children's IQ, the result of the 2-sample unpaired T-test also shows that the children being read to for more than 2.1 hours on average have a significantly higher average IQ than the ones that are read to less than 2.1 hours per week . Furthermore, we have not found reliable evidence for the hypothesis that the time spent watching edutv has a significant impact on the variance of the children's IQ whatsoever. To conclude, it is not possible to affirm with precision that there is a significant relationship between the mother's IQ, the time spent being read to and the children's IQ that applies to the whole population, a better sample and more research are needed for that.

5 Sources

DATASET:
https://www.openintro.org/data/index.php?data=gifted
(will also be uploaded as a seperate file) Other sources:
https://www.davidsongifted.org/gifted-blog/what-is-giftedness/
https://www.investopedia.com/articles/trading/09/linear-regression-time-price.asp

1#Import and read data:

```
library(readr)

dataset <- read_csv("C:/Users/khawl/Desktop/dataset.csv")

View(dataset)
```

2#Clean data:we are filtering our data by removing the independent variables not considered in this study

```
install.packages("dplyr") #I installed a necessary package to use the function select()

library(dplyr)

dataset_filtered <- select(dataset, -c(fatheriq, speak, count,cartoons))

View(dataset_filtered)
```

3#data vizualisation:

```
#first i am naming some variable to make my code look clearer

x <-dataset_filtered$motheriq

y <-dataset_filtered$edutv

z <-dataset_filtered$score

t <-dataset_filtered$read

# code for 3Dscatterplot

#I installed a necessary package for the visualization of this plot

install.packages("scatterplot3d")

library(scatterplot3d)

scatterplot3d(x, y, z, angle = 20, xlab = "Mother's IQ" , ylab ="edutv (hrs/week)" , zlab="IQ score", pch = 16, color = "steelblue")

# other different plots:

plot(x, z, xlab="Mother's IQ" , ylab="kid's IQ score")

abline(lm(z ~ x, data = dataset_filtered), col = "blue") #adding a blue line to make things clearer

plot(y, z, xlab="edutv(hrs/week)" , ylab="kid's IQ score")
```

```
abline(lm(z ~ y, data = dataset_filtered), col = "blue")
plot(t,z, xlab = "read(hrs/week)" , ylab = "IQ score")
abline(lm(z ~ t, data = dataset), col = "blue")
```

4# Multivariate linear regression

```
# Multivariate Linear regression assumptions tests:
#code for residuals vs fitted plot
model <- lm(z~x+y+t, data=dataset_filtered)
res <- resid(model)
plot(fitted(model), res, xlab="predicted values")
abline(0,0) #adding the line
#code for a normal QQ plot
qqnorm(res)
qqline(res) #adding the line
# code for density plot of residuals
plot(density(res), xlab = "residuals")
# code for multivariate linear regression:
fit <- lm(z ~ x+ y + t, data=dataset_filtered)
summary(fit)
```

5# 2 sample Unpaired t-test

```
#first i make my subtables to create our two categories
> data1 <- dataset_filtered[t <= 2.1, ]
> View(data1)
> data2 <- dataset_filtered[t > 2.1, ]
> View(data2)
```

```
#code for the preliminary tests for the t-test assumptions
#shapiro test for each categorie to test the normality
with(data1, shapiro.test(score))
with(data2, shapiro.test(score))
#F-test to test if both categories have a common variance
var.test(data1$score,data2$score)
#t-test code
res <- t.test(data1$score, data2$score, var.equal = TRUE)
res
```