HACETTEPE UNIVERSITY

FACULTY OF ECONOMICS & ADMINISTRATIVE SCIENCES

DEPARTMENT of ECONOMICS

ECO 232

Computer Applications II

FINAL ASSIGNMENT

Data Analysis with R and Excel: “GIFTED” Dataset

Submitted to:

Nilgün Çokça

Büşra Güleryüz

2194168

ANKARA

May 2021

“GIFTED” DATASET DATA ANALYSIS

BUSRA GULERYUZ

2021-06-02

```
## Loading required package: splines

## Loading required package: RcmdrMisc

## Loading required package: car

## Loading required package: carData

## Loading required package: sandwich

## Loading required package: effects

## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

## The Commander GUI is launched only in interactive sessions

## 
## Attaching package: 'Rcmdr'

## The following object is masked from 'package:base':
## 
##     errorCondition

Warning in rgl.init(initValue, onlyNULL): Cannot create Freetype font

Warning in rgl.init(initValue, onlyNULL): font family "sans" not found, using
"bitmap"
```

Numerical Summaries

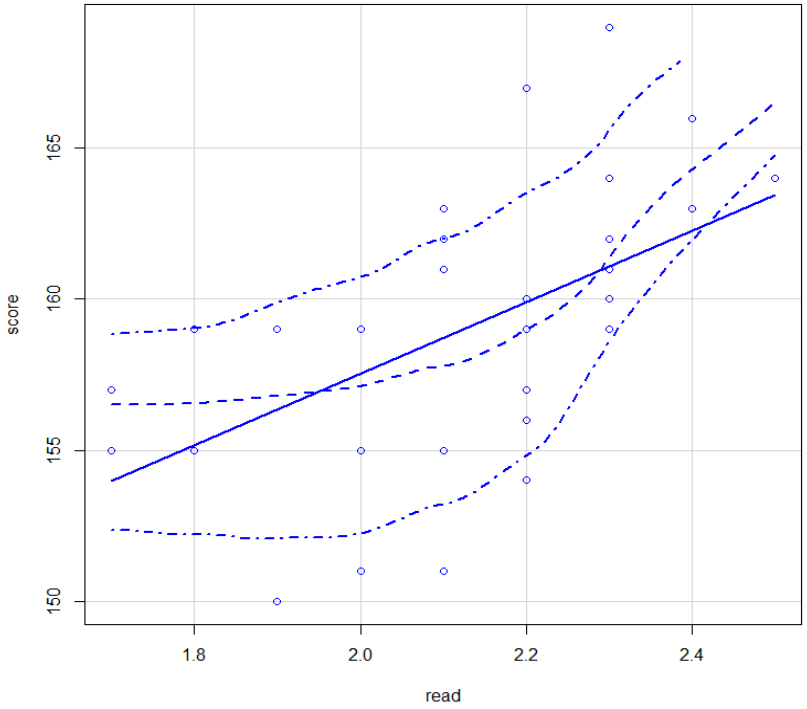
```
> Dataset <- readXL("C:/Users/Güleryüz/Desktop/assignment 2_gifted.xlsx",
+   rownames=FALSE, header=TRUE, na="", sheet="gifted", stringsAsFactors=TRUE
)

> summary(Dataset)

     score          fatheriq        motheriq         speak         count      
 Min.   :150.0   Min.   :110.0   Min.   :101.0   Min.   :10   Min.   :21.00  
 1st Qu.:155.0   1st Qu.:112.0   1st Qu.:113.8   1st Qu.:17   1st Qu.:28.00  
 Median :159.0   Median :115.0   Median :118.0   Median :18   Median :31.00  
 Mean   :159.1   Mean   :114.8   Mean   :118.2   Mean   :18   Mean   :30.69  
 3rd Qu.:162.0   3rd Qu.:116.2   3rd Qu.:122.2   3rd Qu.:20   3rd Qu.:34.25  
 Max.   :169.0   Max.   :126.0   Max.   :131.0   Max.   :23   Max.   :39.00  
      read           edutv          cartoons    
 Min.   :1.700   Min.   :0.750   Min.   :1.750  
 1st Qu.:2.000   1st Qu.:1.750   1st Qu.:2.688  
 Median :2.200   Median :2.000   Median :3.000  
 Mean   :2.136   Mean   :1.958   Mean   :3.062  
 3rd Qu.:2.300   3rd Qu.:2.250   3rd Qu.:3.500  
 Max.   :2.500   Max.   :3.000   Max.   :4.500  

> cor(Dataset)

              score    fatheriq    motheriq       speak       count          re
ad
score     1.0000000  0.18808106  0.57124196  0.26789109  0.54420658  0.525197
37
fatheriq  0.1880811  1.00000000 -0.02481170 -0.03053753 -0.07502148 -0.068218
60
motheriq  0.5712420 -0.02481170  1.00000000  0.07218511  0.02426072 -0.043030
65
speak     0.2678911 -0.03053753  0.07218511  1.00000000  0.05954480  0.185071
29
count     0.5442066 -0.07502148  0.02426072  0.05954480  1.00000000  0.910251
91
read      0.5251974 -0.06821860 -0.04303065  0.18507129  0.91025191  1.000000
00
edutv    -0.3702598  0.11622154 -0.32999908 -0.15452363 -0.21567890 -0.166562
37
cartoons  0.2451001 -0.24835135  0.33841771  0.10936054  0.15490093  0.125733
88
              edutv   cartoons
score    -0.3702598  0.2451001
fatheriq  0.1162215 -0.2483514
motheriq -0.3299991  0.3384177
speak    -0.1545236  0.1093605
count    -0.2156789  0.1549009
read     -0.1665624  0.1257339
```

```
edutv     1.0000000 -0.9234370
cartoons -0.9234370  1.0000000
```

- As can be seen above, minimum and maximum values for each variable are versatile. About the correlation matrix; it shows the correlation between each variable. The variable “score” has a positive correlation with each variable, except for “edutv” variable. Considering that edutv implies the average number of hours that a child watches an educational TV show in a week and score is the score in test of analytical skills, we can say that watching TV educational TV shows reduced the kid’s analytical skills, but this is kind of surprising because watching cartoons caused the opposite effect.

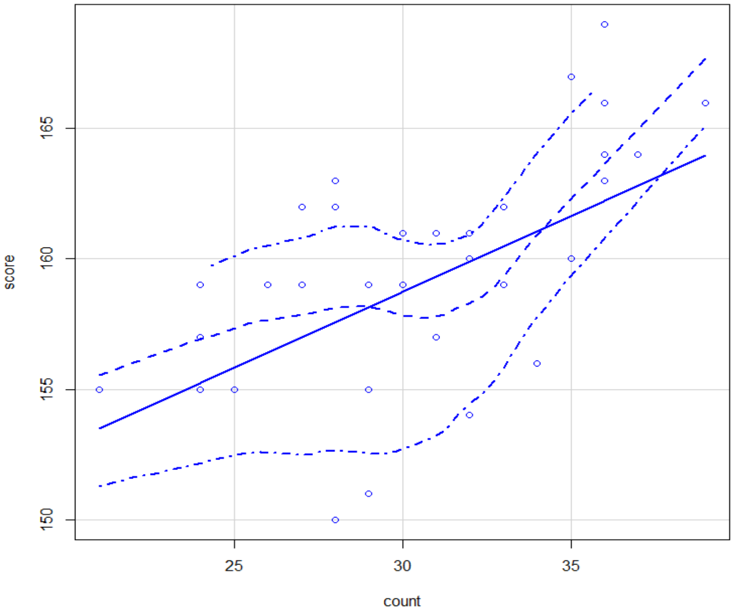
Scatterplots

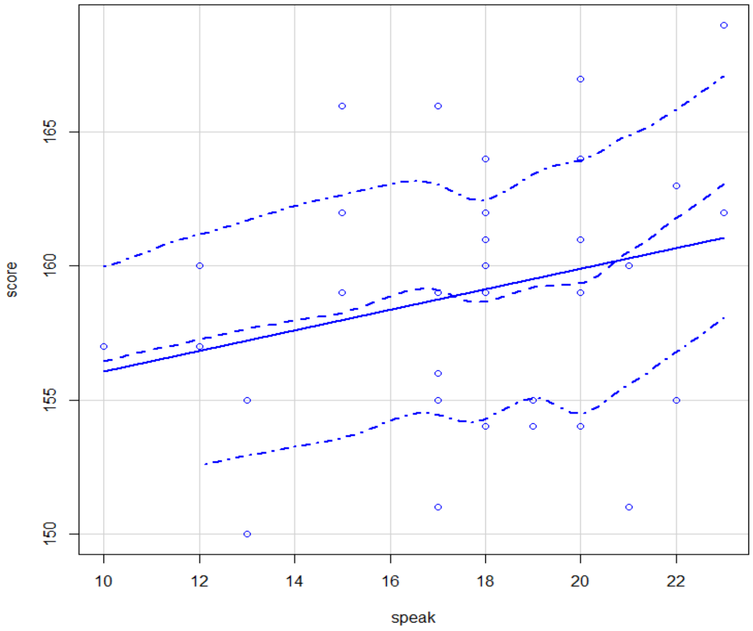
```
> scatterplot(score~read, regLine=TRUE, smooth=TRUE, boxplots=FALSE,
+   data=Dataset)
```



- Read variable (Average number of hours per week the child’s mother or father reads to the child) and the score have a positive relationship. It obviously increases the child’s analytical skills and could be increasing their intelligence level as well. This can be because when someone reads to a child, their imagination skills develop.

```
> scatterplot(score~count, regLine=TRUE, smooth=TRUE, boxplots=FALSE,
+   data=Dataset)

> scatterplot(score~speak, regLine=TRUE, smooth=TRUE, boxplots=FALSE, data=Da
taset)
```

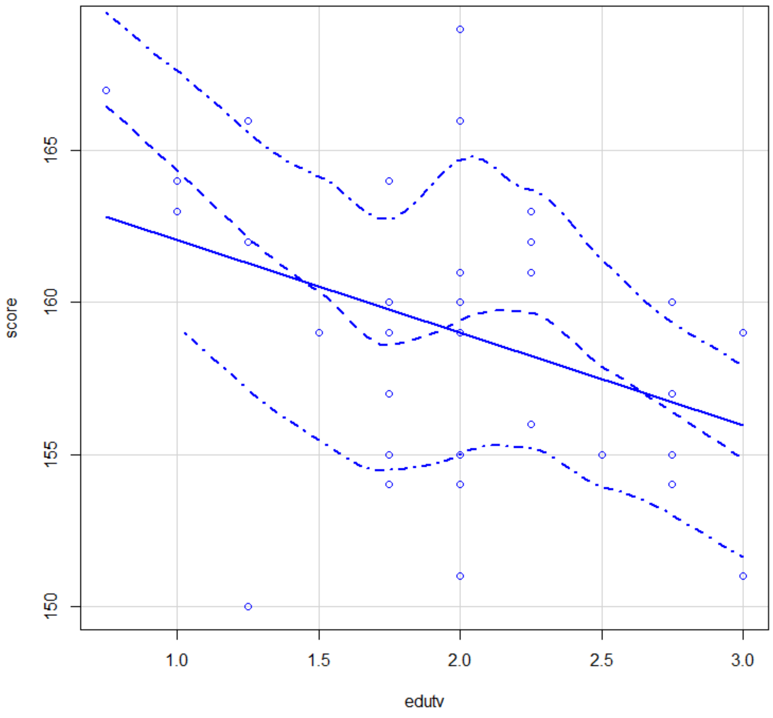


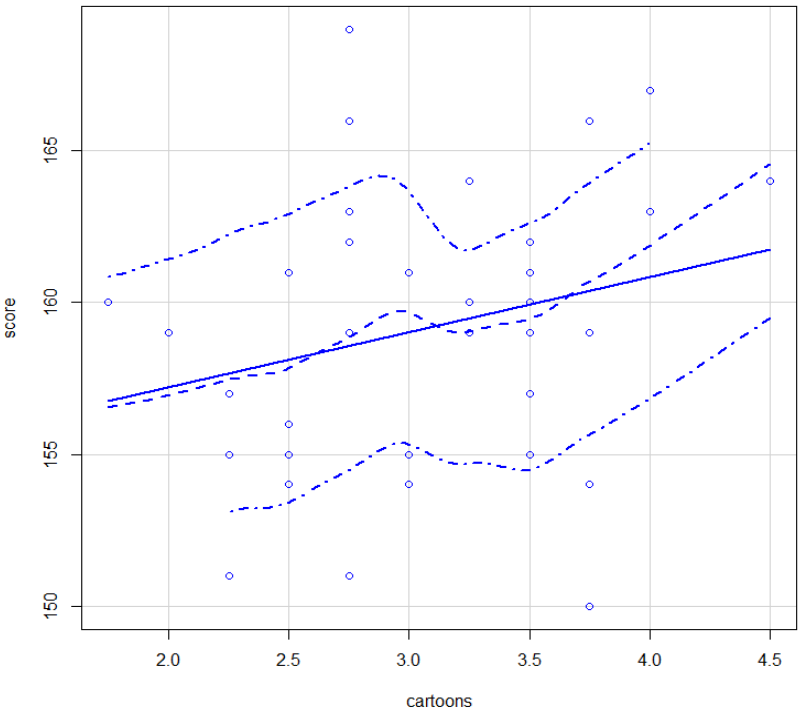


- Count variable (Age in months when the child first counted to 10 successfully) and the score have a positive relationship. It is not very hard to guess that, if a child starts counting earlier than the others, they would turn out to be more skilled analytically. With the same logic, speak variable (Age in months when the child first said “mummy” or “daddy”) has a positive relationship with the score as well.

```
> scatterplot(score~edutv, regLine=TRUE, smooth=TRUE, boxplots=FALSE,
+   data=Dataset)
```

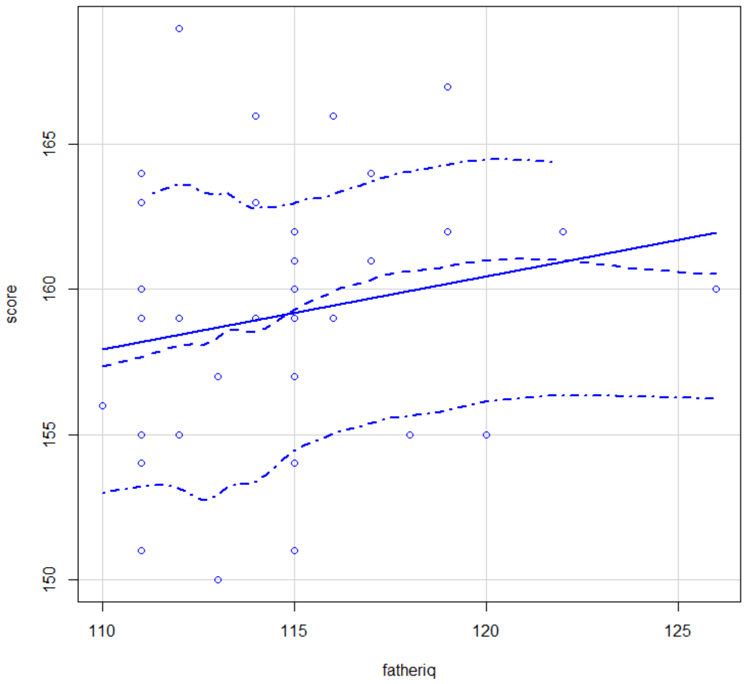
```
> scatterplot(score~cartoons, regLine=TRUE, smooth=TRUE, boxplots=FALSE,
+   data=Dataset)
```

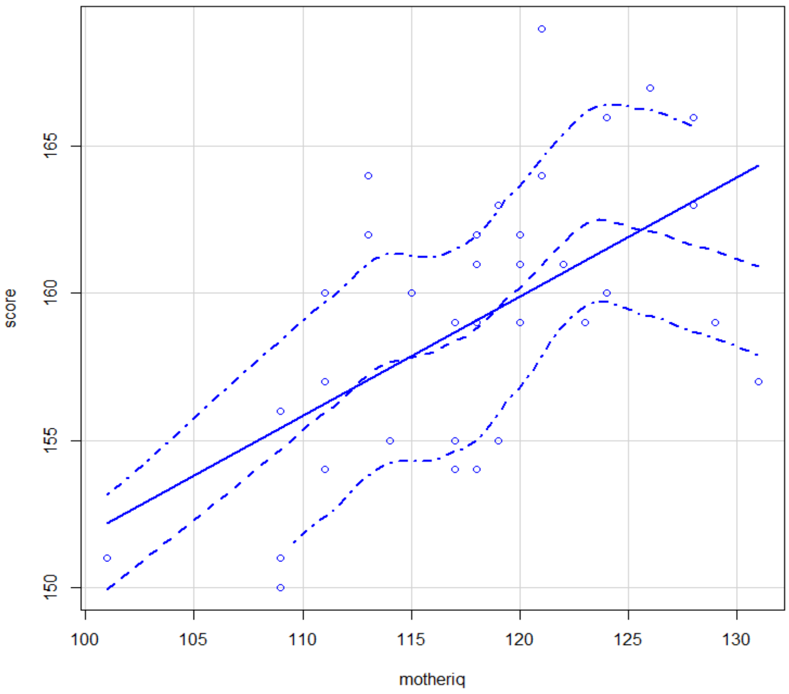




- Edutv variable (Average number of hours per week the child watched an educational program on TV during the past three months.) and the score have a negative relationship while cartoons (Average number of hours per week the child watched cartoons on TV during the past three months.) have it positively. It is very interesting, I think; apparently educational TV shows are not only "not effective" on the child's intelligence - or analytical skills, but it also has a negative effect upon the kid. On the other hand, watching cartoons improves the child's skills. We can conclude that watching cartoons may have a positive effect on the child because it develops the kid's imaginary skills, and educational TV shows might be boring to the kids, therefore not have any positive effect.

```
> scatterplot(score~motheriq, regLine=TRUE, smooth=TRUE, boxplots=FALSE,
+   data=Dataset)

> scatterplot(score~fatheriq, regLine=TRUE, smooth=TRUE, boxplots=FALSE,
+   data=Dataset)
```





- Comparing the effect of father’s and mother’s IQ levels on the child’s analytical skills, we can see that mother’s IQ has a steeper line meaning that it is much more effective. This is probably because the “intelligence” is an x-linked gene, it directly transfers from the mother to the child regardless of its gender, but father can only transfer his “intelligence” gene to his daughter, since he can only transfer Y chromosome to his son.

Regression Analysis

Explanatory variable: cartoons. Response variable: score. I picked these variables because I found it very interesting for watching cartoons to have a positive effect on the child's analytical skills.

```
> Model1 <- lm(score~cartoons, data=Dataset)
> summary(Model1)


Call:
lm(formula = score ~ cartoons, data = Dataset)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3815  -2.9778   0.2481   2.9667  10.4259 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  153.603      3.831  40.095   <2e-16 ***
cartoons       1.807      1.226   1.474     0.15    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.554 on 34 degrees of freedom
Multiple R-squared:  0.06007,   Adjusted R-squared:  0.03243 
F-statistic: 2.173 on 1 and 34 DF,  p-value: 0.1496
```

- The intercept is 153.6, meaning that the kid who watched zero hours of cartoons for 3 months had the score of 153.6. Slope is 1.8, meaning that for each additional hour that a child watches cartoon, their test score increases for 1.8 units. This is very interesting. Apparently, watching cartoons is also helpful for the child's mental development, besides entertaining them.

- The coefficient Standard Error measures the average amount that the coefficient estimates vary from the actual average value of our response variable. We want this to be small for the consistency of our results. In this example, it is 1.226, which is pretty small of a number.

- The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want the t-value to be far from zero as this would indicate that we could reject the null hypothesis, meaning that we could actually declare a relationship between age and circumference. In our example, t-values are far from zero. t-values are also used in calculating the p-values.

- The Pr(>t) acronym found in the model relates to the probability of observing any value equal or larger than t. A small p-value tells us that it is unlikely to observe a relationship between the predictor (cartoons) and response (score) variables. A small p-value for the intercept and the slope proves that we can reject the null hypothesis, and this allows us to declare a relationship between age and circumference.

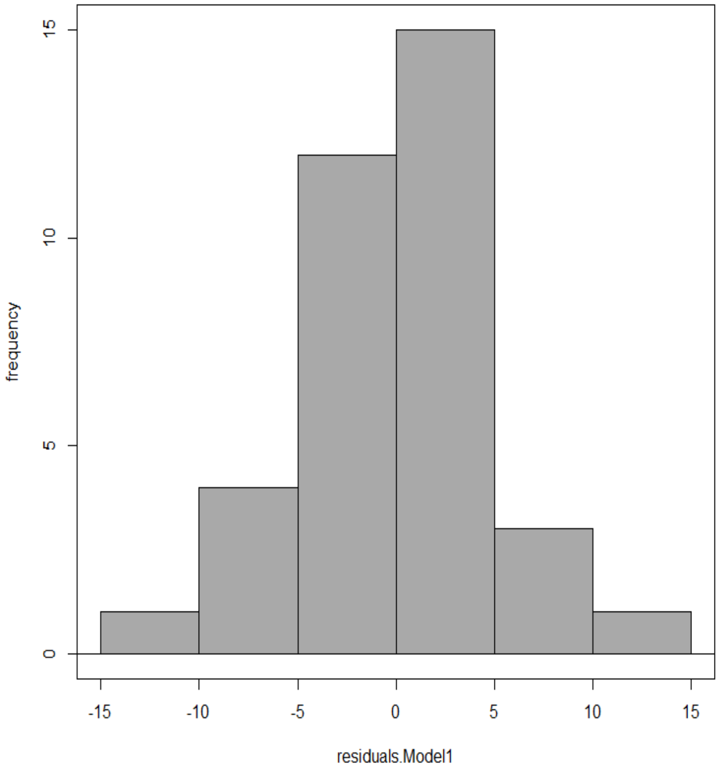
- ➢ Residual standard error is measure of the quality of a linear regression fit. Every linear model is assumed to have an error term "E". because of the existence of this error term, we are not capable of perfectly predicting our response variable (score) from the predictor (cartoons). The residual standard error is 4.554 on 34 degrees of freedom.

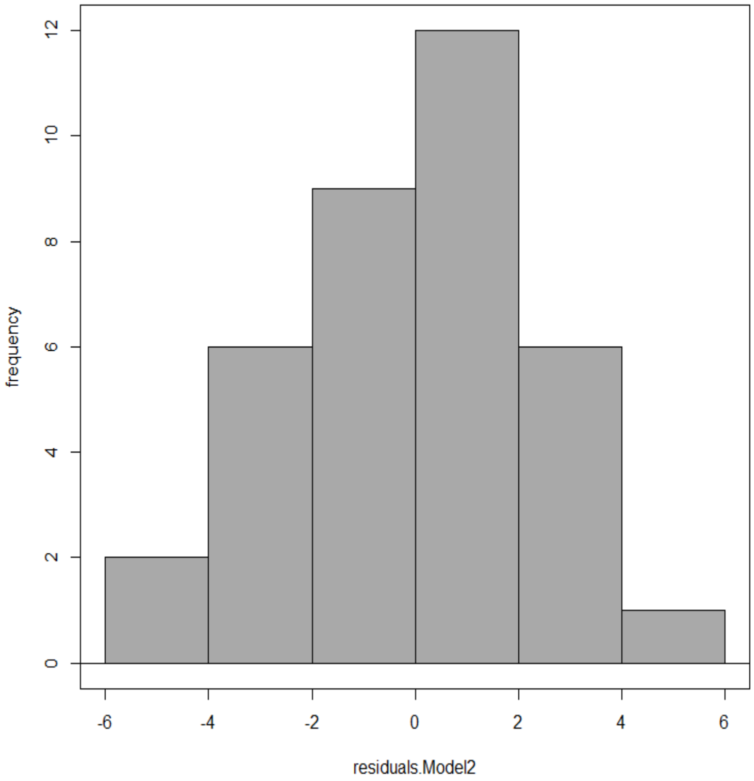
Multivariate Regression Analysis

```
> Model2 <- lm(score~cartoons+count+edutv+fatheriq+motheriq+read+speak,
+   data=Dataset)
> summary(Model2)


Call:
lm(formula = score ~ cartoons + count + edutv + fatheriq + motheriq + 
    read + speak, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8064 -1.5898  0.0479  1.7474  5.2905 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 75.50849   24.02618   3.143  0.00393 ** 
cartoons    -3.33899    2.01808  -1.655  0.10919    
count        0.20649    0.26631   0.775  0.44462    
edutv       -4.20244    2.24503  -1.872  0.07170 .  
fatheriq     0.25249    0.13756   1.835  0.07707 .  
motheriq     0.40007    0.07291   5.488 7.33e-06 ***
read         7.54405    5.58640   1.350  0.18769    
speak        0.18764    0.14767   1.271  0.21429    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.591 on 28 degrees of freedom
Multiple R-squared:  0.7496,    Adjusted R-squared:  0.687 
F-statistic: 11.97 on 7 and 28 DF,  p-value: 5.803e-07
```

- ➢ This model provides a regression analysis for each variable.

```
> Dataset<- within(Dataset, {
+   residuals.Model1 <- residuals(Model1)
+ })

> with(Dataset, Hist(residuals.Model1, scale="frequency", breaks="Sturges",
+   col="darkgray"))

> Dataset<- within(Dataset, {
+   residuals.Model2 <- residuals(Model2)
+ })> with(Dataset, Hist(residuals.Model2, scale="frequency", breaks="Sturges
",
+   col="darkgray"))
```





- ➢ The Histogram of the residuals are made to detect whether the variance is normally distributed or not. A symmetric histogram (which is evenly distributed around zero) indicates that the normality assumption is likely to be true. In our case, histograms are very close to be symmetric. This means that our regression models have turned out to be very useful and close values to the truth.