Linear Analysis and Model choice

Yuhong Chen

First of all, for the question statement, we have a data from the performance of the students, which contains various data about the students in the school, such as gender, ethnicity, education, math scores, reading scores, writing scores, etc. We want to explore the relationship between these data, study whether the characteristics of students or scores are related to other scores by linear regression analysis.

Linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. It measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques. (Kumari, K., & Yadav, S. ,2018).

Now we begin to analyze the data.

See the summary statistics of the data

```
summary(StudentsPerformance)
```

```
##     gender          race/ethnicity     parental level of education
##  Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     lunch           test preparation course   math score     reading score
##  Length:1000        Length:1000             Min.   :  0.00   Min.   : 17.00
##  Class :character   Class :character        1st Qu.: 57.00   1st Qu.: 59.00
##  Mode  :character   Mode  :character        Median : 66.00   Median : 70.00
```

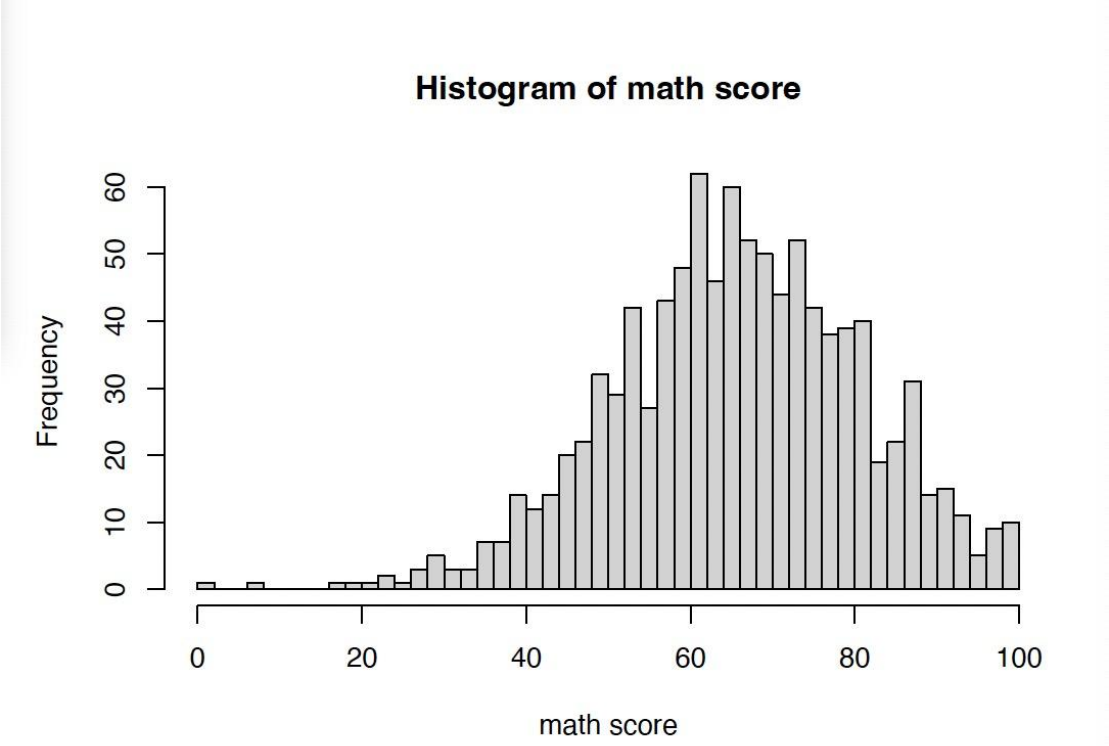


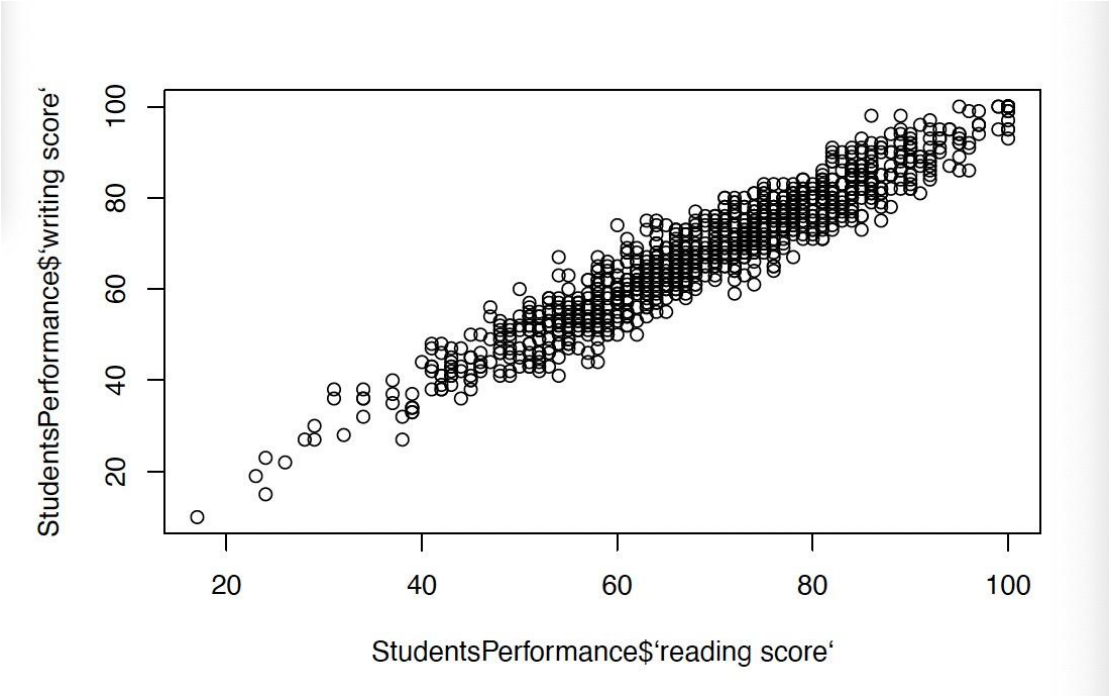
```
##                                          Mean   : 66.09   Mean   : 69.17
##                                          3rd Qu.: 77.00   3rd Qu.: 79.00
##                                          Max.   :100.00   Max.   :100.00
##  writing score
##  Min.   : 10.00
##  1st Qu.: 57.75
##  Median : 69.00
##  Mean   : 68.05
##  3rd Qu.: 79.00
##  Max.   :100.00
```

We can see that every columns have their own data, like the quantitative data,the 4 different score of their min,1st quantile median, mean and max.

Now we have the visualization for the data.

Visualization is a powerful mechanism for extracting information from data. ggplot2 is a contributed visualization package in the R programming language, which creates publication-quality statistical graphics in an efficient, elegant, and systematic manner. (Ito, K., & Murphy, D. ,2013).





For these plots, we could see the math score histogram and reading&writing plot, with their relationship. Also for this part we could use ggplot in tidyverse to have more analysis.

Now we could define the model, first is null model, which

contains only the dependent variable and intercept.

```
n0 <- lm(formula =`math score`~1,data = StudentsPerformance) # fit the model
summary(n0)
```

```
##
## Call:
## lm(formula = `math score` ~ 1, data = StudentsPerformance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.089  -9.089  -0.089  10.911  33.911
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.0890     0.4795   137.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.16 on 999 degrees of freedom
```

```
# Note that the expected value is actually the mean and that the residual error is actually
# standard deviation
mean(StudentsPerformance$`math score`);sd(StudentsPerformance$`math score`)
```

```
## [1] 66.089
```

```
## [1] 15.16308
```

```
# We can use other variables to improve the model and reduce errors
```

So we could use some other variables to improve the model and reduce errors.

Now we Define a simple linear model with one explanatory variable.

```r
#Basic model
m1 <- lm(formula = `math score` ~ `writing score`,data = StudentsPerformance)
summary(m1)
```



```
##
## Call:
## lm(formula = `math score` ~ `writing score`, data = StudentsPerformance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.8467  -6.4600   0.1464   6.4356  25.5515
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.58310    1.31369   8.817   <2e-16 ***
## `writing score`  0.80092    0.01884  42.511   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.049 on 998 degrees of freedom
## Multiple R-squared:  0.6442, Adjusted R-squared:  0.6439
## F-statistic:  1807 on 1 and 998 DF,  p-value: < 2.2e-16
```

Note that the intercept predicts math score when writing score is 0. The slope measures expected change in weight for 1 unit change in height. We need to center explanatory variables to achieve a more meaningful coefficient.

Note that slope is actually correlation coefficient adjusted for relative dispersion of the two variables.

So now we use model with writing squared and cubed.

```r
math<- StudentsPerformance$`math score`
writing<-StudentsPerformance$`writing score`


m2 <- lm(formula = math~I((writing/100)^2),data = StudentsPerformance)
summary(m2)
```

```
##
## Call:
## lm(formula = math ~ I((writing/100)^2), data = StudentsPerformance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.982  -6.486  -0.285   6.432  26.855
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         37.3918     0.7669   48.76   <2e-16 ***
## I((writing/100)^2)  59.0232     1.4560   40.54   <2e-16 ***
## ---
```



```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.325 on 998 degrees of freedom
## Multiple R-squared:  0.6222, Adjusted R-squared:  0.6218
## F-statistic:  1643 on 1 and 998 DF,  p-value: < 2.2e-16
```

```r
m3 <- lm(formula = math~I((writing/100)^3),data = StudentsPerformance)
summary(m3)
```

```
##
## Call:
## lm(formula = math ~ I((writing/100)^3), data = StudentsPerformance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.773  -6.278  -0.184   6.679  28.091
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         46.7192     0.6008   77.76   <2e-16 ***
## I((writing/100)^3)  53.6167     1.4267   37.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.762 on 998 degrees of freedom
## Multiple R-squared:  0.5859, Adjusted R-squared:  0.5855
## F-statistic:  1412 on 1 and 998 DF,  p-value: < 2.2e-16
```

Upon this, we use information criteria to choose the best model. Information criteria such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are commonly used for model selection.

Furthermore, we find that the computational complexity of normalized information criteria methods is exponentially better than that of imputation methods. In a series of simulation studies, we find that normalized-AIC and normalized-BIC outperform previous methods (i.e., normalized-AIC is more efficient, and normalized BIC includes only important variables, although it tends to exclude some of them in cases of large correlation).

(Nitzan, C., & Yakir, B. , 2021).

As we know, smaller of AIC indicate better model.

```
AIC(m1,m2,m3)
```

```
##    df      AIC
## m1  3 7247.121
## m2  3 7307.288
## m3  3 7398.860
```

From above, we choose the model m1 with the smallest AIC.

Now for the mutiple regression model,

```
read <- StudentsPerformance$`reading score`
m4 <- lm(formula = math~writing+read,data = StudentsPerformance)
summary(m4)
```

```
##
## Call:
## lm(formula = math ~ writing + read, data = StudentsPerformance)
##
```



```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.8779  -6.1750   0.2693   6.0184  24.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.52409    1.32823   5.665 1.93e-08 ***
## writing      0.24942    0.06057   4.118 4.14e-05 ***
## read         0.60129    0.06304   9.538  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.667 on 997 degrees of freedom
## Multiple R-squared:  0.674,  Adjusted R-squared:  0.6733
## F-statistic:  1031 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
m5 <- lm(formula = math~writing*read,data = StudentsPerformance)
summary(m5)
```

```
##
## Call:
## lm(formula = math ~ writing * read, data = StudentsPerformance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.052  -6.254   0.230   6.009  24.705
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.7614864  4.1071551   0.916  0.35997
## writing       0.3076971  0.0853945   3.603  0.00033 ***
## read          0.6617578  0.0887452   7.457 1.92e-13 ***
## writing:read -0.0008916  0.0009210  -0.968  0.33321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.667 on 996 degrees of freedom
## Multiple R-squared:  0.6743, Adjusted R-squared:  0.6733
## F-statistic: 687.3 on 3 and 996 DF,  p-value: < 2.2e-16
```

Same we use AIC to compare the 2 models.

```
AIC(m4,m5)
```

```
##    df      AIC
## m4  4 7161.804
## m5  5 7162.863
```

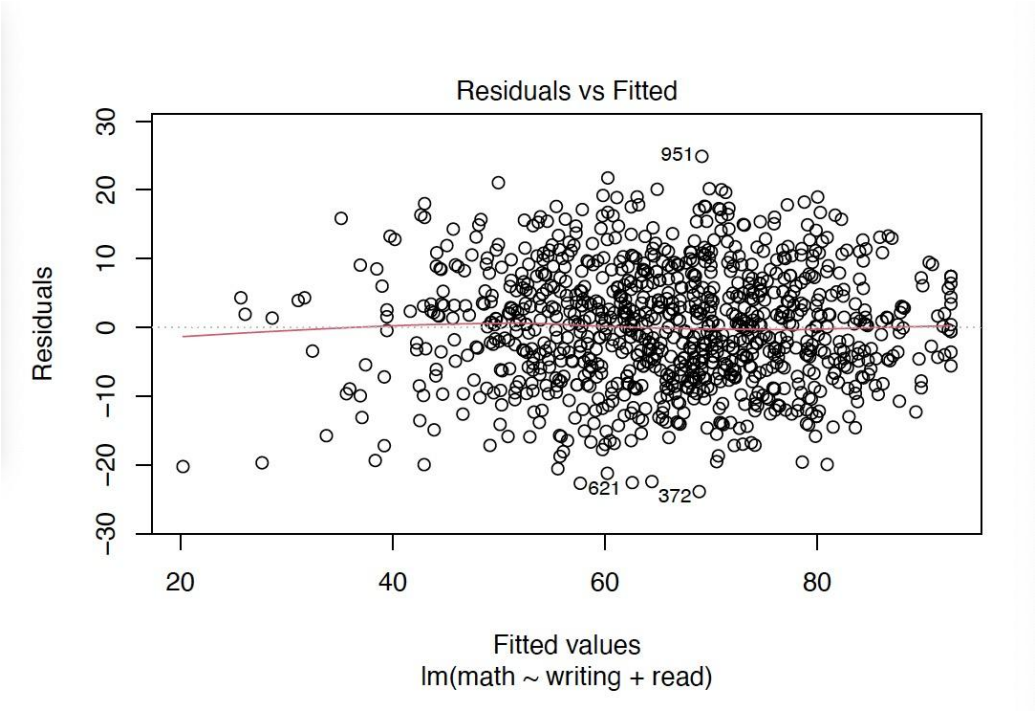
We see from the AIC , the interaction of terms do not improve the model And we compare the models above.Now we compare the models from m1 to m5.

```
AIC(m1,m2,m3,m4,m5)
```

```
##    df      AIC
## m1  3 7247.121
## m2  3 7307.288
## m3  3 7398.860
## m4  4 7161.804
## m5  5 7162.863
```

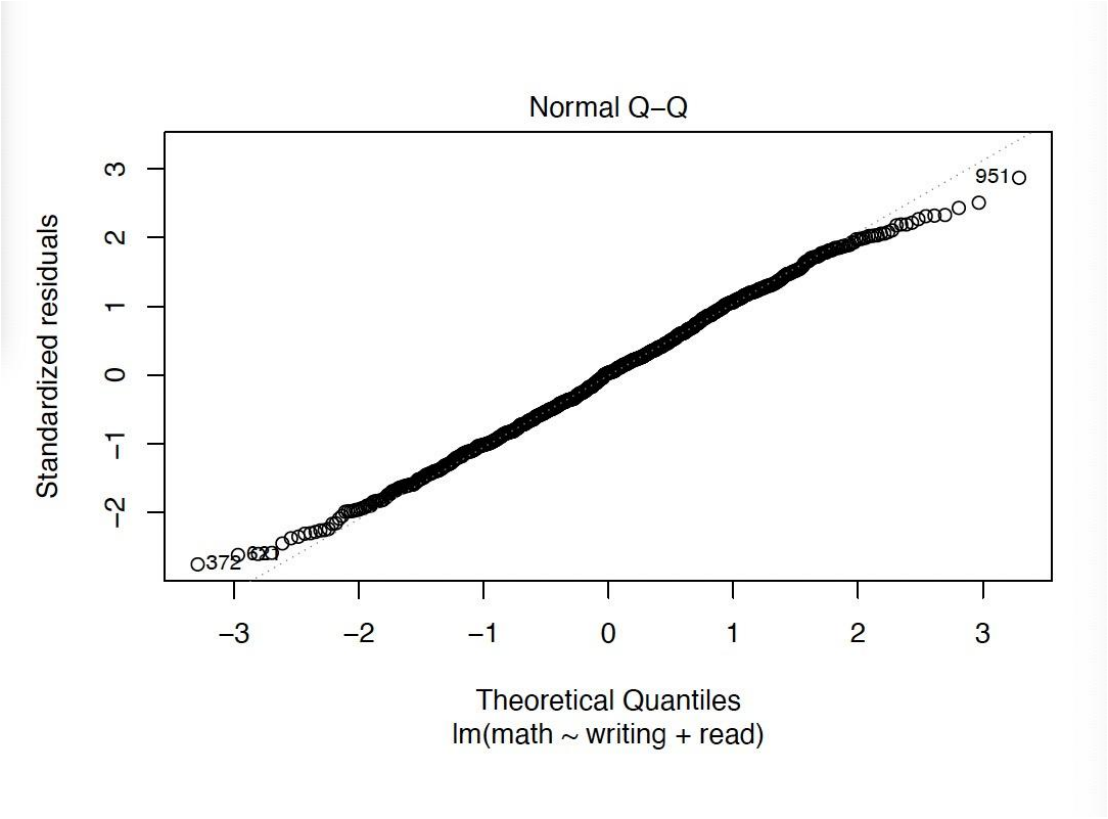
We can see that m4 has the smallest AIC value so the final model is m4 with #{math~writing+read}.

So now we plot the model. We see from the residual VS fitted: this satisfy the linear relationship.

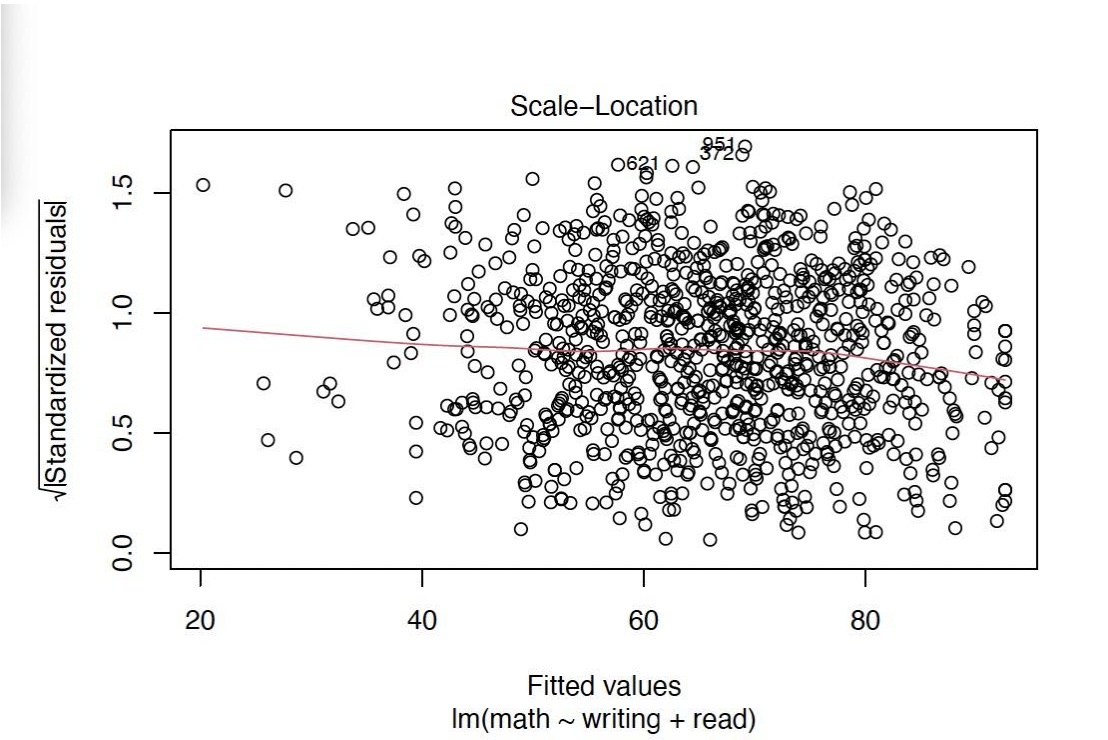


Now for Normal QQ plot. It shows normal distribution.

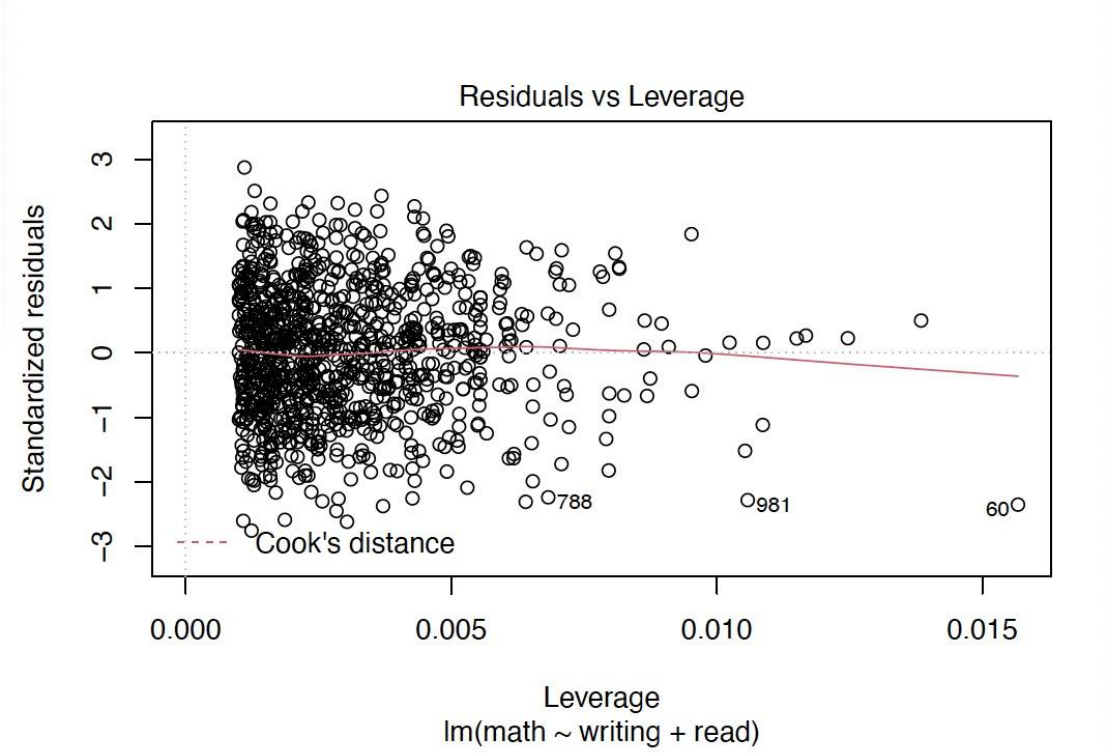
In statistics, a normal distribution or Gaussian distribution is a type of continuous probability distribution for a real-valuedrandom variable. A random variable with a Gaussian distribution is said to be normally distributed, and is called a normal deviate. (Earliest uses of symbols in probability and statistics. (n.d.).)



Now for variance, we see there is a average distribution between the red line.



This is the sr vs leverage.



Now we can make the prediction with regressions.

```
parent<- StudentsPerformance$`parental level of education`
predict(m4,newdata=data.frame(writing=79,read=60),
        type='response')
```

```
##        1
## 63.30597
```

```
predict(m4,newdata=data.frame(writing=79,read=60),
        interval=c('prediction'),level=0.95)
```

```
##        fit      lwr      upr
## 1 63.30597 46.12128 80.49066
```

```
predict(m4,newdata=data.frame(writing=79,read=60),
        interval=c('confidence'),level=0.95)
```



```
##        fit     lwr      upr
## 1 63.30597 60.8389 65.77304
```

The prediction interval is the range in which future random observation can be thought most likely to occur,
whereas the confidence interval is where the mean of future observation is most likely to reside.
The main characteristics of each method
for prediction interval construction are presented and some recommendations are given for selecting the most appropriate method for specific applications. (Cartagena, O., Parra, S., Munoz-Carpintero, D., Marin, L. G., & Saez, D. , 2021.)

The confidence interval is generally much more narrow than the prediction interval and its “narrowness” will increase with increasing numbers of observations, whereas the prediction interval will not decrease in width. So we can
use the model 4 to predict the math score using the two variables which are reading and writing scores with the 95% confidence interval.

In conclusion, from these analysis using linear model, AIC comparation, also some standard deviation and some other data, could help us select the model and make prediction among different variables, and help us to explore the behaviors and problems we are curious.

References:

Kumari, K., & Yadav, S. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences, 4(1), 33–36. https://doi.org/10.4103/jpcs.jpcs_8_18

Ito, K., & Murphy, D. (2013). Application of ggplot2 to pharmacometric graphics. Cpt: Pharmacometrics & Systems Pharmacology, 2(10), 1–16. https://doi.org/10.1038/psp.2013.56

Nitzan, C., & Yakir, B. (2021). Normalized information criteria and model selection in the presence of missing data, 9(2474), 2474–2474. https://doi.org/10.3390/math9192474

Earliest uses of symbols in probability and statistics. (n.d.).https://jeff560.tripod.com/stat.html

Cartagena, O., Parra, S., Munoz-Carpintero, D., Marin, L. G., & Saez, D. (2021). Review on fuzzy and neural prediction interval modelling for nonlinear dynamical systems. Ieee Access, 9. https://doi.org/10.1109/ACCESS.2021.3056003