# **Multicollinearity**

As a data-scientist most of the time we have to work on some unknown problems or data. I am cricket fan and what if someone gave me data of baseball for analysis. It is very important that My inability to understand different terms in the data due to my lack of knowledge about the sport shouldn’t impact the model /analysis. How can we make sure that we can avoid some of the common mistakes while building or planning for our model? Few things that makes a model a parsimonious.

A parsimonious model is a model that accomplishes a desired level of explanation or prediction with as few predictor variables as possible. They explain data with a minimum number of parameters, or predictor variables.

We know the four assumption of Linear Regression lm(X~y) :

Linearity: The relationship between X and the mean of Y is linear.

Homoscedasticity: The variance of residual is the same for any value of X.

Independence: Observations are independent of each other.

Normality: For any fixed value of X, Y is normally distributed.

Multicollinearity: Multicollinearity (also collinearity) is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a non-trivial degree of accuracy.

Multicollinearity is a problem because it undermines the statistical significance of an independent variable as other variables are correlated, higher degree of correlation between variables can’t guarantee that our model can better explain all the variations of the data.

Here I have data about Kids score depending upon moms age and if mom were working(mom\_work) and Mom\_iq. Let’s try to fit the model based on this data.

**Response:** - Kid\_score

**Predicotr**:- mom\_hs , mom\_iq , mom\_work, mom\_age

cognitive <- read.csv("http://bit.ly/dasi\_cognitive")

head(cognitive)

## kid\_score mom\_hs mom\_iq mom\_work mom\_age

## 1 65 yes 121.11753 yes 27

## 2 98 yes 89.36188 yes 25

## 3 85 yes 115.44316 yes 27

## 4 83 yes 99.44964 yes 25

## 5 115 yes 92.74571 yes 27

## 6 98 no 107.90184 no 18

Here I am going to create some colinear predictors then we will see impact of those eon the model.

## # Model1

cog\_full = lm(kid\_score~.,data = cognitive)

summary(cog\_full)

## # Adding new Predictors

cognitive$c\_ageiq <- cognitive$mom\_age\*cognitive$mom\_iq

cognitive$c\_iq <- cognitive$mom\_iq^2

head((cognitive))

## kid\_score mom\_hs mom\_iq mom\_work mom\_age c\_ageiq c\_iq

## 1 65 yes 121.11753 yes 27 3270.173 14669.456

## 2 98 yes 89.36188 yes 25 2234.047 7985.546

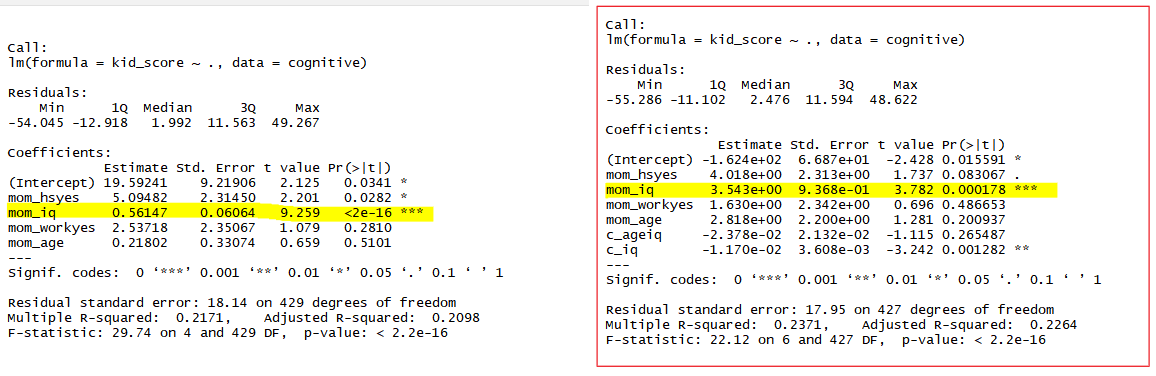
## 3 85 yes 115.44316 yes 27 3116.965 13327.124

## 4 83 yes 99.44964 yes 25 2486.241 9890.231

## 5 115 yes 92.74571 yes 27 2504.134 8601.767

## 6 98 no 107.90184 no 18 1942.233 11642.807

## # Side by Side Mode 1 and Model 2



As we can see the two model1(cog\_full) and Model2 (cog\_full2),

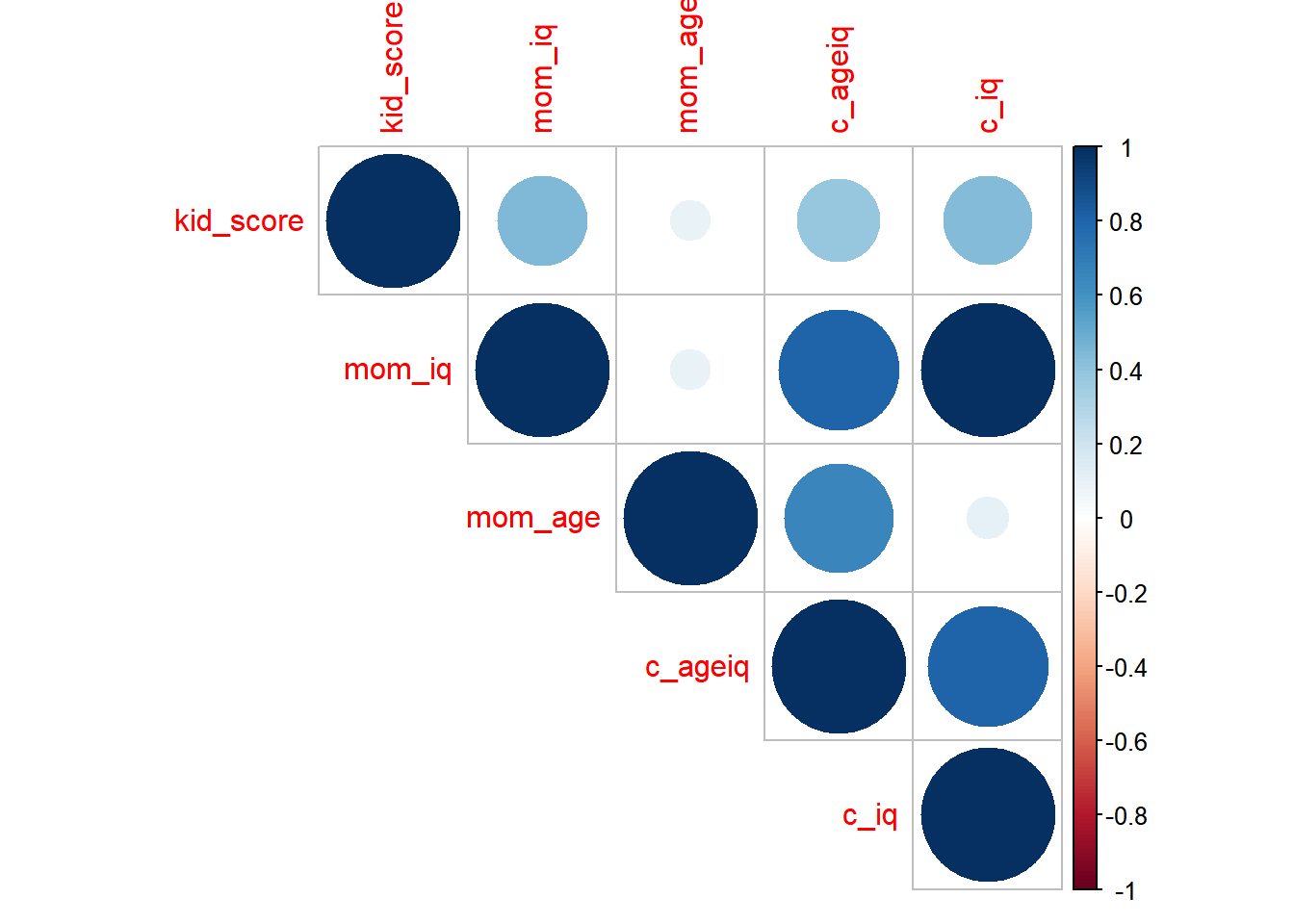
| **Model Name** | **R-squared** | **Adjusted R-squared** | **Significant** | **standard error** |
| --- | --- | --- | --- | --- |
| Model1 | 0.2171 | 0.2098 | mom\_hsyes, mom\_iq | 18.14 |
| Model2 (c) | 0.2371 | 0.2264 | mom\_iq,c\_iq | 17.95 |

From the above two model we see the coefficients of mom\_iq is showing very high std error and its p-value has also gone down. Why we see this problem, it’s because of existing correlated predictor c\_ageiq and c\_iq.

How to catch such collinearity which is present in data and not easy to locate, We can use two ways to test theses:

1. Using the Correlation among the predictors.
2. VIF (variance inflation factor), which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model. The smallest possible value of VIF is one (absence of multicollinearity). Here we will look for VIF value, if that exceeds 5 or 10 indicates a problematic amount of collinearity. [Read More](http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/).

library(corrplot)  
corrplot(cor(cognitive[,-c(2,4)],use="pairwise.complete.obs"),type = 'upper')



Correlation among variables:

cor(cognitive[,-c(2,4)])[,c(2,5)]

## mom\_iq c\_iq

## kid\_score 0.4482758 0.4354671

## mom\_iq 1.0000000 0.9968604

## mom\_age 0.0916084 0.1020082

## c\_ageiq 0.8045191 0.8077851

## c\_iq 0.9968604 1.0000000

Above plots and correlation matrix, very clearly shows that Mom\_iq ia colinear with c\_iq and c\_ageiq

Now we will see Variance Inflation Factors for each predictor in the model.

**library**(car)

vif(cog\_full2)

## mom\_hs mom\_iq mom\_work mom\_age c\_ageiq c\_iq

## 1.212658 265.287084 1.077794 47.457234 132.212711 169.667177

Here again the value from Correlation matrix and Variance Inflation Factors, is giving same result with some correlation that exist between predictors: mom\_iq , mom\_iq , c\_ageiq , c\_iq

In contrast if we check the inflation factor for the Model1 we see that none of the predictors are correlated. Which is good sign for further improving this model.

vif(cog\_full)

## mom\_hs mom\_iq mom\_work mom\_age

## 1.189102 1.088349 1.063187 1.049762

High Variance Inflation Factor (VIF) and Low Tolerance: These two useful statistics are reciprocals of each other. So either a high VIF or a low tolerance is indicative of multicollinearity. VIF is a direct measure of how much the variance of the coefficient (ie. its standard error) is being inflated due to multicollinearity.

Ref:

<https://www.theanalysisfactor.com/eight-ways-to-detect-multicollinearity/>

<http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/>

<http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html>