FinalReport

Gautam Gireesh, Saakshi Shah

22/08/2021

Introduction

As aspiring statisticians, the goal of this study for us was to develop a reasonable regression model using the techniques we have learnt over the course of this semester. Through this, we would be able to understand and determine what factors do in fact impact a student's performance on quiz 4. At a glance, it would be easy for us to assume that all the aspects - country, quiz grades, hours spent on COVID and stats - must make an impact. But having been students of this course, we know better and will put our intuition to the test to determine whether or not all the factors make an impact. The following report will conduct an in-depth analysis of the data and explain how and why we chose the model.

Choice of Method

The method we thought would work best for our analysis was forward selection. With the forward selection method, we add predictor variables one after the other and examine the SSres, Rˆ2, Rˆ2 adj and AIC at every stage to determine whether or not the added component substantially improves or worsens our model. This includes examining whether the SSres and AIC have decreased and whether the Rˆ2 and Rˆ2 adj have increased.

For us, the first step involved examing the response variable against our respective three predictor variables:

- X1 = COVID hours average
- X2 = Stats studying hours average
- X3 = Quiz average

We then made different combinations of two and one combination of all three variables against the response. In the end, we created a table (see below: Process of Obtaining Final Model) with a sequence of models that allowed us to move forward in selecting our final model.

Variable Selection

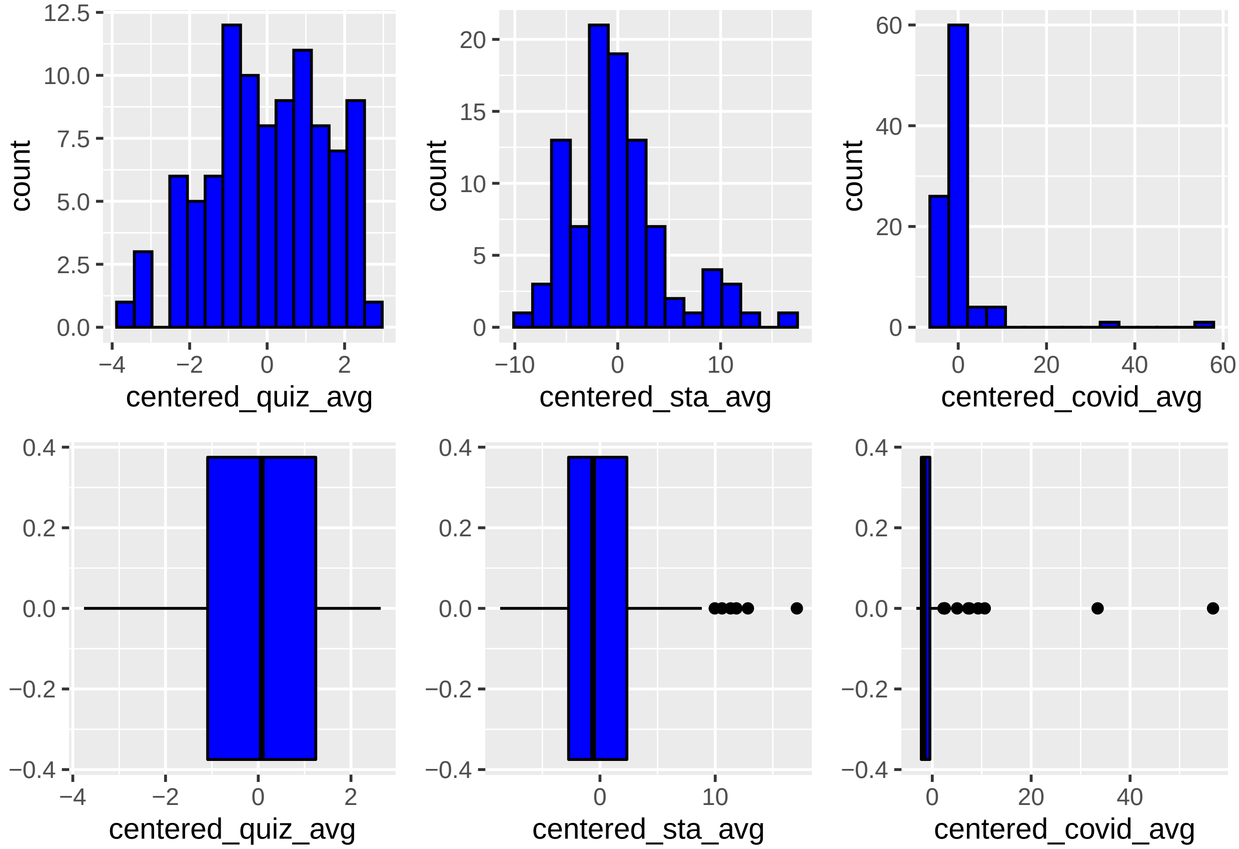
The variables we chose to include are the average hours spent on covid, average hours spent on sta302, quiz averages and country. We incorporated country as a variable by creating two different models: an "American" model and an "Asian" model. The American model encompasses Canada and the USA, and the Asian model dealt with the remaining countries. As for the other variables, we took the average as we felt it best represented the variable instead of dealing with it individually. For example, for the sta302 studying hours, we noticed an average worked better instead of having three data points for each individual. It is important here to note that the average for the quiz scores does not include quiz four grades as it is our response variable.

Data Cleaning

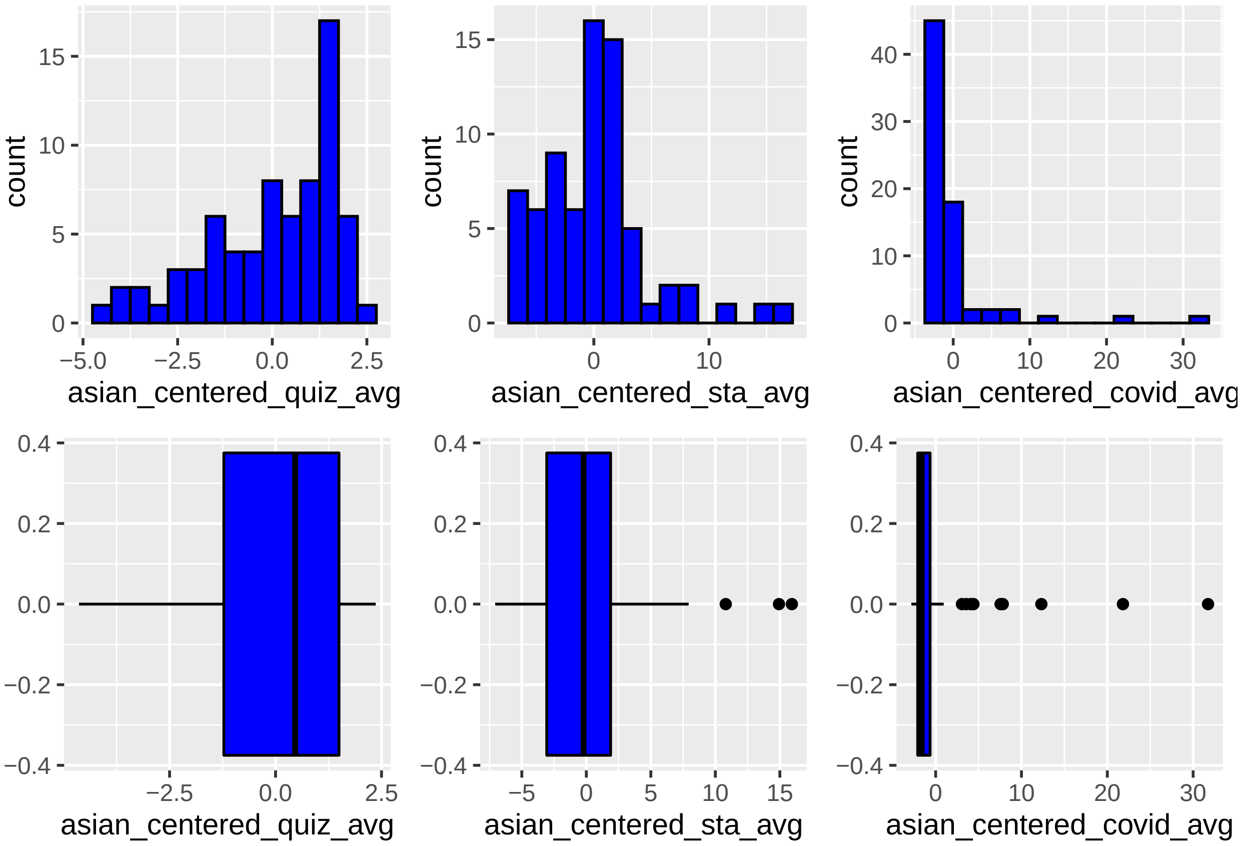
As a group, we decided we wanted to clean two important things before using our data for analysis. First and foremost, it was essential to remove rows of data that had a “NA” in the quiz 4 column. This is because quiz 4 was our response variable, and it would not make sense for that part of our data to have “NA”. We then proceeded to check within each row if the student had more than one “NA” in any of the three categories: STA302 studying, COVID19 hours and quiz grades. If the student had two “NA’s” in the category, their average would significantly skew the data, which is why we thought it would be best to remove those rows. At the end of our pre-processing stage, we were left with 182 rows of data; this was perfect as we were both able to clean our data and ensure we still had significant data left to work with.

Description of Data

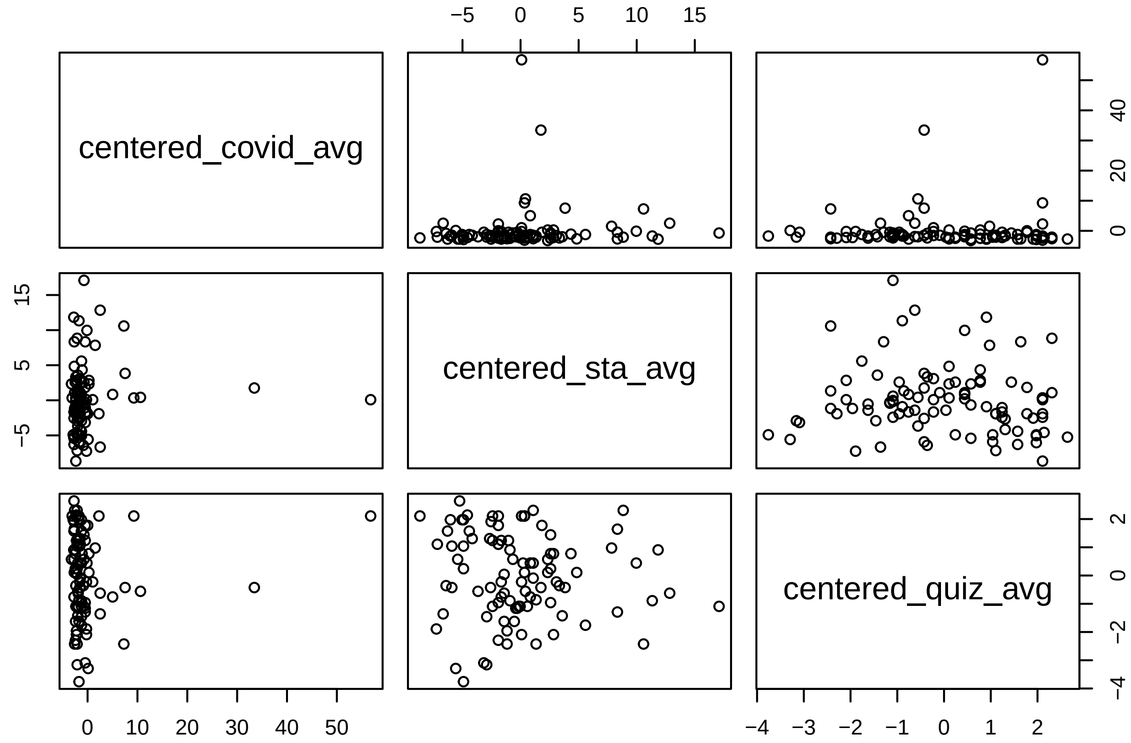
Histogram of american predictors

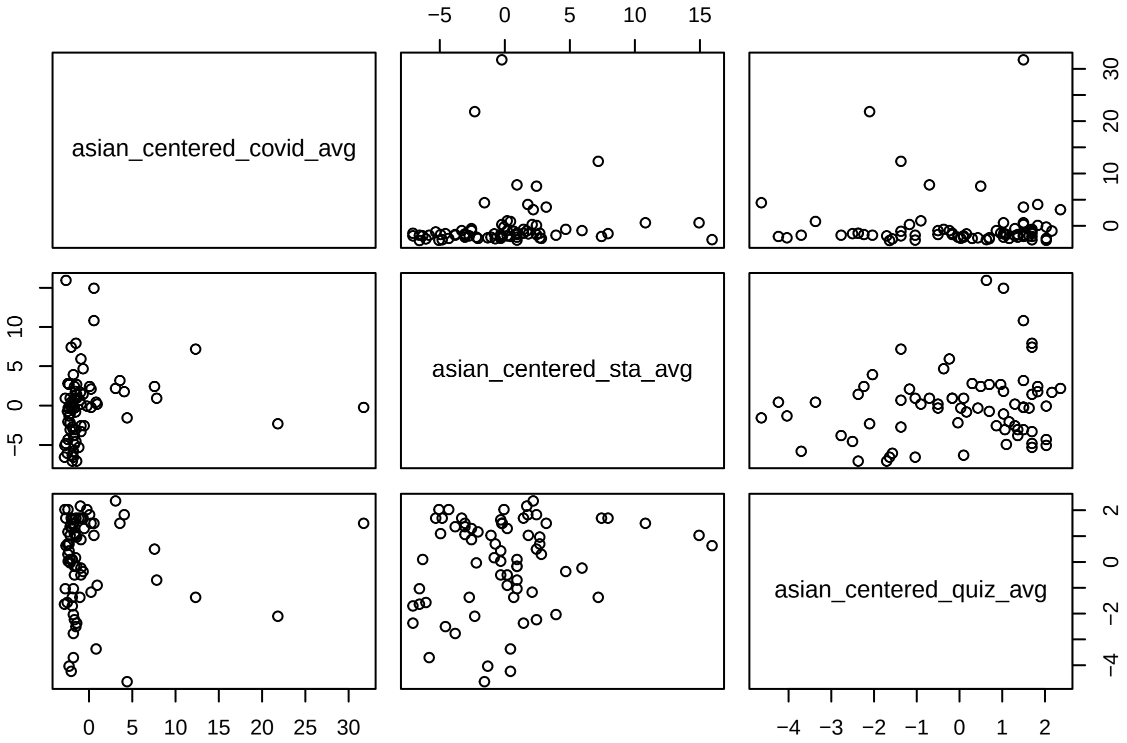


Histogram of asian predictors



Visualizing relationship between predictor variables in both the different models


centered_covid_avg
centered_sta_avg
centered_quiz_avg
−5
0
5
10
15
0
20
40
15
5
−5
2
0
−2
−4
0
10
20
30
40
50
−4
−3
−2
−1
0
1
2



Processing of Obtaining Final Model

Goodness of Final Model

We picked the following models from the American and Asian dataset

- American Data set: We chose the model with 1 predictor variable which is the quiz averages.

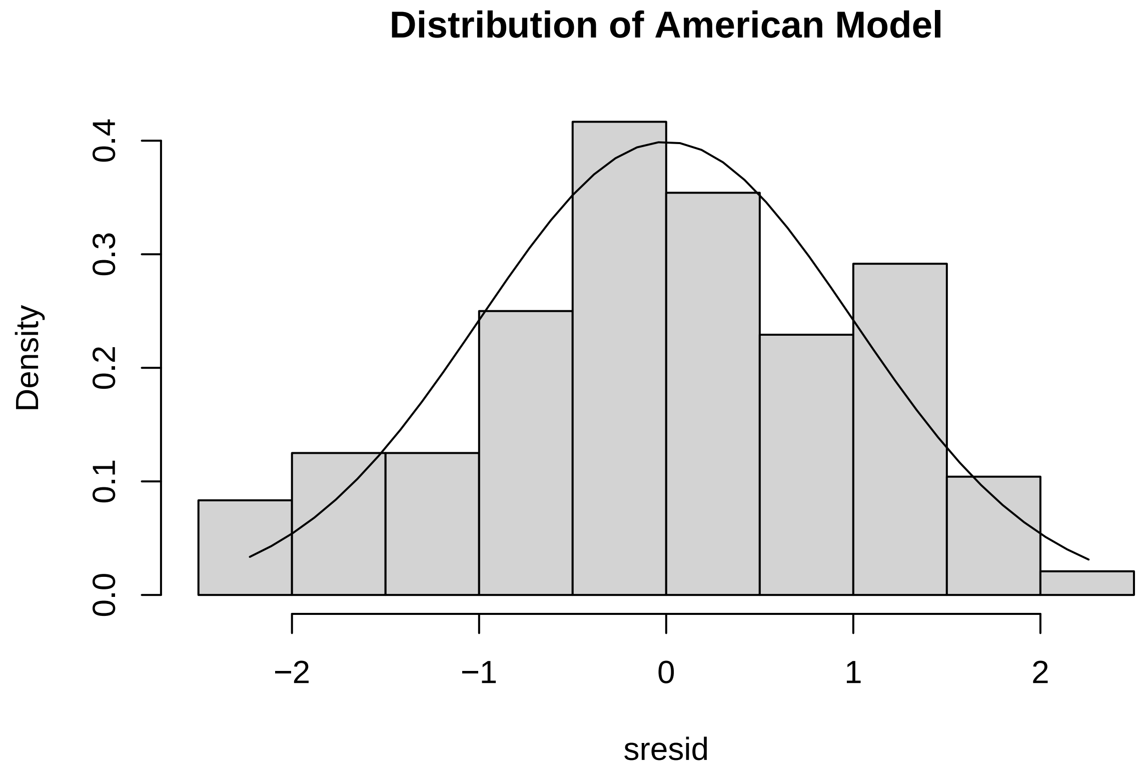
$$y_i = 7.0104 + 0.5836(quizAverage) + \epsilon$$

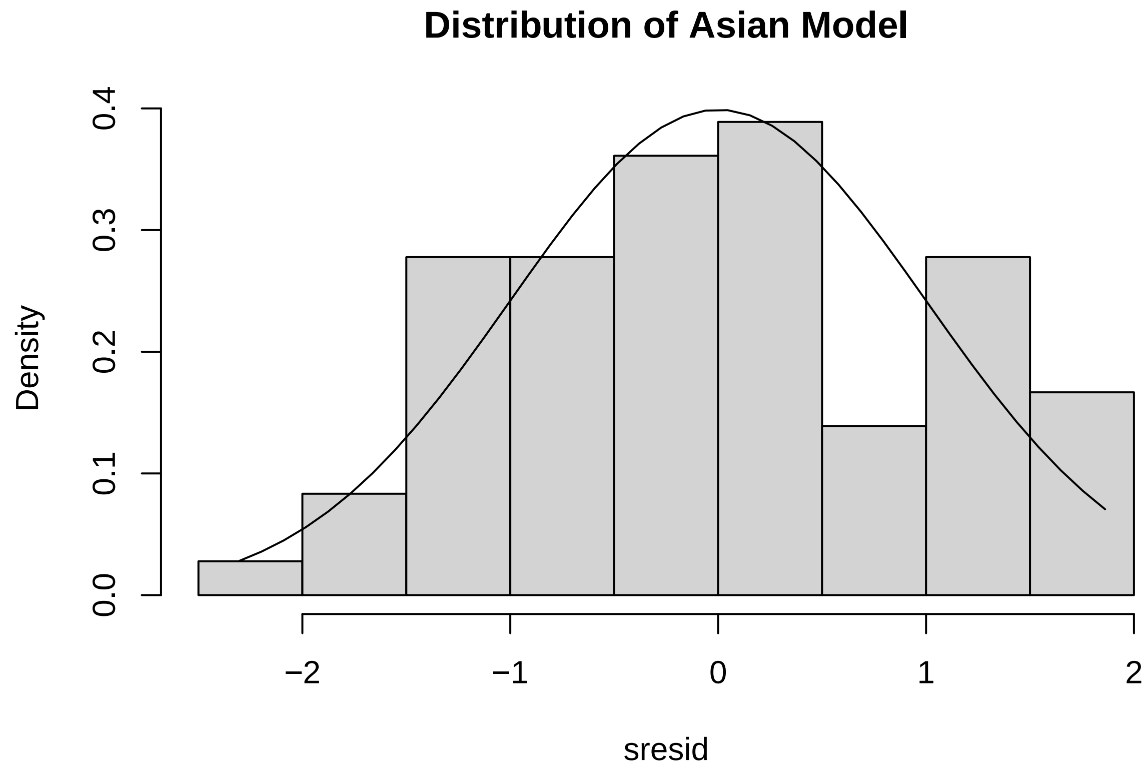
- Asian Data set: We chose the model with 2 predictor variables namely quiz average and stats hours average.

$$y_i = 7.90295 + -0.02265(statsHoursAverage) + 0.54094(quizAverage)\epsilon$$

To ensure the goodness in these final models, we have to make sure our models don't break any regression assumptions

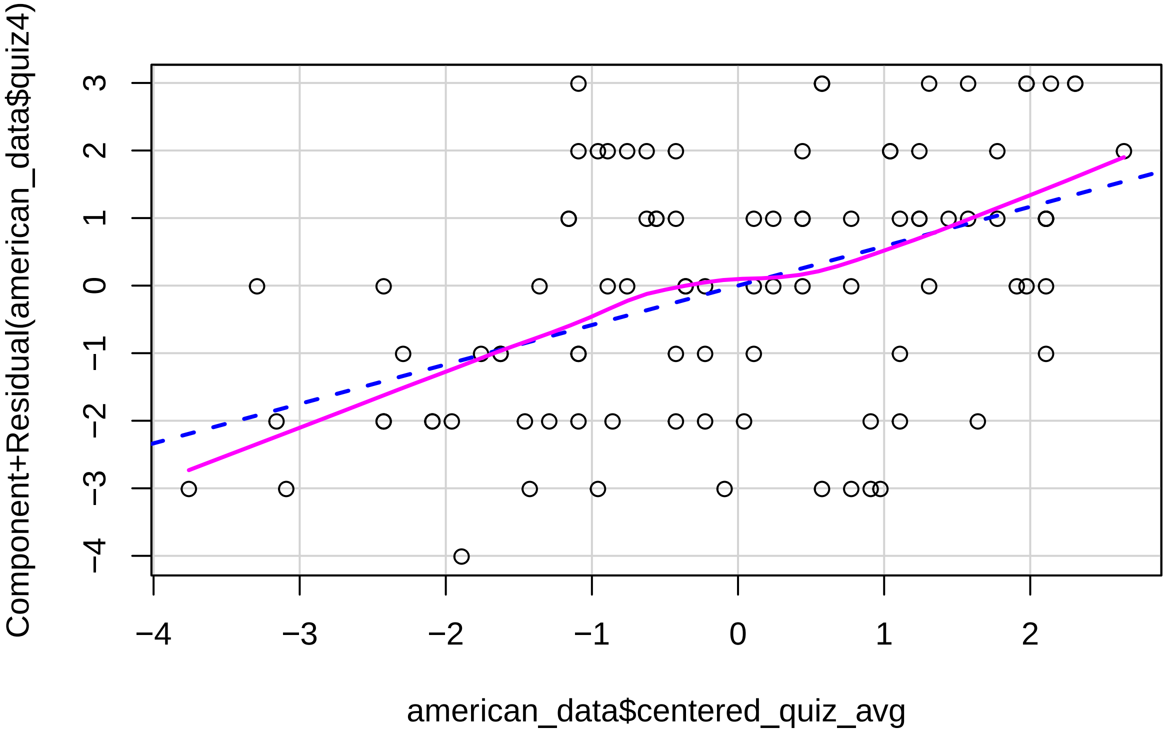
Testing Normality of American model

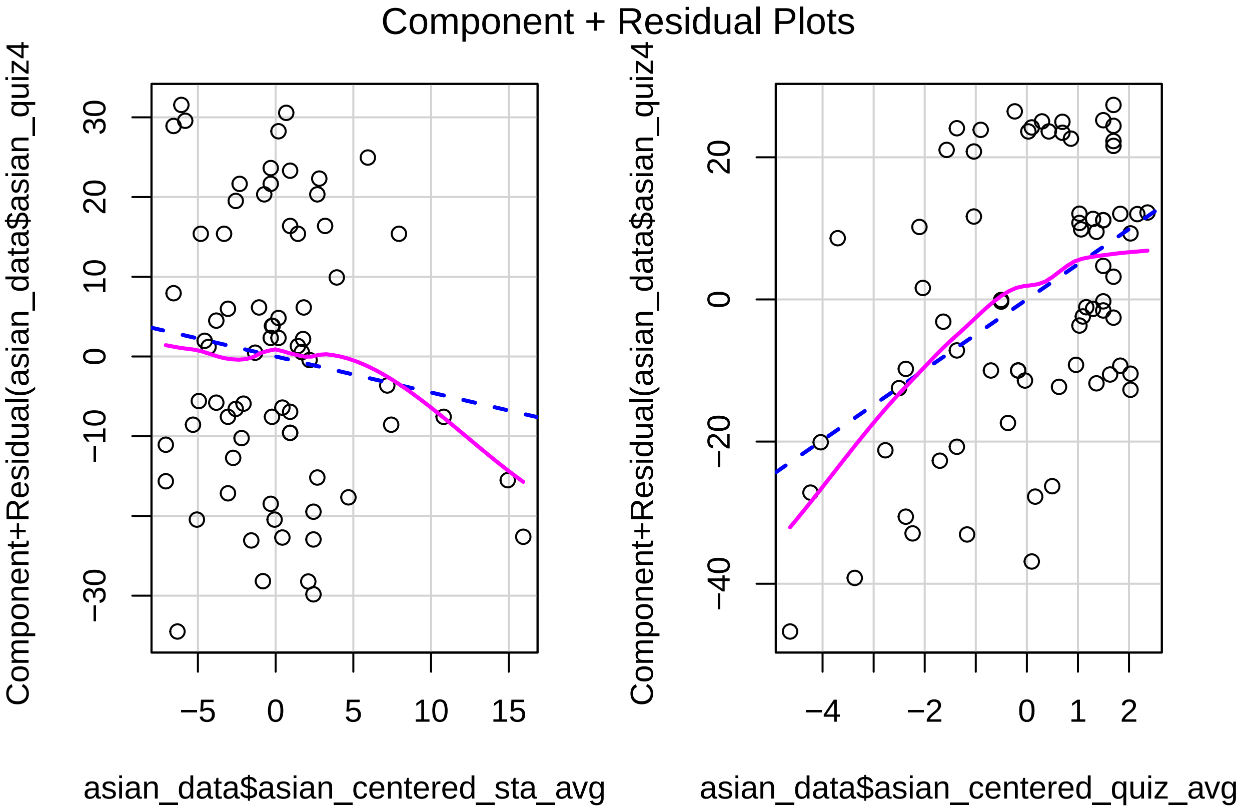




Testing the Linearity

The predictors have an almost linear relationship with the dependent variable as seen in the graphs below





Testing the Independence Assumption

We need to check if the errors are autocorrelated with themselves

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.09799564      1.803139   0.348
##  Alternative hypothesis: rho != 0
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.06184895      1.873156   0.592
##  Alternative hypothesis: rho != 0
```

Since, both the p values are greater than 0.05, this tells us that the errors are not autocorrelated and we haven't violated the independence assumption

Final Model Interpretation and Importance:

Limitations of Analysis:

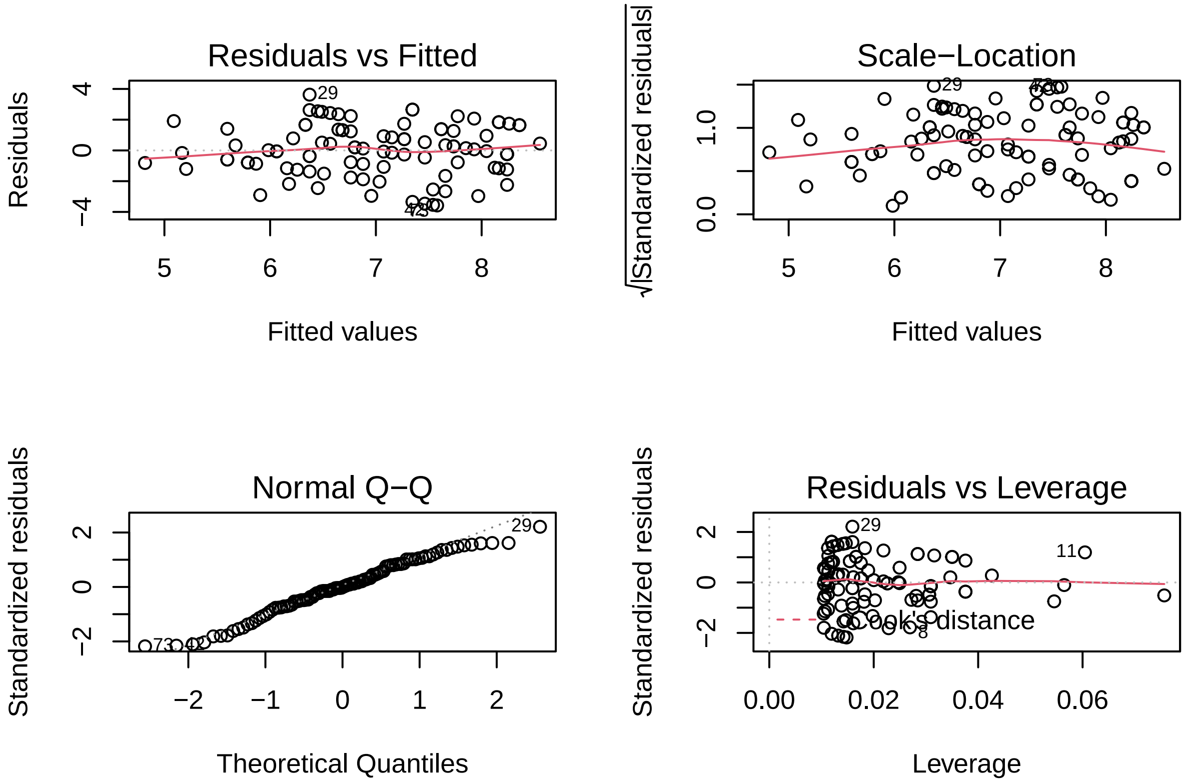
Citations

Mueller, A. B. (n.d.). Regression: Basics, Assumptions, & Diagnostics. Chapter 12: Regression: Basics, assumptions, & diagnostics. https://ademos.people.uic.edu/Chapter12.html.

Appendix

American Dataset modelling

```
n<-length(quiz4)

full_model<-lm(american_data$quiz4 ~ american_data$centered_covid_avg + american_data$centered_sta_avg +

model1<-lm(american_data$quiz4~american_data$centered_covid_avg, data=american_data)
p_prime1 <- length(model1$coefficients)
summary(model1)
SSres1<-deviance(model1)
AIC1<-n*log(SSres1) -n*log(n) + 2*p_prime1
print(AIC1)
mallow_cp1<-ols_mallows_cp(model1, full_model)

model2<-lm(american_data$quiz4~american_data$centered_sta_avg, data=american_data)
p_prime2 <- length(model2$coefficients)
summary(model2)
SSres2<-deviance(model2)
AIC2<-n*log(SSres2) -n*log(n) + 2*p_prime2
print(AIC2)
mallow_cp2<-ols_mallows_cp(model2, full_model)

model3<-lm(american_data$quiz4~american_data$centered_quiz_avg, data=american_data)
p_prime3 <- length(model3$coefficients)
summary(model3)
SSres3<-deviance(model3)
AIC3<-n*log(SSres3) -n*log(n) + 2*p_prime3
print(AIC3)
mallow_cp3<-ols_mallows_cp(model3, full_model)
layout(matrix(c(1,2,3,4),2,2))
plot(model3)
```

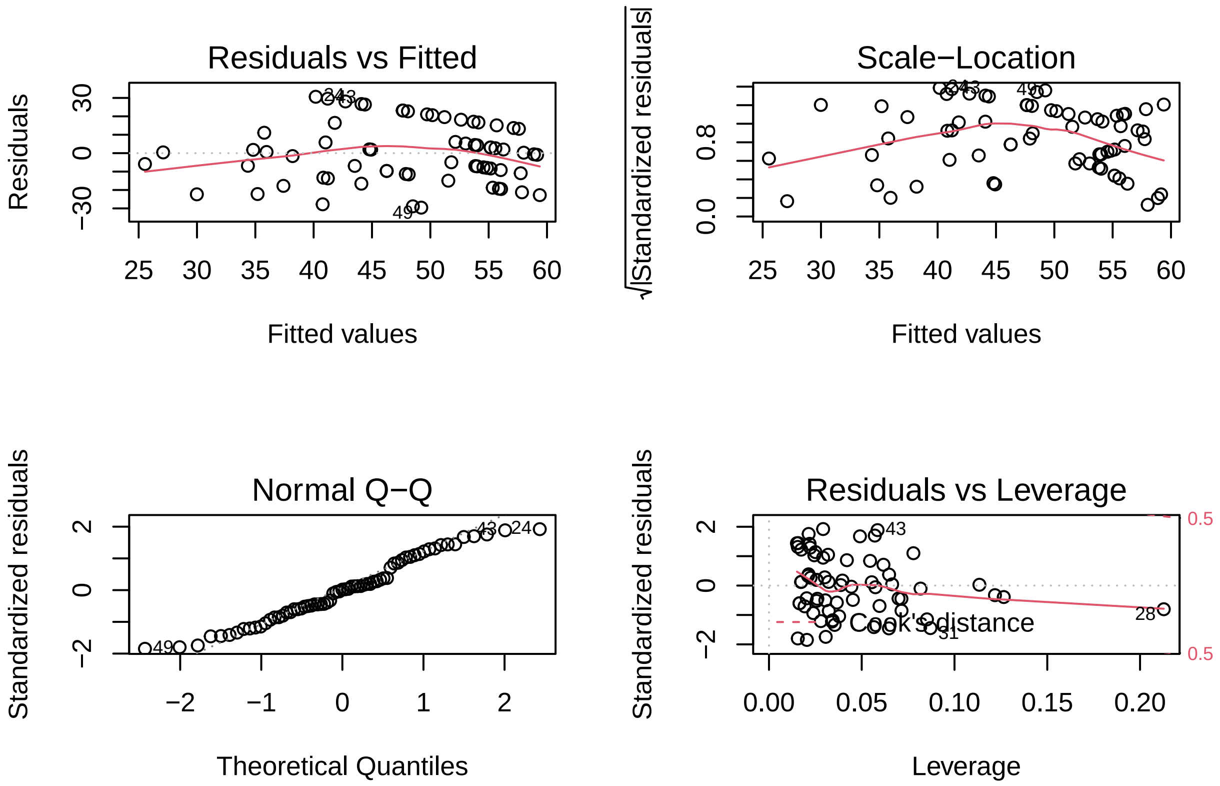


```
model4<-lm(american_data$quiz4~american_data$centered_covid_avg + american_data$centered_quiz_avg, data=
p_prime4 <- length(model4$coefficients)
summary(model4)
SSres4<-deviance(model4)
AIC4<-n*log(SSres4) -n*log(n) + 2*p_prime4
print(AIC4)
mallow_cp4<-ols_mallows_cp(model4, full_model)

model5<-lm(american_data$quiz4~american_data$centered_sta_avg + american_data$centered_quiz_avg, data=am
p_prime5 <- length(model5$coefficients)
summary(model5)
ssres5<-deviance(model5)
aic5<-n*log(ssres5) -n*log(n) + 2*p_prime5
print(aic5)
mallow_cp5<-ols_mallows_cp(model5, full_model)


model6<-lm(american_data$quiz4~american_data$centered_sta_avg + american_data$centered_covid_avg, data=a
p_prime6 <- length(model6$coefficients)
summary(model6)
SSres6<-deviance(model6)
AIC6<-n*log(SSres6) -n*log(n) + 2*p_prime6
print(AIC6)
mallow_cp6<-ols_mallows_cp(model6, full_model)

model7<-lm(american_data$quiz4~american_data$centered_quiz_avg + american_data$centered_covid_avg + ame
p_prime7 <- length(model7$coefficients)
summary(model7)
SSres7<-deviance(model7)
AIC7<-n*log(SSres7) -n*log(n) + 2*p_prime7
print(AIC7)
mallow_cp7<-ols_mallows_cp(model7, full_model)
```

Asian Dataset modelling

```
asian_n<-length(asian_quiz4)

asian_full_model<-lm(asian_data$asian_quiz4~asian_data$asian_centered_covid_avg + asian_data$asian_cente
asian_CD <- cooks.distance(asian_full_model)
n = nrow(asian_data)
asian_influential <- influential <- as.numeric(names(asian_CD)[(asian_CD > (4/n))])
asian_data <- asian_data[-asian_influential,]

asian_model1<-lm(asian_data$asian_quiz4~asian_data$asian_centered_covid_avg, data=asian_data)
asian_p_prime1 <- length(asian_model1$coefficients)
summary(asian_model1)
asian_SSres1<-deviance(asian_model1)
asian_AIC1<-asian_n*log(asian_SSres1) -asian_n*log(asian_n) + 2*asian_p_prime1
print(asian_AIC1)
asian_mallow_cp1<-ols_mallows_cp(asian_model1, asian_full_model)

asian_model2<-lm(asian_data$asian_quiz4~asian_data$asian_centered_sta_avg, data=asian_data)
asian_p_prime2 <- length(asian_model2$coefficients)
summary(asian_model2)
asian_SSres2<-deviance(asian_model2)
asian_AIC2<-asian_n*log(asian_SSres2) -asian_n*log(asian_n) + 2*asian_p_prime2
print(asian_AIC2)
asian_mallow_cp2<-ols_mallows_cp(asian_model2, asian_full_model)

asian_model3<-lm(asian_data$asian_quiz4~asian_data$asian_centered_quiz_avg, data=asian_data)
asian_p_prime3 <- length(asian_model3$coefficients)
summary(asian_model3)
asian_SSres3<-deviance(asian_model3)
asian_AIC3<-asian_n*log(asian_SSres3) -asian_n*log(asian_n) + 2*asian_p_prime3
print(asian_AIC3)
asian_mallow_cp3<-ols_mallows_cp(asian_model3, asian_full_model)

asian_model4<-lm(asian_data$asian_quiz4~asian_data$asian_centered_covid_avg + asian_data$asian_centered_
asian_p_prime4 <- length(asian_model4$coefficients)
summary(asian_model4)
asian_SSres4<-deviance(asian_model4)
asian_AIC4<-asian_n*log(asian_SSres4) -asian_n*log(asian_n) + 2*asian_p_prime4
print(asian_AIC4)
asian_mallow_cp4<-ols_mallows_cp(asian_model4, asian_full_model)

asian_model5<-lm(asian_data$asian_quiz4~asian_data$asian_centered_sta_avg + asian_data$asian_centered_qu
asian_p_prime5 <- length(asian_model5$coefficients)
```

```
summary(asian_model5)
asian_SSres5<-deviance(asian_model5)
asian_AIC5<-asian_n*log(asian_SSres5) -asian_n*log(asian_n) + 2*asian_p_prime5
print(asian_AIC5)
asian_mallow_cp5<-ols_mallows_cp(asian_model5, asian_full_model)
layout(matrix(c(1,2,3,4),2,2))
plot(asian_model5)
```



```
asian_model6<-lm(asian_data$asian_quiz4~asian_data$asian_centered_sta_avg + asian_data$asian_centered_c
asian_p_prime6 <- length(asian_model6$coefficients)
summary(asian_model6)
asian_SSres6<-deviance(asian_model6)
asian_AIC6<-asian_n*log(asian_SSres6) -asian_n*log(asian_n) + 2*asian_p_prime6
print(asian_AIC6)
asian_mallow_cp6<-ols_mallows_cp(asian_model6, asian_full_model)

asian_model7<-lm(asian_data$asian_quiz4~asian_data$asian_centered_quiz_avg + asian_data$asian_centered_
asian_p_prime7 <- length(asian_model7$coefficients)
summary(asian_model7)
asian_SSres7<-deviance(asian_model7)
asian_AIC7<-asian_n*log(asian_SSres7) -asian_n*log(asian_n) + 2*asian_p_prime7
print(asian_AIC7)
asian_mallow_cp7<-ols_mallows_cp(asian_model7, asian_full_model)
```