Ivan Malig
EDX Capstones Project

A Regression Analysis and Attempted Prediction on How Travel Time Affects Test Scores

1. Introduction

This project is the last requirement for the Data Science course on EDX. Using the Cebu Child Follow Up-Survey 1994, Cebu Community Follow-Up Survey 1994, and Cebu Mother Follow Up-Survey 1994, I plan to see how travel time affects a student's academic performance. For this particular study, I will be looking at travel time's effect on primary school students. The motivation of the study is the fact that the country has since been having problems with travel time, be it because of the lack of available means of transportation or because of traffic. Helping clear the effect of travel on education, be it positive or negative, can be a significant step in trying to fix problems with traffic. The dataset used contains different information such as current barangay where the student lives, distance from the nearest public/private school, travel time to school in hours, and of course test scores in different subject areas.

A sample of size of 2,829 children is used for the mathematics, English, and Cebu reading achievement tests regression. A sample size of 2,859 children is used for the non-verbal intelligence test regression. Sample size is determined after datasets have been merged before regression.

The dependent variables used in the regression are the following: iqscore and SCORE2. The child’s total score for nonverbal intelligence is given through iqscore. The score, iqscore, is measured out of 100 items. SCORE2 shows the score of child in Cebuano reading, mathematics, and English achievement tests. The same variable name, SCORE2, is used for the three different achievement. The score, SCORE2, is measured out of 60 items.

The main independent variable included in the regressions is TRAVHR2. Variable, TRAVHR2, tells us how many hours the index child spends travelling to school on a regular school day. This particular variable is measured in hours.

Datasets were uploaded on my github repository and imported into r using this code

```
#Download datasets
engldata<- read_dta("https://github.com/igmalig/P/blob/master/engl.dta?raw=true")
community<-
read_dta("https://github.com/igmalig/P/blob/master/comunit1.dta?raw=true")
childdata<- read_dta("https://github.com/igmalig/P/blob/master/child2.dta?raw=true")
cebudata<- read_dta("https://github.com/igmalig/P/blob/master/cebu.dta?raw=true")
childfile<- read_dta("https://github.com/igmalig/P/blob/master/cdietiq.dta?raw=true")
motherdata<- read_dta("https://github.com/igmalig/P/blob/master/mother.dta?raw=true")
```

```
mathdata<- read_dta("https://github.com/igmalig/P/blob/master/math.dta?raw=true")
```

In the 2nd part where I attempted to predict (rmse model) an iqscore based on a student's travel time, distance from public school, and distance from private school, I introduced IQset data frame which was created from the MERGECHILDFINAL. The merged dataframe was then cleaned by the na.omit function since there are many rows and columns with na values. The idea was to remove and row that had at least one na value. This trimmed down our dataset a lot. The challenge with this is the fact that some samples which could be really important in our study are automatically removed if they had an NA value for travel time distance from public schools, or distance from private schools.

2. Method

In examining the relationship between travel time and academic performance, I used multiple linear regression.

The **multiple linear regression models** are written as:

iqscore = 0 + 1 TRAVHR2 + 2 SEXCHLD2 + 3 CURSTRA2 + 4agechild + 5 CAREKID2 + 6 ILLNESC2 + 7 MISSEDA2 + 8NUMSIBLINGS + 9 WITHSPO2 + u

SCORE2 = 0 + 1 TRAVHR2 + 2 SEXCHLD2 + 3agechild + 4CAREKID2 + 5 ILLNESC2 + 6 MISSEDA2 + 7NUMSIBLINGS + 8 WITHSPO2 + u

Where *0* is the intercept, 1 is the parameter associated with x1,2is the parameter associated with x2, and so on. u is the error term and contains all the other explanatory variables that are not included in the equation.

This is the step by step on how we came up with the merged data sets that was used. INcluded also are details on what the specific columns showed.

```
#To seethe headings of our community dataset
colnames(community)

#Rename barangay column to "current barangay"
names(community)[2] <- "CURBRGY2"

#Choose which columns or the community dataset to keep
community1 <- community[c(2,26,42,43,44,45,54,55,56,57)]
head(community1)
#2 CURBRGY 2 - Current Barangay
#26 density - POPULATION DENSITY PERSONS PER SQUARE KILOMETER
#42 compubel - Are there Complete Public Elementary Schools in the Barangay? If yes how many?
```

```
#43 macpubel - TOTAL NUMBER OF MALE STUDENTS IN THE COMPLETE
PUBLIC ELEMENTARY SCHOOL(S)
#44 fecpubel - TOTAL NUMBER OF FEMALE STUDENTS IN THE COMPLETE
PUBLIC ELEMENTARY SCHOOL(S)
#45 dicpel- How far in (km) is the brgy. center to the closest complete public school not
found in the barangay?
#54 cprvtels - Are there Complete Private Elementary Schools Barangay? If yes how
many?
#55 macprvte - TOTAL NUMBER OF MALE STUDENTS IN THE COMPLETE
PRIVATE ELEMENTARY SCHOOL(S)
#56 fecprvte - TOTAL NUMBER OF FEMALE STUDENTS IN THE COMPLETE
PRIVATE ELEMENTARY SCHOOL(S)
#57 dicprvte- HOW FAR (IN KM.) IS THE BRGY CENTER TO THE CLOSEST
COMPLETE PRIVATE ELEMENTARY SCHOOL NOT FOUND IN BARANGAY?

#Count the distance of barangay center to the closest complete public schol and
population density in communit1
count(community1,dicpel, wt = NULL, sort = FALSE)
count(community1,density, wt = NULL, sort = FALSE)



#Choose columns from the Math dataset
mathdata1 <- mathdata[c(1,14,15,80)]
head(mathdata1)
#1 CURBRGY2 - CURRENT BARANGAY OF RESIDENCE OF CHILD
#2 SCHOLBA2 - BARANGAY IN WHICH SCHOOL IS LOCATED X
#14 HHNUM942 - CHILD'S 1991 HOUSEHOLD ID NUMBER
#15 WOMAN942 - 1994 WOMAN ID NUMBER
#16 CHLD CODE - CHILD's CODE
#80 SCORE2 - SCORE OF CHILD IN MATHEMATICS ACHIEVEMENT TEST

#Choose columns from the English dataset
engldata1 <- engldata[c(1,14,15,80)]
head(engldata1)
#1 CURBRGY2 - CURRENT BARANGAY OF RESIDENCE OF CHILD
#2 SCHOLBA2 - BARANGAY IN WHICH SCHOOL IS LOCATED X
#14 HHNUM942 - CHILD'S 1991 HOUSEHOLD ID NUMBER
#15 WOMAN942 - 1994 WOMAN ID NUMBER
#16 CHLDCODE - CHILD's CODE
#80 SCORE2 - SCORE OF CHILD IN ENGLISH ACHIEVEMENT TEST

#Choose columns from the English dataset
cebudata1 <- cebudata[c(1,14,15,50)]
head(cebudata1)
#1 CURBRGY2 - CURRENT BARANGAY OF RESIDENCE OF CHILD
```

```
#2 SCHOLBA2 - BARANGAY IN WHICH SCHOOL IS LOCATED X
#3 SCHLCODE - CODED NAME OF SCHOOL X
#14 HHNUM942 - CHILD'S 1991 HOUSEHOLD ID NUMBER
#15 WOMAN942 - 1994 WOMAN ID NUMBER
#16 CHLDCODE - CHILD's CODE
#50 SCORE 2 - SCORE OF CHILD IN CEBUANO READING ACHIEVEMENT TEST

#We check here how many people there are for different birth years (agechild <- (94-childfile$YRBIRTC2))
count(childfile,YRBIRTC2, wt = NULL,sort=FALSE)
childfile$CURSTRA2[childfile$CURSTRA2==1] <- 0
childfile$CURSTRA2[childfile$CURSTRA2==2] <- 1

#Choose which columns to keep for childfile dataset
childfile1 <- childfile[c(2,3,14,22,45,232)]
head(childfile1)
#childfile1 <- cbind(childfile1, agechild)
#2 CURBRGY2 - CURRENT BARANGAY
#3 CURSTRA2 - CURRENT STRATUM (SEE APPENDIX 1)
#14 HHNUM942 - 1994 HOUSEHOLD ID NUMBER
#20 YRBIRTC2 - YEAR OF BIRTH OF CHILD -> Determine Age
#22 CHLNO942 - CHILD'S LINE NUMBER IN 1994
#45 MO1AB942 - NUMBER OF DAYS CHILD ABSENT IN MONTH1 (SY 1994-1995)
#232 IQSCORE - CHILD'S TOTAL SCORE FOR NONVERBAL INTELLIGENCE TEST
#agechild - age of child

#Merge the exam data sets with community1
mergemathdata <- merge(mathdata1,community1)

mergeengldata <- merge(engldata1,community1)

mergecebudata <- merge(cebudata1,community1)

mergechildfiledata <- merge(childfile1,community1)

#Compute age of children
agechild <- (94-childdata$YRBIRTC2)
colnames(childdata)[1] <- "CURBRGY2"
childdata$SEXCHLD2[childdata$SEXCHLD2==1] <- 0
childdata$SEXCHLD2[childdata$SEXCHLD2==2] <- 1

#Choose columns from the childdata dataset
childdata1 <- childdata[c(1,11,12,14,19,90,91,92,93,94,101,144)]
head(childdata1)
```

```
childdata1 <- cbind(childdata1,agechild)
#1 BASEBRGY to CURBRGY2 - BASE BARANGAY - Change name to CURBRGY2
#11 HHNUM942 - 1994 HOUSEHOLD ID NUMBER
#12 WOMAN942 - 1994 WOMAN ID NUMBER
#14 SEXCHLD2 - SEX OF INDEX CHILD
#16 YRBIRTC2 - YEAR OF BIRTH OF INDEX CHILD X
#17 CHLDCODE - CHILD'S CODE X
#18 CHLNO942- LINE NUMBER OF INDEX CHILD (1994) X
#19 SCHOLIN2 - HAS INDEX CHILD EVER ATTENDED SCHOOL?
#90 TRAVHR2 - ON A REGULAR SCHOOLDAY, HOW MANY HOURS DOES INDEX CHILD SPEND TRAVELLING TO SCHOOL?
#91 WORKHR2 - ON A REGULAR SCHOOLDAY, HOW MANY HOURS DOES INDEX CHILD SPEND WORKING FOR PAY OR ON FARM OR FAMILY BUSINESS?
#92 CHOREHR2 - ON A REGULAR SCHOOLDAY, HOW MANY HOURS DOES INDEX CHILD SPEND HELPING WITH HOUSEHOLD CHORES?
#93 CARESIB2 - ON A REGULAR SCHOOLDAY, HOW MANY HOURS DOES INDEX CHILD SPEND CARING OF YOUNGER SIBLINGS?
#94 PLAYHR2 - ON A REGULAR SCHOOLDAY, HOW MANY HOURS DOES INDEX CHILD SPEND PLAYING?
#101 MISSEDA2 - IN THE PAST MONTH, HOW MANY DAYS HAS INDEX CHILD MISSED SCHOOL WHEN IT WAS IN SESSION?
#129 CAUSEH12 - CAUSE OF INDEX CHILD'S FIRST HOSPITALIZATION X
#144 ILLNESC2 - DOES INDEX CHILD HAVE ANY CHRONIC ILLNESS/DISABILITY?

#Merge exam data sets with the created childdata1
MERGEMATH <- merge(mergemathdata,childdata1)
MERGEENGL <- merge(mergeengldata,childdata1)
MERGECEBU <- merge(mergecebudata,childdata1)
MERGECHILDFILE <- merge(mergechildfiledata,childdata1)

count(MERGEENGL,SEXCHLD2, wt = NULL,sort=FALSE)
colnames(motherdata)[1] <- "CURBRGY2"
motherdata$CAREKID2[motherdata$CAREKID2==2] <- 0
motherdata$CAREKID2[motherdata$CAREKID2==3] <- 0
motherdata$CAREKID2[motherdata$CAREKID2==4] <- 0
motherdata$WITHSPO2[motherdata$WITHSPO2==3] <- 0
motherdata$WITHSPO2[motherdata$WITHSPO2==2] <- 0
motherdata$LIVTODA2 <- motherdata$LIVTODA2-1

#Change column name
colnames(motherdata)[32] <- "NUMSIBLINGS"

#Select which columns to keep for motherdata
```

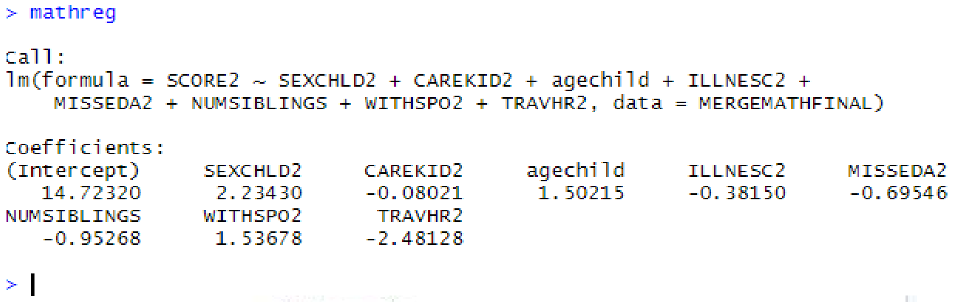
```
motherdata1 <- motherdata[c(1,2,12,11,32,43,130)]
head(motherdata1)
#1 BASEBRG2 to CURBRGY2 - BASELINE BARANGAY ID NUMBER
#11 HHNUM942 - 1994 HOUSEHOLD ID NUMBER
#14 LINENUM2 - LINE NUMBER OF MOTHER/CARETAKER X
#12 WOMAN942 - 1994 WOMAN ID NUMBER
#21 TYPEPA12 - HOW IS MOTHER DOING HER MAIN JOB? X
#16 MARSTAT2 - MARITAL STATUS OF MOTHER/CARETAKER X
#25 TYPEPA22 - HOW IS MOTHER PAID ON HER SECOND JOB? X
#32 LIVTODA2 - TOTAL NUMBER OF CHILDREN MOTHER HAVE GIVEN BIRTH TO WHO ARE STILL ALIVE TODAY
#47 WITHSPO2 - IS MOTHER CURRENTLY LIVING WITH HUSBAND NOW?
#130 CAREKID2 - CAN MOTHER STILL TAKE CARE OF THE CHILDREN?

sum(motherdata$CAREKID2)
count(motherdata,CAREKID2, wt = NULL, sort = FALSE)

#Final merging
MERGEMATHFINAL <- merge(motherdata1,MERGEMATH)
MERGEENGLFINAL <- merge(motherdata1,MERGEENGL)
MERGECEBUFINAL <- merge(motherdata1,MERGECEBU)
MERGECHILDFILEFINAL <- merge(motherdata1,MERGECHILDFILE)
```

Finally, these were the codes used for plotting the different results in our regression

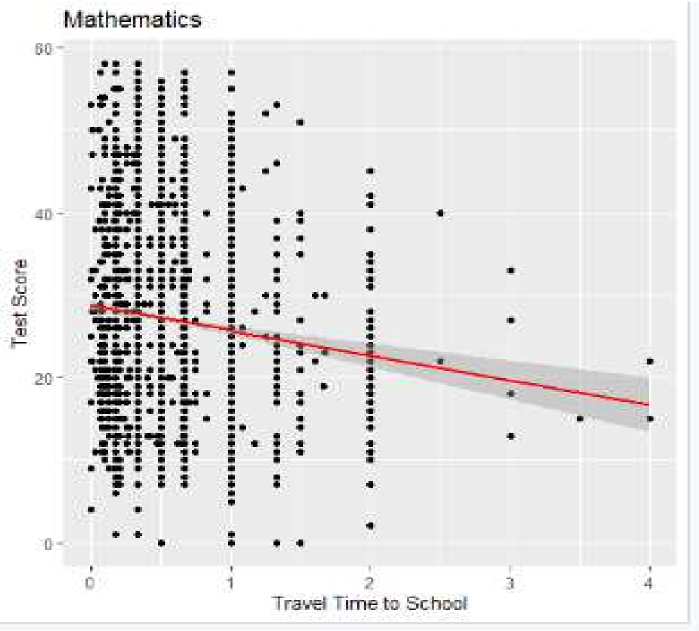
```
ggplot(MERGECHILDFILEFINAL, aes(x = TRAVHR2, y = iqscore)) +
  geom_point() + ggtitle("Non-Verbal Intelligence") +
  ylab("Test Score") + xlab("Travel Time to School") +
  stat_smooth(method = "lm", col = "red")

ggplot(MERGEMATHFINAL, aes(x = TRAVHR2, y = SCORE2)) +
  geom_point() + ggtitle("Mathematics") +
  ylab("Test Score") + xlab("Travel Time to School") +
  stat_smooth(method = "lm", col = "red")

ggplot(MERGEENGLFINAL, aes(x = TRAVHR2, y = SCORE2)) +
  geom_point() + ggtitle("English") +
  ylab("Test Score") + xlab("Travel Time to School") +
  stat_smooth(method = "lm", col = "red")

ggplot(MERGECEBUFINAL, aes(x = TRAVHR2, y = SCORE2)) +
  geom_point() + ggtitle("Cebuano Reading") +
  ylab("Test Score") + xlab("Travel Time to School") +
  stat_smooth(method = "lm", col = "red")
```

3. Results

mathreg <- lm(SCORE2 ~ SEXCHLD2 + CAREKID2 + agechild + ILLNESC2 + MISSEDA2 + NUMSIBLINGS + WITHSPO2 + TRAVHR2, data=MERGEMATHFINAL)

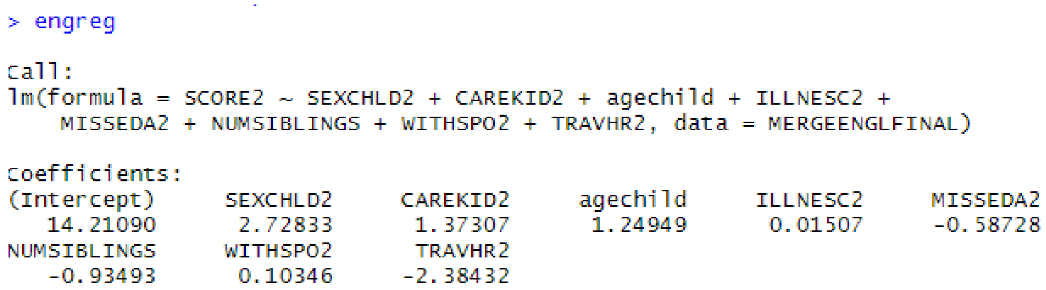


```
> mathreg

Call:
lm(formula = SCORE2 ~ SEXCHLD2 + CAREKID2 + agechild + ILLNESC2 +
    MISSEDA2 + NUMSIBLINGS + WITHSPO2 + TRAVHR2, data = MERGEMATHFINAL)

Coefficients:
(Intercept)     SEXCHLD2     CAREKID2     agechild     ILLNESC2     MISSEDA2
   14.72320      2.23430     -0.08021      1.50215     -0.38150     -0.69546
NUMSIBLINGS     WITHSPO2      TRAVHR2
   -0.95268      1.53678     -2.48128

> |
```

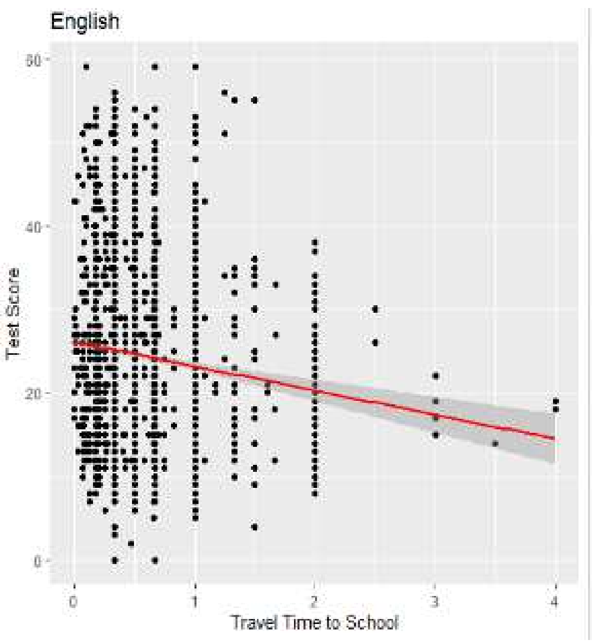


We can see the effects of different coefficients on mathematics test scores however we are most interested in travel time. We see that travel time has a -2.48 coefficient, signifying that an extra hour of travel reduces math scores by 2.48. We can also see from the plot that the maximum scores for math test decreased for every additional hour of travel.

engreg <- lm(SCORE2 ~ SEXCHLD2 + CAREKID2 + agechild + ILLNESC2 + MISSEDA2 + NUMSIBLINGS + WITHSPO2 + TRAVHR2, data=MERGEENGLFINAL)

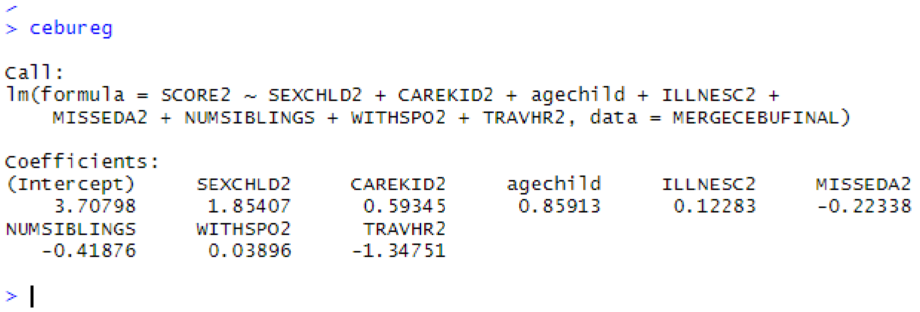


```
> engreg

Call:
lm(formula = SCORE2 ~ SEXCHLD2 + CAREKID2 + agechild + ILLNESC2 +
    MISSEDA2 + NUMSIBLINGS + WITHSPO2 + TRAVHR2, data = MERGEENGLFINAL)

Coefficients:
(Intercept)     SEXCHLD2     CAREKID2     agechild     ILLNESC2     MISSEDA2
   14.21090      2.72833      1.37307      1.24949      0.01507     -0.58728
NUMSIBLINGS     WITHSPO2      TRAVHR2
   -0.93493      0.10346     -2.38432
```

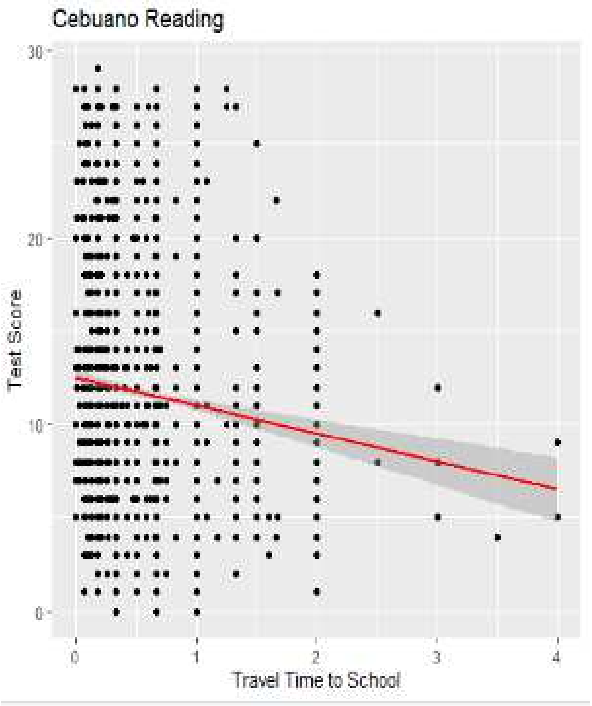


We can see here that the variable for tavel time, similar to the one in mathreg, is negative. This means that the longer the the travel, the lower the resulting score in English test.

cebureg <- lm(SCORE2 ~ SEXCHLD2 + CAREKID2 + agechild + ILLNESC2 + MISSEDA2 + NUMSIBLINGS + WITHSPO2 + TRAVHR2, data=MERGECEBUFINAL)

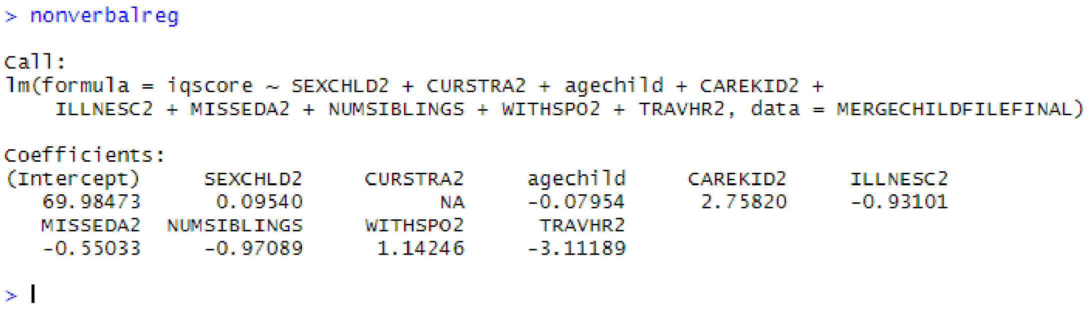


```
> cebureg

Call:
lm(formula = SCORE2 ~ SEXCHLD2 + CAREKID2 + agechild + ILLNESC2 +
    MISSEDA2 + NUMSIBLINGS + WITHSPO2 + TRAVHR2, data = MERGECEBUFINAL)

Coefficients:
(Intercept)     SEXCHLD2     CAREKID2     agechild     ILLNESC2     MISSEDA2
    3.70798      1.85407      0.59345      0.85913      0.12283     -0.22338
NUMSIBLINGS     WITHSPO2      TRAVHR2
   -0.41876      0.03896     -1.34751

> |
```

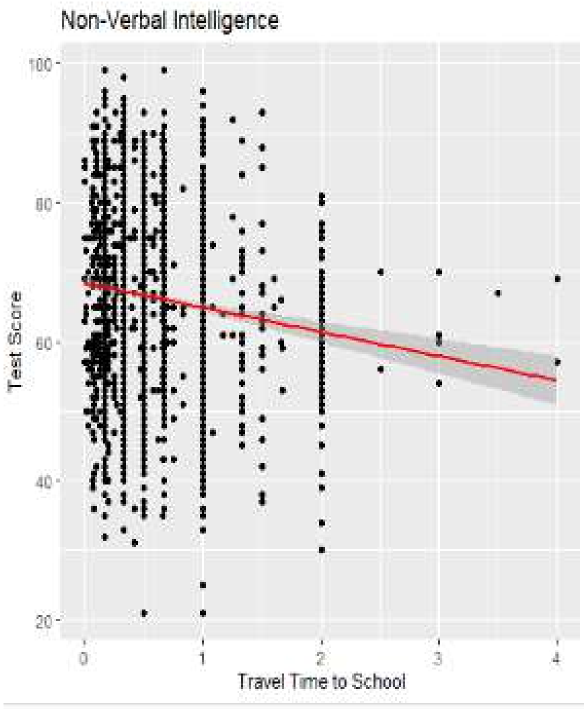


As has been the case, longer travel time results in a lower test score. However, there is an interesting result here. We can see that the coefficent for travel (-1.34) is significantly lower than the coefficients for math and english. A good reason for this is the fact that the samples are cebuano; therefore they are bound to do better in cebuano reading as compared to other subjects.

nonverbalreg <- lm(iqscore ~ SEXCHLD2 + CURSTRA2 + agechild + CAREKID2 + ILLNESC2 + MISSEDA2 + NUMSIBLINGS + WITHSPO2 + TRAVHR2, data=MERGECHILDFILEFINAL)

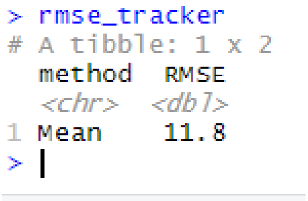


```
> nonverbalreg

Call:
lm(formula = iqscore ~ SEXCHLD2 + CURSTRA2 + agechild + CAREKID2 +
    ILLNESC2 + MISSEDA2 + NUMSIBLINGS + WITHSPO2 + TRAVHR2, data = MERGECHILDFILEFINAL)

Coefficients:
(Intercept)     SEXCHLD2     CURSTRA2     agechild     CAREKID2     ILLNESC2
   69.98473      0.09540           NA     -0.07954      2.75820     -0.93101
   MISSEDA2  NUMSIBLINGS     WITHSPO2      TRAVHR2
   -0.55033     -0.97089      1.14246     -3.11189

> |
```



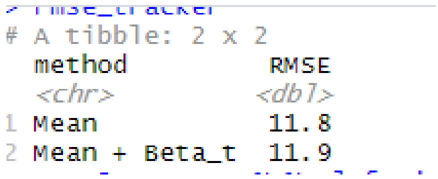
Finally, as expected, the coefficient for nonverbal also resulted in a negative value.

For the 2nd part, the goal is to run an rmse simulation in hopes of predicting iqscores from information such as travel time, distance from public elementary school, and distance from private school. Similar to the movielensproject, the simplest prediction we do is predicting by the iqscore mean. Here is the resulting rmse from this method.



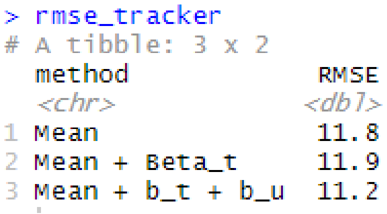
```
> rmse_tracker
# A tibble: 1 x 2
  method  RMSE
  <chr>  <dbl>
1 Mean    11.8
> |
```

The Next method, I tried incorporating is predicting with a mean and a penalty term for travel hours beta_t. Unfortunately, the results were a bit confusing since rmse actually increased



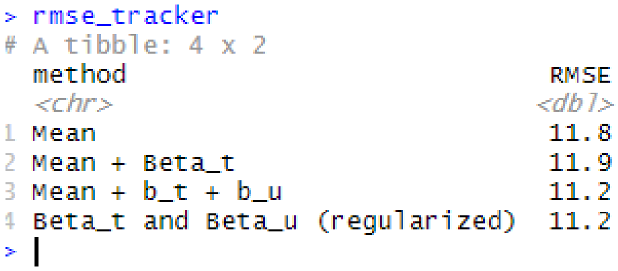
```
> rmse_tracker
# A tibble: 2 x 2
  method          RMSE
  <chr>          <dbl>
1 Mean            11.8
2 Mean + Beta_t   11.9
```

The next model has a penalty term beta_u included. This term refers to the distance from a public school. The resulting rmse as we see below has been reduced however not by much.



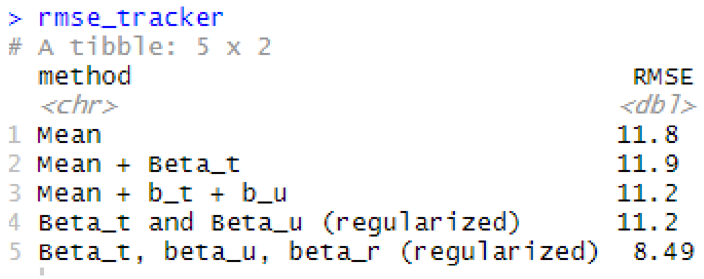
```
> rmse_tracker
# A tibble: 3 x 2
  method             RMSE
  <chr>             <dbl>
1 Mean               11.8
2 Mean + Beta_t      11.9
3 Mean + b_t + b_u   11.2
```

I next tried regularizing the factors beta_t and beta_u. As we can see it did not do anything to our rmse. The lambda obtained that minimizes rmse



```
> rmse_tracker
# A tibble: 4 x 2
  method                             RMSE
  <chr>                             <dbl>
1 Mean                               11.8
2 Mean + Beta_t                      11.9
3 Mean + b_t + b_u                   11.2
4 Beta_t and Beta_u (regularized)   11.2
> 
```

Finally, I tried regularizing while including another variable, beta_r, which is a variable for distance from private schools. We can see that once we included beta_r, the resulting rmse suddenly dropped to around 8.49



```
> rmse_tracker
# A tibble: 5 x 2
  method                                  RMSE
  <chr>                                  <dbl>
1 Mean                                   11.8 
2 Mean + Beta_t                          11.9 
3 Mean + b_t + b_u                       11.2 
4 Beta_t and Beta_u (regularized)        11.2 
5 Beta_t, beta_u, beta_r (regularized)    8.49
```

Our final rmse was still high, although it did decrease substantially from our initial rmse of 11.8. Some of my ideas as to why we got the results we did was the fact that our dataset contained a lot of NA values.

4. Conclusion

After this project, we were able to establish the fact that longer travel hours lead to lower scores across all subject areas used. This is intuitive in itself since longer travel hours would mean less time for other matters such as studying. The fast that we also saw a lower coefficient for cebuano learning was also intuitive since cebuanos are expected to be naturally better in that aspect. For the second prediction, our results were a bit confusing since the resulting rmse was not as low as I wanted it to be however we were able to improve from the initial prediction of using the mean only. The project does have a lot of limitations. For starters, the dataset used in the second one was trimmed down since there were a lot of incomplete information. Another thing is that I was only able to consider some out of the many possible predictors available meaning I could have gotten a much lower rmse had I used more predictors. Overall, the project was able to confirm my intuition and while the score prediction in the latter part was not as good as expected, it wan better than the basic mean prediction method.