DSC 423

Assignment 3

Based on Modules 5 and 7

Ronaldlee Ejalu

StudentId: 2020637

I have completed this work independently. The solutions given are entirely my own work

Your submission must include your name and student ID. Your submission must include the honor statement: "I have completed this work independently. The solutions given are entirely my own work."

I created two dummy variables on raceeth and grade:

Set 10 and white as reference levels for grade and raceeth respectively in my data cleaning process as

shown below:

```
## data cleaning
```{r}
pisa_clean <- pisa_select %>%
  transmute(grade = as.factor(grade)
            , male = male
            , raceeth = as.factor(raceeth)
            , preschool = preschool
            , expectbachelors = expectbachelors
            , motherhs = motherhs
            , motherbachelors = motherbachelors
            , motherwork = motherwork
            , fatherhs = fatherhs
            , fatherbachelors = fatherbachelors
            , fatherwork = fatherwork
            , selfbornus = selfbornus
            , motherbornus = motherbornus
            , fatherbornus = fatherbornus
            , englishathome = englishathome
            , computerforschoolwork = computerforschoolwork
            , read30minsaday = read30minsaday
```
```

```
            , minutesperweekenglish = minutesperweekenglish
            , studentsinenglish = studentsinenglish
            , schoolhaslibrary = schoolhaslibrary
            , publicschool = publicschool
            , urban = urban
            , schoolsize = schoolsize
            , readingscore = readingscore) %>%

mutate(raceeth = relevel(raceeth, ref = 'White')
       , grade = relevel(grade, ref = '10')
       )

## we create a matrix for raceeth
raceethdummies.matrix <- model.matrix(~pisa_clean$raceeth)
## convert the model matrix into a data frame
raceethdummies.frame <- data.frame(raceethdummies.matrix)
pisa_clean <- cbind(pisa_clean, raceethdummies.frame)

##Create a matrix for grade
gradedummies.matrix <- model.matrix(~pisa_clean$grade)
## convert the model matrix into a data frame
gradedummies.frame <- data.frame(gradedummies.matrix)
pisa_clean <- cbind(pisa_clean, gradedummies.frame)
```

I also created a matrix for both raceeth and grade, converted both matrices into data frames which I combined to my existing dataset. By doing this I coded all my dummy variables into 0's and 1's.

The code below partitions my data set into train and test by performing an eighty-twenty split.

```
## Perform an eighty-twenty split to partition the data set
## into train and test data sets.
```{r}
set.seed(123)
partition <- sample(2,nrow(pisacleansed),replace = TRUE, prob= c(0.80,0.20))
```
```

```
train <- pisacleansed[partition==1, ]
test <-  pisacleansed[partition==2, ]
```

Running a full model on the train dataset:

```
## full model
## Adjusted R-squared:  0.302 for A and correlation of 0.576
```{r}
modelA <- lm(readingscore ~ ., data = train)
summary(modelA)
prediction <- predict(modelA,test)
actual =test$readingscore
cor(prediction, actual)
vif(modelA)
```
```

Inspecting the model with the summary() function and using the vif() function to detect multicollinearity, results into the following summary stastics:

```
Call:
lm(formula = readingscore ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-251.61  -48.36    0.86   49.42  238.16

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   450.68863   15.35974   29.34  < 2e-16 ***
male                          -11.62048    2.97191   -3.91  9.5e-05 ***
preschool                       0.63580    3.33402    0.19   0.8488
expectbachelors                48.83817    4.00362   12.20  < 2e-16 ***
motherhs                        4.48293    5.64568    0.79   0.4272
motherbachelors                 9.15130    3.69139    2.48   0.0132 *
```

```
motherwork                          -3.96281    3.31500   -1.20   0.2320
fatherhs                             4.13945    5.17969    0.80   0.4243
fatherbachelors                     18.58506    3.80868    4.88  1.1e-06 ***
fatherwork                           3.19744    4.14032    0.77   0.4400
selfbornus                          -0.34568    6.65335   -0.05   0.9586
motherbornus                        -5.98379    6.29368   -0.95   0.3418
fatherbornus                         3.53586    5.89831    0.60   0.5489
englishathome                       10.62721    6.43702    1.65   0.0989 .
computerforschoolwork               23.48165    5.42273    4.33  1.5e-05 ***
read30minsaday                      32.21396    3.21504   10.02  < 2e-16 ***
minutesperweekenglish                0.02296    0.01026    2.24   0.0254 *
studentsinenglish                   -0.07578    0.21472   -0.35   0.7242
schoolhaslibrary                    -2.61366    8.23219   -0.32   0.7509
publicschool                       -17.89477    6.19895   -2.89   0.0039 **
urban                               -0.87291    3.70269   -0.24   0.8136
schoolsize                           0.00602    0.00205    2.94   0.0033 **
american_indian_alaska_native      -65.77260   14.60942   -4.50  7.0e-06 ***
asian                               -7.09277    8.59439   -0.83   0.4093
black                              -65.40745    5.23104  -12.50  < 2e-16 ***
hispanic                           -34.01712    5.06246   -6.72  2.2e-11 ***
morethanonerace                    -13.52083    8.08919   -1.67   0.0947 .
nativehawaiianOtherPacificIslander -12.30972   16.36644   -0.75   0.4520
G8                                 -90.67649   52.69481   -1.72   0.0854 .
G9                                 -51.44170    5.44736   -9.44  < 2e-16 ***
G11                                 17.22593    3.62163    4.76  2.1e-06 ***
G12                                 93.91797   43.29161    2.17   0.0301 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74.2 on 2689 degrees of freedom
Multiple R-squared:  0.31,  Adjusted R-squared:  0.302
F-statistic:   39 on 31 and 2689 DF,  p-value: <2e-16

[1] 0.576
                            male                          preschool
expectbachelors
```

```
                            1.09                            1.09
1.13
                        motherhs                motherbachelors
motherwork
                            1.55                            1.55
1.07
                        fatherhs                fatherbachelors
fatherwork
                            1.55                            1.61
1.05
                      selfbornus                  motherbornus
fatherbornus
                            1.46                            3.34
2.98
                  englishathome           computerforschoolwork
read30minsaday
                            2.21                            1.12
1.07
           minutesperweekenglish              studentsinenglish
schoolhaslibrary
                            1.01                            1.12
1.05
                   publicschool                           urban
schoolsize
                            1.48                            1.56
1.49
    american_indian_alaska_native                           asian
black
                            1.04                            1.47
1.13
                        hispanic                morethanonerace
nativehawaiianOtherPacificIslander
                            2.05                            1.04
1.06
                              G8                              G9
G11
```

```
                              1.01                                    1.07
1.06
                               G12
                              1.02
```

When you look at the vif values of the different independent variables in the model, there are all less than 10 so, there is no need to worry about multicollinearity.

Using all subsets selection method to select my features, I trained several models using the train data set two different models.

```
## full model
## Adjusted R-squared:  0.302 for A and correlation of 0.576
```{r}
modelA <- lm(readingscore ~ ., data = train)
summary(modelA)
prediction <- predict(modelA,test)
actual =test$readingscore
cor(prediction, actual)
vif(modelA)
```


Pruned selfbornus
Adjusted R-squared: 0.302 and correlation of 0.576
```{r}
modelB <- lm(readingscore ~
male+preschool+expectbachelors+motherhs+motherbachelors+motherwork+fatherhs+f
atherbachelors
+ fatherwork +
motherbornus+fatherbornus+englishathome+computerforschoolwork+read30minsaday
+
minutesperweekenglish+studentsinenglish+schoolhaslibrary+publicschool+urban+s
choolsize
```
```

```
+
american_indian_alaska_native+asian+black+hispanic+morethanonerace+nativehawa
iianOtherPacificIslander
+ G8 + G9 +G11 + G12, data = train)
summary(modelB)
prediction <- predict(modelB,test)
actual =test$readingscore
cor(prediction, actual)
```


Prune out school has a library
Adjusted R-squared: 0.303 and correlation of 0.576
```{r}
modelC <- lm(readingscore ~
male+preschool+expectbachelors+motherhs+motherbachelors+motherwork+fatherhs+f
atherbachelors
+ fatherwork +
motherbornus+fatherbornus+englishathome+computerforschoolwork+read30minsaday
+ minutesperweekenglish+studentsinenglish + publicschool+urban+schoolsize
+
american_indian_alaska_native+asian+black+hispanic+morethanonerace+nativehawa
iianOtherPacificIslander
+ G8 + G9 +G11 + G12, data = train)
summary(modelC)
prediction <- predict(modelC,test)
actual =test$readingscore
cor(prediction, actual)
```


Prune out urban
Adjusted R-squared: 0.303 and correlation of 0.576
```{r}
modelD <- lm(readingscore ~ male + preschool + expectbachelors + motherhs +
motherbachelors + motherwork + fatherhs + fatherbachelors + fatherwork +
```

```
motherbornus + fatherbornus + englishathome + computerforschoolwork +
read30minsaday + minutesperweekenglish+studentsinenglish + publicschool +
schoolsize + american_indian_alaska_native + asian + black + hispanic +
morethanonerace + nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12,
data = train)
summary(modelD)
prediction <- predict(modelD,test)
actual =test$readingscore
cor(prediction, actual)
```

Pruned studentsinenglish
Adjusted R-squared: 0.303 and correlation of 0.576
```{r}
modelE <- lm(readingscore ~ male + preschool + expectbachelors + motherhs +
motherbachelors + motherwork + fatherhs + fatherbachelors + fatherwork +
motherbornus + fatherbornus + englishathome + computerforschoolwork +
read30minsaday + minutesperweekenglish + publicschool + schoolsize +
american_indian_alaska_native + asian + black + hispanic + morethanonerace +
nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12, data = train)
summary(modelE)
prediction <- predict(modelE,test)
actual =test$readingscore
cor(prediction, actual)
```

Pruned Preschool
Adjusted R-squared: 0.303 and correlation of 0.576
```{r}
modelF <- lm(readingscore ~ male  + expectbachelors + motherhs +
motherbachelors + motherwork + fatherhs + fatherbachelors + fatherwork +
motherbornus + fatherbornus + englishathome + computerforschoolwork +
read30minsaday + minutesperweekenglish + publicschool + schoolsize +
american_indian_alaska_native + asian + black + hispanic + morethanonerace +
nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12, data = train)
summary(modelF)
```

```
prediction <- predict(modelF,test)
actual =test$readingscore
cor(prediction, actual)
```

Pruned fatherbornus
Adjusted R-squared: 0.304 and correlation of 0.576
```{r}
modelG <- lm(readingscore ~ male  + expectbachelors + motherhs +
motherbachelors + motherwork + fatherhs + fatherbachelors + fatherwork +
motherbornus  + englishathome + computerforschoolwork + read30minsaday +
minutesperweekenglish + publicschool + schoolsize +
american_indian_alaska_native + asian + black + hispanic + morethanonerace +
nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12, data = train)
summary(modelG)
prediction <- predict(modelG,test)
actual =test$readingscore
cor(prediction, actual)
```

pruned fatherwork
Adjusted R-squared: 0.304 and correlation of 0.576
```{r}
modelH <- lm(readingscore ~ male  + expectbachelors + motherhs +
motherbachelors + motherwork + fatherhs + fatherbachelors  + motherbornus  +
englishathome + computerforschoolwork + read30minsaday +
minutesperweekenglish + publicschool + schoolsize +
american_indian_alaska_native + asian + black + hispanic + morethanonerace +
nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12, data = train)
summary(modelH)
prediction <- predict(modelH,test)
actual =test$readingscore
cor(prediction, actual)
```

pruned motherhs

```
## Adjusted R-squared:  0.304 and correlation of 0.576
```{r}
modeli <- lm(readingscore ~ male  + expectbachelors  + motherbachelors +
motherwork + fatherhs + fatherbachelors  + motherbornus  + englishathome +
computerforschoolwork + read30minsaday + minutesperweekenglish + publicschool
+ schoolsize + american_indian_alaska_native + asian + black + hispanic +
morethanonerace + nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12,
data = train)
summary(modeli)
prediction <- predict(modeli,test)
actual =test$readingscore
cor(prediction, actual)
```

Pruned motherbornus
Adjusted R-squared: 0.304 and 0.575
```{r}
modelk <- lm(readingscore ~ male  + expectbachelors  + motherbachelors +
motherwork + fatherhs + fatherbachelors  + englishathome +
computerforschoolwork + read30minsaday + minutesperweekenglish + publicschool
+ schoolsize + american_indian_alaska_native + asian + black + hispanic +
morethanonerace + nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12,
data = train)
summary(modelk)
prediction <- predict(modelk,test)
actual =test$readingscore
cor(prediction, actual)
```

Prunned motherwork
```{r}
modell <- lm(readingscore ~ male  + expectbachelors  + motherbachelors  +
fatherhs + fatherbachelors  + englishathome + computerforschoolwork +
read30minsaday + minutesperweekenglish + publicschool + schoolsize +
american_indian_alaska_native + asian + black + hispanic + morethanonerace +
nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12, data = pisacleansed)
```
```

```
summary(model1)
prediction <- predict(modelk,test)
actual =test$readingscore
cor(prediction, actual)
```

5 fold cross validation to validate modelA
Average mean square error is 210964
```{r}
library(DAAG)
outA <- cv.lm(data = test
              , form.lm = formula((readingscore ~ .))
              , plotit = "Observed", m=5)
```

5 fold cross validation to validate our modelB
Average mean square error is 5741
```{r}
library(DAAG)
outB <- cv.lm(data = test
              , form.lm = formula((readingscore ~ male + preschool +
expectbachelors + motherhs + motherbachelors + motherwork +
fatherhs+fatherbachelors
+ fatherwork + motherbornus + fatherbornus + englishathome +
computerforschoolwork + read30minsaday
+ minutesperweekenglish + studentsinenglish + schoolhaslibrary + publicschool
+ urban + schoolsize
+ american_indian_alaska_native + asian + black + hispanic + morethanonerace
+ nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12))
              , plotit = "Observed", m=5)
```

5 fold cross validation to validate our modelC
Average mean square error is 5709
```{r}
```

```
library(DAAG)
outC <- cv.lm(data = test
              , form.lm = formula((readingscore ~ male + preschool +
expectbachelors + motherhs + motherbachelors + motherwork + fatherhs +
fatherbachelors
+ fatherwork + motherbornus + fatherbornus + englishathome +
computerforschoolwork + read30minsaday
+ minutesperweekenglish + studentsinenglish + publicschool + urban +
schoolsize
+ american_indian_alaska_native + asian + black + hispanic + morethanonerace
+ nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12))
              , plotit = "Observed", m=5)
```

5 fold cross validation to validate our modelD
Average mean square error is 5668
```{r}
library(DAAG)
outD <- cv.lm(data = test
              , form.lm = formula((readingscore ~ male + preschool +
expectbachelors + motherhs + motherbachelors + motherwork + fatherhs +
fatherbachelors + fatherwork + motherbornus + fatherbornus + englishathome +
computerforschoolwork + read30minsaday +
minutesperweekenglish+studentsinenglish + publicschool + schoolsize +
american_indian_alaska_native + asian + black + hispanic + morethanonerace +
nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12))
              , plotit = "Observed", m=5)
```

5 fold cross validation to validate our modelE
Average mean square error is 5692
```{r}
outE <- cv.lm(data = test
              , form.lm = formula((readingscore ~ male + preschool +
expectbachelors + motherhs + motherbachelors + motherwork + fatherhs +
fatherbachelors + fatherwork + motherbornus + fatherbornus + englishathome +
```

```
computerforschoolwork + read30minsaday + minutesperweekenglish + publicschool
+ schoolsize + american_indian_alaska_native + asian + black + hispanic +
morethanonerace + nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12))
             , plotit = "Observed", m=5)
```

5 fold cross validation to validate our modelF
Average mean square error of 5690
```{r}
outF <- cv.lm(data = test
             , form.lm = formula((readingscore ~ male  + expectbachelors +
motherhs + motherbachelors + motherwork + fatherhs + fatherbachelors +
fatherwork + motherbornus + fatherbornus + englishathome +
computerforschoolwork + read30minsaday + minutesperweekenglish + publicschool
+ schoolsize + american_indian_alaska_native + asian + black + hispanic +
morethanonerace + nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12))
             , plotit = "Observed", m=5)
```

5 fold cross validation to validate our modelG
Average mean square error of 5648
```{r}
outF <- cv.lm(data = test
             , form.lm = formula((readingscore ~ male  + expectbachelors +
motherhs + motherbachelors + motherwork + fatherhs + fatherbachelors +
fatherwork + motherbornus  + englishathome + computerforschoolwork +
read30minsaday + minutesperweekenglish + publicschool + schoolsize +
american_indian_alaska_native + asian + black + hispanic + morethanonerace +
nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12))
             , plotit = "Observed", m=5)
```

5 fold cross validation to validate our modelk
Average mean square error of
```{r}
outK <- cv.lm(data = test
```

```
            , form.lm = formula((readingscore ~ male  + expectbachelors  +
motherbachelors  + fatherhs + fatherbachelors  + englishathome +
computerforschoolwork + read30minsaday + minutesperweekenglish + publicschool
+ schoolsize + american_indian_alaska_native + asian + black + hispanic +
morethanonerace + nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12))
            , plotit = "Observed", m=5)
```

Recorded Adjusted R-squared and the average mean square error in the table

Model	Adjusted R-Squared	Average mean Square error
modelA	0.302	210964
modelB	0.302	5741
modelC	0.302	5709
modelD	0.303	5668
modelE	0.303	5692
modelF	0.303	5690
modelG	0.304	5648
ModelK	0..304	5572

I built a second order term on minutesperweekenglishS and I went a head to use all subset selection to prune more variables whose p-values were more the default level of significance, 0.05.

```
# model selection for second oder terms
## Adjusted R-squared:  0.316
```{r}
pisacleansed$minutesperweekenglishSQ <- pisacleansed$minutesperweekenglish^2
modeli <- lm(readingscore ~ male  + expectbachelors  + motherbachelors +
motherwork + fatherhs + fatherbachelors  + motherbornus  + englishathome +
```
```

```
computerforschoolwork + read30minsaday + minutesperweekenglish + publicschool
+ schoolsize + american_indian_alaska_native + asian + black + hispanic +
morethanonerace + nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12 +
minutesperweekenglishSQ, data = pisacleansed)
summary(modeli)
```


Prune motherwork
Adjusted R-squared: 0.316
```{r}
pisacleansed$minutesperweekenglishSQ <- pisacleansed$minutesperweekenglish^2
modeli <- lm(readingscore ~ male  + expectbachelors  + motherbachelors  +
fatherhs + fatherbachelors  + motherbornus  + englishathome +
computerforschoolwork + read30minsaday + minutesperweekenglish + publicschool
+ schoolsize + american_indian_alaska_native + asian + black + hispanic +
morethanonerace + nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12 +
minutesperweekenglishSQ, data = pisacleansed)
summary(modeli)
```


##Prune asian
Adjusted R-squared: 0.316
```{r}
pisacleansed$minutesperweekenglishSQ <- pisacleansed$minutesperweekenglish^2
modeli <- lm(readingscore ~ male  + expectbachelors  + motherbachelors  +
fatherhs + fatherbachelors  + motherbornus  + englishathome +
computerforschoolwork + read30minsaday + minutesperweekenglish + publicschool
+ schoolsize + american_indian_alaska_native  + black + hispanic +
morethanonerace + nativehawaiianOtherPacificIslander + G8 + G9 +G11 + G12 +
minutesperweekenglishSQ, data = pisacleansed)
summary(modeli)
```

```
## Pruned nativehawaiianOtherPacificIslander
## Adjusted R-squared:  0.316
```{r}
pisacleansed$minutesperweekenglishSQ <- pisacleansed$minutesperweekenglish^2
modeli <- lm(readingscore ~ male  + expectbachelors  + motherbachelors  +
fatherhs + fatherbachelors  + motherbornus  + englishathome +
computerforschoolwork + read30minsaday + minutesperweekenglish + publicschool
+ schoolsize + american_indian_alaska_native  + black + hispanic +
morethanonerace  + G8 + G9 +G11 + G12 + minutesperweekenglishSQ, data =
pisacleansed)
summary(modeli)
```

Pruned motherbornus
Adjusted R-squared: 0.316
```{r}
pisacleansed$minutesperweekenglishSQ <- pisacleansed$minutesperweekenglish^2
modeli <- lm(readingscore ~ male  + expectbachelors  + motherbachelors  +
fatherhs + fatherbachelors    + englishathome + computerforschoolwork +
read30minsaday + minutesperweekenglish + publicschool + schoolsize +
american_indian_alaska_native  + black + hispanic + morethanonerace  + G8 +
G9 +G11 + G12 + minutesperweekenglishSQ, data = pisacleansed)
summary(modeli)
```

Prune fatherhs
Adjusted R-squared: 0.315
```{r}
pisacleansed$minutesperweekenglishSQ <- pisacleansed$minutesperweekenglish^2
modeli <- lm(readingscore ~ male  + expectbachelors  + motherbachelors   +
fatherbachelors    + englishathome + computerforschoolwork + read30minsaday +
minutesperweekenglish + publicschool + schoolsize +
american_indian_alaska_native  + black + hispanic + morethanonerace  + G8 +
G9 +G11 + G12 + minutesperweekenglishSQ, data = pisacleansed)
summary(modeli)
```
```

I built two interaction terms:

- combining read30minsaday and minutesperweekenglish through a multiplication process.
- combining motherbachelors and fatherbachelors through a multiplication process.

This is demonstrated below:

```
## Check for interaction terms
## Adjusted R-squared:  0.317
## two interactio terms
```{r}
pisacleansed$minutesperweekenglishSQ <- pisacleansed$minutesperweekenglish^2
modeli <- lm(readingscore ~ male  + expectbachelors  + motherbachelors   +
fatherbachelors    + englishathome + computerforschoolwork + read30minsaday +
minutesperweekenglish + publicschool + schoolsize +
american_indian_alaska_native  + black + hispanic + morethanonerace  + G8 +
G9 +G11 + G12 + minutesperweekenglishSQ + read30minsaday
*minutesperweekenglish + motherbachelors* fatherbachelors, data =
pisacleansed)
summary(modeli)
```
```

using the plot(modeli) function to graph residual plot and using the leverage Vs residuals graph they were three observations 2657,507 and 1879 which I endup removing from my data sets since they acted as outliers

```
## Removing specific rows in r
```{r}
## Adding an index column to my data set
pisacleansed$generated_uid <- 1:nrow(pisacleansed)
pisacleansed <- pisacleansed[-c(2657,507,1879),]
```
```

After pruning the interaction term, read30minsaday:minutesperweekenglish, this is my final model:

```
##read30minsaday:minutesperweekenglish
```{r}
pisacleansed$minutesperweekenglishSQ <- pisacleansed$minutesperweekenglish^2
```
```

```
modeli <- lm(readingscore ~ male  + expectbachelors  + motherbachelors   +
fatherbachelors    + englishathome + computerforschoolwork + read30minsaday +
minutesperweekenglish + publicschool + schoolsize +
american_indian_alaska_native  + black + hispanic + morethanonerace  + G8 +
G9 +G11 + G12 + minutesperweekenglishSQ  + motherbachelors* fatherbachelors,
data = pisacleansed)
summary(modeli)
```

Running the summary function to inspect the model, below

are the summary statistics:

```
Call:
lm(formula = readingscore ~ male + expectbachelors + motherbachelors +
    fatherbachelors + englishathome + computerforschoolwork +
    read30minsaday + minutesperweekenglish + publicschool + schoolsize +
    american_indian_alaska_native + black + hispanic + morethanonerace +
    G8 + G9 + G11 + G12 + minutesperweekenglishSQ + motherbachelors *
    fatherbachelors, data = pisacleansed)

Residuals:
    Min      1Q  Median      3Q     Max
-252.22  -48.22    1.06   48.84  248.95

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.32e+02   9.28e+00   46.61  < 2e-16 ***
male                           -1.25e+01   2.61e+00   -4.78  1.8e-06 ***
expectbachelors                 5.33e+01   3.52e+00   15.14  < 2e-16 ***
motherbachelors                 4.26e+00   4.25e+00    1.00  0.31664
fatherbachelors                 9.95e+00   4.52e+00    2.20  0.02781 *
englishathome                   1.06e+01   4.46e+00    2.37  0.01793 *
computerforschoolwork           2.08e+01   4.74e+00    4.39  1.2e-05 ***
read30minsaday                  3.29e+01   2.83e+00   11.62  < 2e-16 ***
```

```
minutesperweekenglish            1.21e-01   2.01e-02    6.04  1.7e-09 ***
publicschool                    -1.77e+01   4.94e+00   -3.59  0.00034 ***
schoolsize                       6.55e-03   1.61e-03    4.07  4.9e-05 ***
american_indian_alaska_native   -6.43e+01   1.34e+01   -4.80  1.7e-06 ***
black                           -6.44e+01   4.54e+00  -14.18  < 2e-16 ***
hispanic                        -3.26e+01   3.79e+00   -8.60  < 2e-16 ***
morethanonerace                 -1.83e+01   7.02e+00   -2.60  0.00932 **
G8                              -8.89e+01   5.21e+01   -1.70  0.08840 .
G9                              -4.92e+01   4.86e+00  -10.12  < 2e-16 ***
G11                              1.51e+01   3.20e+00    4.71  2.6e-06 ***
G12                              5.35e+01   5.22e+01    1.03  0.30518
minutesperweekenglishSQ         -1.12e-04   1.82e-05   -6.15  8.5e-10 ***
motherbachelors:fatherbachelors  1.75e+01   6.42e+00    2.72  0.00653 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73.5 on 3380 degrees of freedom
Multiple R-squared:  0.323, Adjusted R-squared:  0.319
F-statistic: 80.6 on 20 and 3380 DF,  p-value: <2e-16
```

The F-tests look good which shows that something in the model is working. Adjusted R-squared increased slightly to 0.319 .

Instead of predicting readingscore, I will transform my dependent variable to the log of the reading score + 1 , this is what we are going to use in the final model.

```
## Variable transformation
## log of both independent variable
## Adjusted R-squared:  0.322
```{r}
pisacleansed$minutesperweekenglishSQ <- pisacleansed$minutesperweekenglish^2
modeli <- lm(log(readingscore + 1) ~ male  + expectbachelors  +
motherbachelors   + fatherbachelors    + englishathome +
computerforschoolwork + read30minsaday + minutesperweekenglish + publicschool
+ schoolsize + american_indian_alaska_native  + black + hispanic +
```
```

```
morethanonerace  + G8 + G9 +G11 + G12 + minutesperweekenglishSQ  +
motherbachelors* fatherbachelors, data = pisacleansed)
summary(modeli)
```

Running the summary function to inspect the final model:

```
Call:
lm(formula = log(readingscore + 1) ~ male + expectbachelors +
    motherbachelors + fatherbachelors + englishathome + computerforschoolwork
+
    read30minsaday + minutesperweekenglish + publicschool + schoolsize +
    american_indian_alaska_native + black + hispanic + morethanonerace +
    G8 + G9 + G11 + G12 + minutesperweekenglishSQ + motherbachelors *
    fatherbachelors, data = pisacleansed)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6238 -0.0874  0.0118  0.0985  0.4151

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    6.06e+00   1.86e-02  325.10  < 2e-16 ***
male                          -2.58e-02   5.25e-03   -4.92  9.2e-07 ***
expectbachelors                1.13e-01   7.08e-03   15.99  < 2e-16 ***
motherbachelors                4.35e-03   8.55e-03    0.51  0.61073
fatherbachelors                1.85e-02   9.08e-03    2.04  0.04152 *
englishathome                  2.11e-02   8.97e-03    2.35  0.01900 *
computerforschoolwork          4.22e-02   9.53e-03    4.42  1.0e-05 ***
read30minsaday                 6.43e-02   5.69e-03   11.30  < 2e-16 ***
minutesperweekenglish          2.46e-04   4.04e-05    6.09  1.2e-09 ***
publicschool                  -3.48e-02   9.93e-03   -3.50  0.00047 ***
schoolsize                     1.28e-05   3.24e-06    3.96  7.8e-05 ***
american_indian_alaska_native -1.32e-01   2.69e-02   -4.88  1.1e-06 ***
black                         -1.34e-01   9.13e-03  -14.63  < 2e-16 ***
hispanic                      -6.55e-02   7.63e-03   -8.59  < 2e-16 ***
```

```
morethanonerace                 -3.40e-02   1.41e-02   -2.41  0.01615 *
G8                              -2.02e-01   1.05e-01   -1.93  0.05430 .
G9                              -1.07e-01   9.78e-03  -10.93  < 2e-16 ***
G11                              2.89e-02   6.43e-03    4.50  7.2e-06 ***
G12                              1.19e-01   1.05e-01    1.13  0.25688
minutesperweekenglishSQ         -2.23e-07   3.66e-08   -6.10  1.2e-09 ***
motherbachelors:fatherbachelors  3.28e-02   1.29e-02    2.54  0.01110 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.148 on 3380 degrees of freedom
Multiple R-squared:  0.326, Adjusted R-squared:  0.322
F-statistic: 81.8 on 20 and 3380 DF,  p-value: <2e-16
```

Interpreting the results of final model.

Looking at the betas, we can define the regression line as:

```
Readingscore = 6.06e+00 + (-2.58e-02)male +  (1.13e-01)expectbachelors +
4.35e-03motherbachelors + 1.85e-02fatherbachelors
+ 2.11e-02englishathome + 4.22e-02computerforschoolwork + 6.43e-
02read30minsaday + 2.46e-04minutesperweekenglish + (-3.48e-02)publicschool
+ 1.28e-05schoolsize + (6.06e+00-1.32e-01)american_indian_alaska_native +
(6.06e+00 + (-1.34e-01))black + (6.06e+00-6.55e-02)hispanic + (6.06e+00-
3.40e-02)morethanonerace
+ (6.06e+00 + (-2.02e-01))G8 + (6.06e+00 + (-1.07e-01))G9 + (6.06e+00 +
(2.89e-02))G11 + (6.06e+00 + 1.19e-01 )G12 + (-2.23e-
07)minutesperweekenglishSQ + 3.28e-02motherbachelors:fatherbachelors
```

This is the model that minimizes the sum of the square of errors.

Looking at the F value of 81.8 and the P-Value of less than 2e-16. The P-Value is the probability that given the null hypothesis, that all the Betas associated with the independent variables are equal to zero. We would observe the data as extreme as it is. Since the P-value is very small, so we are going to reject the null hypothesis and accept the alternative, that at least of the Betas is not equal to 0. We don’t know

which Beta or they are all not equal to zero. It is not what the F-TEST tells us. This is a test of the model itself, which tells me that something in my model is working. Adjusted R-Squared is 0.322; 32.2 percent of the variability in the readingscore is explained by the model.

Looking at the individual P-Value for male, expectbachelors, fatherbachelors, englishathome, computerschoolwork, read30minsaday, minutesperweekenglish, publicschool, schoolsize, American_indian_alaska_native, black, Hispanic, morethanonerace, G8, G9, G11, minutesperweekenhglishSQ and motherbachelors:fatherbachelors from the t-test, those are the p-values for the null hypothesis that Betas associated with those variables are equals zero. Because the P-Values are low, we are going to reject that null hypothesis and accept the alternative that the Betas associated with those parameters are not equal to zero and then use their estimations.

G10 will have an estimation of 6.06e+00, G8 will have an estimation of 6.06e+00 + (-2.02e-01), G9 will have an estimation of 6.06e+00 + (-1.07e-01), G11 will have an estimation of 6.06e+00 + (2.89e-02), white will have an estimation of 6.06e+00, black will have an estimation of 6.06e+00 + (-1.34e-01).

Something to note here the child,motherbachelors, of the interaction term motherbachelors:fatherbachelor is not significant since P-value don't look good. Because we always keep the children of the interaction term regardless of their P-value, I will motherbachelor in the model.

Also, I left G12 in the model because if I prune it out it affects the adjusted R-squared by decreasing it.

In summary whites have higher reading scores than blacks or hispanics, also kids whose parents have both attained bachelors' degrees perform better. Lastly, kids who spend 30 mins reading a day achieve good reading scores.

Also attached is my code.