Exercise4

Build a model to predict the IQ of a child based on covariates about his/her mom using the data set kidiq.csv. Below are the variables contained therein

**kid score:** IQ score of the kid

**mom hs:** whether the mom has attained high school (1/0)

**mon iq:** mom’s IQ score

**mon work:** a numerical variable ranges from 1 to 4,

– 1 =did not work in the first three years of the child’s life

– 2 =worked in the 2nd or 3rd year of child’s life

– 3 =worked part-time in the first year of child’s life

– 4 =worked full-time in the first year of child’s life

**mom age:** age of the mom when delivering the child.

Which variables do you recommend to be included into the independent variables? Please report all the steps and results of your analysis. Please also interpret your results.

After testing several models, we have decided to choose the following model: kid score =-11.48+51.27mom hs+0.97mom iq-0.48mom hs\*mom iq

Following are the steps and analysis:

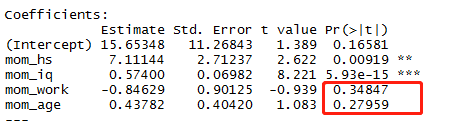
1. Plot Scatter between kid score and each variable.

2. Split data into training set and testing set

3. Construct model: regress kid score on the rest of variables and observe the result.

> model1 = lm(kid\_score ~ mom\_hs+mom\_iq+mom\_work+mom\_age , data=Train)

> summary(model1)

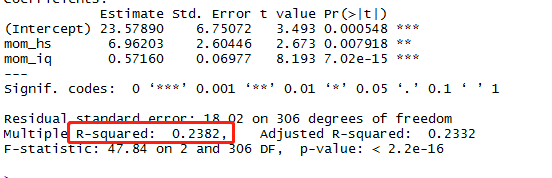


From above picture, we find that the P-value of the mom work and mom age are greater than 0.05. It implies that mom work and mom age are insignificant variables. We build a new model by drop these two variables.

4. Regress kid score on mom hs and mom iq.

> model2=lm(kid\_score ~ mom\_hs+mom\_iq,data=Train)

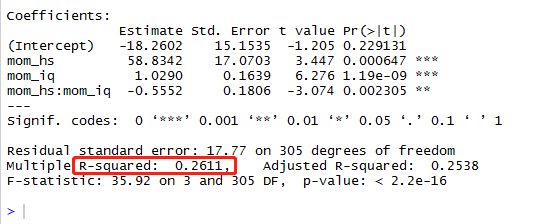
> summary(model2)



5. We find variables are siganificant. Then, we want to improve the R squared. The interaction term is taken into consideration.

> model3=lm(kid\_score ~ mom\_hs+mom\_iq+mom\_hs\*mom\_iq,data=Train)

> summary(model3)

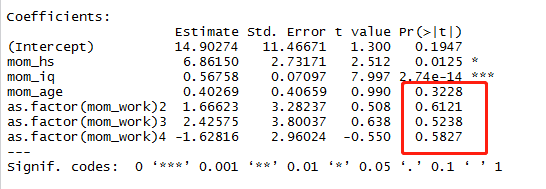


R squared is improved.

6. We observerd that mom\_work can be regraded as factor variables, then we could construct a new model.

> model4=lm(kid\_score ~ mom\_hs+mom\_iq+mom\_age+as.factor(mom\_work),data=Train)

> summary(model4)



mom\_work are also not significant variables.

7, We use testing set to test model2 and model3.

> predictTest2 = predict(model2, newdata=Test) # model2

> predictTest2

> # Compute R-squared

> SSE\_kidIQ2 = sum((Test$kid\_score - predictTest2)^2)

> SST\_kidIQ2 = sum((Test$kid\_score - mean(Test$kid\_score))^2)

> 1 - SSE\_kidIQ2/SST\_kidIQ2

[1] 0.1415191

> # Make test set predictions

> predictTest3 = predict(model3, newdata=Test) # model3

> predictTest3

> # Compute R-squared

> SSE\_kidIQ3 = sum((Test$kid\_score - predictTest3)^2)

> SST\_kidIQ3 = sum((Test$kid\_score - mean(Test$kid\_score))^2)

> 1 - SSE\_kidIQ3/SST\_kidIQ3

[1] 0.1393573

From the test data, we should choose the model:

kid\_score= 23.57890+ 6.96203\*mom\_hs+ 0.57160\* mom\_iq

In fact, the difference between model2 and model3 is very small. You also could choose model3.

kid\_score= -18.2602+ 58.8342\*mom\_hs+ 1.0290\* mom\_iq -0.5552\* mom\_hs\*mom\_iq.